Colecția *UNIVERSITARIA*

Bogdan IONESCU

ANALIZA și PRELUCRAREA SECVENȚELOR VIDEO

Indexarea automată după conținut

Bogdan IONESCU

ANALIZA şi PRELUCRAREA SECVENŢELOR VIDEO

Indexarea automată după conţinut

EDITURA TEHNICA

Bucureşti, 2009

Adresă: **Editura TEHNICĂ**
Str. Olari, nr. 23, sector 2
Bucureşti, România
cod 024056

www.tehnica.ro

Referenţi ştiinţifici:

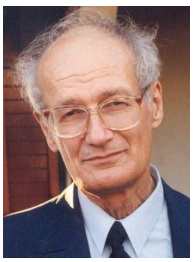
Prof. univ. dr. ing. Constantin VERTAN
Conf. univ. dr. ing. Mihai CIUC

Lucrare editată cu sprijinul
Consiliului Naţional al Cercetării Ştiinţifice
din Învăţământul Superior

Descrierea CIP a Bibliotecii Naţionale a României
IONESCU, BOGDAN
Analiza şi prelucrarea secvenţelor video: indexarea automată după conţinut / Bogdan Ionescu. – Bucureşti: Editura Tehnică, 2009
Bibliogr.
ISBN 978-973-31-2354-5

621.397.42
681.772.7

Prefaţă



Fostul meu student şi actualul coleg dr. Bogdan Ionescu m-a rugat să-i prefaţez această primă carte, ceea ce fac cu plăcere - el fiind dintr-o pleiadă de tineri care aduc cinste şcolii româneşti. Bogdan Ionescu şi-a terminat de curând (un an) un doctorat în cotutelă (România - Franţa) în domeniul relativ nou şi foarte important al prelucrării semnalelor multidimensionale, şi a avut şansa unui subiect a cărui stringenţă creşte pe zi ce trece: indexarea după conţinut a bazelor de date video (mai pe româneşte: a bibliotecilor de filme - în cazul lui, de filme artistice de animaţie).

Ce vrea să zică asta - indexarea după conţinut - cititorul va găsi în primul capitol, dar sunt tentat să zic şi aici, în aceste rânduri, câteva cuvinte: problema nu e chiar nouă. Cu ceva zeci de ani în urmă am aflat că pe alte meleaguri oamenii se ocupau, pentru cuvinte, cu alcătuirea unor asemenea dicţionare. Cele alfabetice, pe care le avem şi noi, îţi explică ce vrea să zică un cuvânt pe care îl ai dar al cărui sens nu îl ştii; dar sunt şi probleme de alt fel: acolo era un exemplu de întâmplare în academia spaniolă - un vorbitor nu-şi aducea aminte cum se cheamă un om născut pe vapor (noi n-avem cuvânt pentru acest concept). Ne trebuie dicţionare care să ne ducă de la concept la cuvânt. Despre unele popoare primitive se zice că aveau zeci de cuvinte pentru a denumi diferite tipuri de nori; noi n-avem, dar am putea eventual descrie forma lor, mişcarea lor, ca să precizăm la care ne referim când vrem să povestim o întâmplare concretă.

Într-o bibliotecă de un miliard de cărţi, cu câte 500 de pagini fiecare şi cu 2000 de semne pe pagină avem nevoie doar de 50 de cifre binare pentru a identifica orice literă, ceea ce mi se pare extrem de puţin - la îndemâna

umanului: le cuprindem cu ochiul dintr-o privire, pe un rând. Oare nu e posibil să avem căi/o cale de a ajunge la "obiectul" dorit dintr-o colecţie vastă, cunoscându-l prin calităţile sale (făcute cumva măsurabile: da-nu, roşu-albastru-galben-verde, o valoare întreagă între 1 şi 100, 17 grade de turtire a unui cerc în elipsă, etc.)? "Obiectele" de care vorbeam pot fi entităţi foarte complexe: o imagine, o secvenţă de film mut, entităţi "multimodale" (vorbă, sunete, imagini, text, etc.). Parcă suntem tentaţi a zice da. Dar acum vine partea dificilă a problemei, şi în acelaşi timp frumoasă prin efortul de creaţie pe care ni-l cere (aspectul care ne provoacă, ne desfide, englezul ar zice "chalenging"): pe de o parte, în cazul concret al unei colecţii de un tip dat (de pietre, de gâze, de filme), care sunt atributele, cum le definim ca să caracterizăm cât mai compact şi mai corect, acea colecţie; pe de altă parte, în faţa unui obiect din colecţie, cum măsurăm *automat*, adică nu prin intervenţia omului (în cazul ăsta avem nevoie de un specialist în domeniu!), aceste atribute.

Fără acest mic amănunt aici, "automat", suntem pierduţi fiindcă operaţia manuală de adnotare cu atribute a obiectelor este consumatoare de timp în aşa măsură că ne face întreprinderea lipsită de sens.

În momentul de faţă al scurtei noastre istorii de câteva sute de ani, suntem în pericol de a fi "înecaţi în informaţii" care pe de o parte multe ne sunt vitale şi pe de alta, în ansamblul lor ne copleşesc, fără a putea ajunge la cele de care avem nevoie suntem ca însetatul din pustiu peste care năvăleşte marea. Indexarea automată după conţinut ne poate salva.

La laboratorul nostru din Politehnica bucureşteană, aceste preocupări sunt de dată mai veche (aş menţiona aici preocupările prof. Constantin Vertan în timpul unor stagii în Franţa şi apoi aici), dar tomul lui Bogdan Ionescu este prima carte dedicată acestui subiect, şi în particular indexării video, şi cred că trebuie s-o salutăm cu entuziasm fiind sosită într-un moment când e nevoie de ea. Sper că o vor urma altele şi că subiectul va atrage şi pe alţi tineri cercetători spre binele nostru al tuturor. Felicitări autorului pentru munca asiduă depusă şi calitatea lucrării rezultate.

Prof. univ. dr. ing. Vasile BUZULOIU
Bucureşti 17 noiembrie 2008

Cuvântul autorului

Indexarea automată după conţinut a datelor este un domeniu ce câştigă din ce în ce mai mult teren, datorită necesităţii crescânde de exploatare a volumelor mari de date. Dacă, nu până demult, puteam vorbi de o lipsă informaţională, progresul tehnologic a făcut ca în zilele noastre să ne confruntăm cu o adevărată explozie de informaţie.

Din acest amalgam informaţional, un interes aparte il au *informaţiile multimedia*, ce sunt definite ca fiind o combinaţie de tipuri de conţinut, printre care cele mai uzuale sunt: textul, sunetul şi imaginile.

În societatea modernă, informaţia multimedia face parte din viaţa noastră cotidiană şi imi este greu să-mi imaginez că vom mai putea vreodată să ne lipsim de ea. De exemplu, telefonul portabil a devenit indispensabil şi ne însoţeşte pretutindeni, acesta fiind un adevărat centru multimedia în miniatură. Prin intermediul acestuia, putem accesa informaţiile multimedia din reţeaua Internet, putem folosi mesageria electronică, putem înregistra, stoca, reda şi distribui filme sau imagini în orice moment. Fiecare persoană, a devenit astfel, cu voie sau fără voie, un "consumator" de date multimedia.

Motivată în principal de un interes comercial, dezvoltarea infrastructurii de stocare şi transmisie a datelor a dus la apariţia unei noi probleme, şi anume: *Cum accesăm informaţia multimedia utilă dintr-un vast amalgam de date? Cum facem să găsim aceea informaţie pe care o dorim?* Problema ar fi una simplă în cazul a câtorva date, dar când o astfel de colecţie poate conţine la ordinul a sute de mii de documente video, de exemplu, care la rândul lor conţin sute de mii de imagini, problema pare imposibil de rezolvat.

Soluţia actuală existentă este dată de *sistemele de indexare după conţinut* a bazelor de date. Conceptul de indexare este definit ca fiind procesul de adnotare a informaţiei existente într-o colecţie de date, prin adăugarea de

informaţii suplimentare despre conţinutul acesteia, informaţii numite şi *indici* de conţinut. Pe baza indicilor, sistemul poate grupa datele în funcţie de similaritate, în categorii, subcategorii şi aşa mai departe. De exemplu, dacă dispunem de o bază de documente video, ideal, în urma indexării automate, acestea pot fi regrupate în funcţie de gen în: filme, muzică, desene animate, ştiri, etc., sau la un nivel de detaliu mai ridicat, în subcategorii precum: film de ficţiune, dramă, documentar, etc. În acest fel, căutarea informaţiei dorite este restrânsă la căutarea într-o subcategorie din care aceasta face parte, reducând astfel timpul de căutare şi totodată imbunătăţind precizia căutării.

Pe de altă parte, procesul de indexare nu este opţional, ci este strict necesar într-o colecţie mare de date. În acest caz, o informaţie care nu a fost indexată este practic inexistentă pentru utilizator, cu toate că aceasta este prezentă în bază. Să luăm exemplul simplu al unui sistem de indexare a fişierelor, prezent în orice sistem de operare. Acesta ordonează datele în funcţie de nume, tipul conţinutului, data creării etc., în directoare şi subdirectoare. Dacă un anumit fişier nu a fost indexat, cu toate că acesta se află fizic pe dispozitivul de stocare, acesta este transparent pentru utilizator, fiind imposibil de localizat.

Sistemele de indexare, pe parcursul evoluţiei, au trecut de la o abordare sintactică a procesului de adnotare la o abordare semantică, cum este cazul sistemelor actuale. Diferenţa dintre acestea este una semnificativă. *Adnotarea sintactică* se limitează la caracterizarea conţinutului datelor cu atribute numerice de nivel scăzut, precum măsuri statistice, diverşi parametri, etc. Din păcate, o astfel de abordare este implicit adresată unui public avizat în domeniu, căutarea informaţiei necesitând cunoştinţe tehnice. Pe de altă parte, *adnotarea semantică* are ca scop descrierea conţinutului datelor într-un mod cât mai apropiat de modul de percepţie uman. Astfel, localizarea datelor devine naturală şi accesibilă publicului larg, fiind ghidată de un limbaj textual. De exemplu, căutarea filmelor în funcţie de valorile vitezei medii de deplasare a obiectelor în scenă nu este evidentă, pe când o căutare în funcţie de conţinutul de acţiune (redus, ridicat) este pe înţelesul tuturor.

Această lucrare vine să adreseze tocmai această problematică a indexării automate după conţinut a datelor multimedia, punând accentul pe secvenţele de imagini, domeniu de mare actualitate în străinătate în acest moment, dar încă la începuturi în România.

Această lucrare propune un studiu bibliografic detaliat al literaturii de specialitate din acest domeniu, abordând direcţiile fundamentale de analiză şi prelucrare a secvenţelor de imagini în contextul sistemelor de indexare după conţinut. Astfel, sunt prezentate atât aspecte teoretice (principii şi metode), cât şi exemple concrete (sisteme, aplicaţii), punând la dispoziţia cititorului

o bibliografie mai mult decât generoasă (peste 280 de citări ale unor articole din reviste şi conferinţe internaţionale de specialitate). Cartea este adresată atât începătorilor în domeniul prelucrării şi analizei de imagini şi video, cât şi celor deja experimentaţi, constituind un ghid de bună practică şi totodată un sistem de indexare a realizărilor semnificative din domeniu.

Ideea scrierii acestui manuscris, a apărut în urmă cu mai bine de cinci ani, odată cu demararea tezei mele de doctorat realizată în cotutelă, pe de-o parte la laboratorul LAPI - Laboratorul de Analiza şi Prelucrarea Imaginilor din Universitatea "Politehnica" din Bucureşti, sub îndrumarea Domnului Profesor Vasile Buzuloiu, cât şi la laboratorul, la vremea respectivă, LAMII - Laboratoire d'Automatique et de Micro-Informatique Industrielle din Université de Savoie, sub îndrumarea Domnului Profesor Patrick Lambert. Tematica abordată a constat în studiul şi dezvoltarea unui sistem de indexare automată după conţinut a secvenţelor de animaţie din cadrul Festivalului Internaţional al Filmului de Animaţie de la Annecy, echivalentul în domeniul animaţiei al festivalului de film de la Cannes. Studiul bibliografic şi cercetarea detaliată realizată cu această ocazie, precum şi faptul că doar o parte din acestea au putut fi valorificate în teza de doctorat (din motive obiective de spaţiu), m-au condus spre ideea unei posibile redactări ulterioare a unei cărţi dedicate.

Această idee avea să se concretizeze după susţinerea tezei de doctorat, când am participat la competiţia de granturi de Resurse Umane, organizată de CNCSIS - Consiliului Naţional al Cercetării Ştiinţifice din Învăţământul Superior, programul RP de stimulare a revenirii în ţară a tinerilor cercetători români. Proiectul propus venea să continue natural cercetarea realizată în străinătate până în acel moment, şi anume propunea extinderea studiului indexării spre baze de date generice de documente video, precum şi dezvoltarea unei aplicaţii software de adnotare şi navigare virtuală în baza de date. Obţinerea grantului RP-2 mi-a permis actualizarea studiului bibliografic realizat anterior, îmbunătăţirea acestuia, precum şi dezvoltarea de noi direcţii de studiu.

Astfel, rezultatele cercetării până în acest moment s-au concretizat în şapte capitole. În primul capitol am detaliat problematica indexării după conţinut a datelor multimedia, punând accentul pe metodele de analiză şi adnotare de conţinut a secvenţelor de imagini, ce fac subiectul acestei lucrări. De asemenea, am realizat o trecere în revistă a tehnicilor de indexare a imaginilor, sunetului, secvenţelor de imagini şi, respectiv, video.

Capitolul al doilea abordează o problemă de prelucrare a secvenţelor de imagini ce este premergătoare adnotării propriu-zise a conţinutului, dar totodată necesară, şi anume segmentarea temporală a secvenţei, atât în unităţi sintactice (plane video), cât şi semantice (scene video). Segmentarea

temporală, prin detecţia schimbărilor de plan, permite înţelegerea structurii temporale a secvenţei, necesară în etapele ulterioare de prelucrare, indiferent că este vorba de o indexare sintactică sau de nivel semantic superior.

Capitolul al treilea propune o analiză a metodelor de caracterizare a informaţiei fundamentale a secvenţelor de imagini şi anume mişcarea. Pornind de la studierea problematicii estimării mişcării la nivel de imagine, am realizat o trecere în revistă a diverselor direcţii de studiu abordate de metodele de analiză şi caracterizare a conţinutului de mişcare din secvenţă.

Capitolul al patrulea abordează o altă informaţie reprezentativă a secvenţelor de imagini, ce joacă un rol important în percepţia vizuală, şi anume, conţinutul de culoare. Pornind de la modalităţile clasice de reprezentare a culorilor folosind spaţiile de culoare, şi ajungând până la o descriere perceptuală cu ajutorul teoriei culorilor, am realizat o trecere în revistă a modalităţilor de caracterizare a conţinutului de culoare, atât static, la nivel de imagine, cât şi dinamic, la nivel de secvenţă de imagini.

Capitolul al cincilea propune un studiu al metodelor de rezumare automată a conţinutului secvenţelor de imagini, atât statică (în imagini) cât şi dinamică (în mişcare). Rezumarea de conţinut joacă un rol important pentru indexare, deoarece permite pe de-o parte reducerea drastică a timpului vizualizării datelor dintr-o bază mare de date, cât şi reducerea redundanţei informaţionale pentru alte etape de prelucrare.

Capitolul al şaselea face trecerea dintre nivelul sintactic de adnotare şi cel semantic, prin abordarea tehnicilor de formalizare cu concepte fuzzy a datelor numerice de nivel scăzut.

În final, capitolul al şaptelea prezintă un studiu al tehnicilor de clasificare nesupervizată a datelor (automată), cât şi al tehnicilor de clasificare supervizată (ce folosesc o etapă de învăţare). Tehnicile de clasificare prezintă un real interes pentru procesul de indexare, deoarece pe baza atributelor de conţinut, determinate în etapa de adnotare, acestea pot grupa datele în colecţii de date omogene.

Pentru mai multe detalii referitor la aspecte aplicative ale prelucrării şi analizei secvenţelor de imagini în contextul indexării, cititorul poate consulta site-ul proiectului de indexare RP-2, şi anume: `http://alpha.imag.pub.ro/VideoIndexingRP2/`.

Sper sincer ca această lucrare să constituie un ajutor şi o referinţă pentru cei interesaţi de problemele prelucrării secvenţelor de imagini, şi că alţii o să-mi urmeze exemplul şi o să ducă mai departe cercetarea românească din acest domeniu.

Ş.l. univ. dr. ing. Bogdan IONESCU
Bucureşti 30 noiembrie 2008

Mulţumiri

Această lucrare nu s-ar fi concretizat fără suportul grantului de cercetare CNCSIS - Consiliului Naţional al Cercetării Ştiinţifice din Învăţământul Superior, Resurse Umane, RP-2 (2007-2009), intitulat "Dezvoltarea de Metode de Indexare Semantică după Conţinut a Bazelor de Documente Video: Aplicaţii la Navigare, Căutare şi Rezumare Automată a Conţinutului"[1]. În acest sens, ţin să mulţumesc *Domnului Preşedinte CNCSIS Profesor Ioan Dumitrache*, *Domnului Vicepreşedinte Profesor Mihai Gîrţu* şi *Domnului Director Profesor Adrian Curaj*, iniţiatorii programului RP de reintegrare. De asemenea, ţin să mulţumesc *Doamnei Director Adjunct Magdalena Crîngaşu* şi *Doamnei Consilier Adriana Rotar*, pentru ajutorul acordat cât şi pentru informaţiile preţioase oferite pe durata desfăşurării proiectului.

Ţin să mulţumesc laboratorului LAPI - Laboratorul de Analiza şi Prelucrarea Imaginilor, din Universitatea "Politehnica" din Bucureşti, şi astfel *Domnului Profesor Vasile Buzuloiu*, pentru acceptarea mea în colectivul de cercetare, pentru prietenia arătată de-a lungul timpului cât şi pentru încadrarea preţioasă acordată pe parcursul formării mele profesionale. Mulţumesc colegilor mei profesori, *Constantin Vertan* şi *Mihai Ciuc*, pentru ajutorul important, pentru atmosfera cordială din cadrul laboratorului precum şi pentru modelul de conduită arătat.

Mulţumesc în mod special *Domnului Profesor Adrian Badea* şi *Domnului Profesor Corneliu Burileanu*, pentru prietenia acordată, pentru ajutorul preţios, pentru sugestiile valoroase şi pentru suportul constant de-a lungul formării mele ştiinţifice.

Ţin să mulţumesc *Domnului Profesor Nicolae Vasiliu* pentru ajutorul

[1] vezi site-ul proiectului "`http://alpha.imag.pub.ro/VideoIndexingRP2/`".

acordat publicării acestei lucrări şi pentru sfaturile preţioase. De asemenea, mulţumesc *Domnului Profesor Ilie Prisecaru* pentru suportul acestuia şi pentru ajutorul acordat.

Mulţumesc de asemenea *Domnului Profesor Teodor Petrescu* şi *Domnului Profesor Dan Stoichescu* pentru sprijinirea activităţii mele de cercetare în cadrul Facultăţii de Electronică, Telecomunicaţii şi Tehnologia Informaţiei şi a Catedrei de Electronică Aplicată şi Ingineria Informaţiei, precum şi pentru ajutorul acordat.

Vreau să mulţumesc laboratorului LISTIC - Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance, Annecy, Franţa, şi astfel *Domnului Profesor Philippe Bolon* pentru co-finanţarea tezei mele de doctorat realizată în domeniul analizei şi prelucrării secvenţelor de imagini. De asemenea, mulţumesc călduros *Domnului Profesor Patrick Lambert* şi *Domnului Profesor Didier Coquin* pentru încadrarea mea pe parcursul tezei de doctorat, pentru sfaturile acordate cât şi pentru suportul constant al acestora pe toată durata cercetării efectuate la Annecy.

De asemenea, ţin să mulţumesc tuturor colaboratorilor externi ce au susţinut proiectul de indexare video pe care l-am iniţiat:

- *Domnul Profesor Daniel Bouillot*, IMUS - Institut de Management de l'Université de Savoie şi CITIA - Cité de l'Image en Mouvement,

- *Domnul Profesor Patrick Lambert* şi *Domnul Profesor Philippe Bolon* - LISTIC, Polytech'Savoie, Annecy-Franţa,

- *Domnul Profesor Robert Laganière* - VIVA - The Video, Image, Vision and Autonomous Systems Research Laboratory, Ottawa-Canada,

- *Domnul Emmanuel Quillet* şi *Domnul Director Hervé Lièvre* CERIMES - Centre de Ressources et d'Information sur les Multimédias pour L'Enseignement Supérieur din Ministère Enseignement Supérieur et Recherche Français.

Mulţumesc în mod special prietenei mele Monica care m-a sprijinit în tot ce am întreprins până în prezent şi care a avut răbdarea să corecteze acest manuscris.

Cu această ocazie, ţin să mulţumesc călduros *Doamnei Eugenia Burcea* pentru tot sprijinul preţios acordat şi *Doamnei Director Diana Cocârţă* ce mi-a sugerat redactarea acestei cărţi.

Nu în ultimul rând, vreau să mulţumesc călduros Editurii Tehnice şi astfel *Domnului Director Roman Chirilă* pentru acceptarea publicării acestei cărţi, pentru finanţarea a o parte din costurile de publicare, precum şi pentru munca deloc neglijabilă depusă pentru aducerea la viaţă a manuscrisului acestei cărţi.

Remerciements

Ces travaux ne se seraient véritablement concrétisés sans le soutien assuré par le Grant de recherche du CNCSIS, le Conseil National de la Recherche Scientifique de l'Enseignement Supérieure de la Roumanie, Ressources Humaines, RP-2 (2007-2009), intitulé "Le Développement de Méthodes d'Indexation Sémantique du Contenu des Documents Vidéo: Application à la Navigation, Recherche et Résumé Automatique du Contenu"[2]. A ce titre là, je tiens à remercier *M. Ioan Dumitrache*, *M. Mihai Gîrţu* et *M. Adrian Curaj*, les promoteurs du programme RP. Je tiens également à remercier *Mme Adriana Rotar* et *Mme Magdalena Crîngaşu* pour leur aide et pour les informations précieuses fournies pendant le déroulement du projet.

Je tiens à remercier le laboratoire LAPI de l'Université "Politehnica" de Bucarest, Laboratoire d'Analyse et Traitement d'Images, et particulièrement *M. Vasile Buzuloiu*, pour m'avoir accueilli au sein de son équipe de recherche, pour son amitié et son encadrement précieux pendant ma formation professionnelle. Je remercie également mes collègues professeurs, *M. Constantin Vertan* et *M. Mihai Ciuc*, pour leur aide précieuse, pour la bonne ambiance qu'ils ont su créer au sein du laboratoire et pour l'exemple de conduite qu'ils ont été.

Je remercie particulièrement à *M. Adrian Badea* et *M. Corneliu Burileanu*, pour leur amitié, leur aide précieuse, leurs nombreux conseils et leur soutien constant tout au long de ma formation scientifique.

Je remercie *M. Nicolae Vasiliu* pour m'avoir aidé dans la publication de ce livre et pour ses précieux conseils. Egalement, je remercie *M. Ilie Prisecaru* pour son soutien et son aide.

[2]voir la page web du projet `http://alpha.imag.pub.ro/VideoIndexingRP2`.

Je veux également adresser tous mes remerciements à *M. Teodor Petrescu* et *M. Dan Stoichescu* pour leur aide et pour avoir encouragé et soutenu mon activité de recherche à la Faculté d'Electronique, Télécommunications et Technologie de l'Information et au département d'Electronique Appliquée et d'Ingénierie de l'Information.

Je tiens à remercier le laboratoire LISTIC, Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance d'Annecy, ainsi que *M. Philippe Bolon*, directeur du LISTIC, pour le cofinancement de ma thèse de doctorat portant sur l'analyse et le traitement de séquences d'image. Je remercie également chaleureusement *M. Patrick Lambert* et *M. Didier Coquin* pour leur encadrement pendant ma thèse, leurs nombreux conseils et leur soutien constant tout au long de mes stages de recherche à Annecy.

J'adresse tous mes remerciements aux collaborateurs étrangers qui ont soutenu le projet d'indexation vidéo que j'ai monté à Bucarest:

- *M. Daniel Bouillot* de l'IMUS - Institut de Management de l'Université de Savoie et CITIA - Cité de l'Image en Mouvement,

- *M. Patrick Lambert* et *M. Philippe Bolon* du LISTIC, Polytech'Savoie, Annecy-France,

- *M. Robert Laganière* de VIVA - The Video, Image, Vision and Autonomous Systems Research Laboratory, Ottawa-Canada,

- *M. Emmanuel Quillet* et *M. Hervé Lièvre* du CERIMES - Centre de Ressources et d'Information sur les Multimédias pour l'Enseignement Supérieur du Ministère Enseignement Supérieur et Recherche Français.

Je remercie tout particulièrement mon amie Monica pour son soutien, pour la patience dont elle a fait preuve pour corriger ce manuscrit et pour avoir été constamment proche de moi.

A cette occasion, je remercie chaleureusement *Mme Eugenia Burcea* pour son aide très précieuse et *Mme Diana Cocârţă* qui m'a suggéré de réaliser ce livre.

Enfin, je tiens à remercier chaleureusement la Maison d'Edition "Editura Tehnică" ainsi que *M. Roman Chirilă* pour avoir accepté la publication de ces travaux, pour le financement d'une partie des frais et aussi pour l'aide apportée pour donner vie à ce manuscrit.

Cuprins

CAPITOLUL 1

Conceptul de indexare după conţinut

Rezumat: *Un rol important în societatea modernă il are informaţia multimedia (imagini, sunet, text, video). Datorită exploziei tehnologice cu care ne confruntăm, volumul informaţional a devenit foarte mare. Problema actuală nu este lipsa de informaţie, ci găsirea informaţiei utile într-un vast amalgam de informaţii. Soluţia este dată de sistemele de indexare după conţinut a datelor. În acest capitol vom face o trecere în revistă a literaturii de specialitate din domeniu, prezentând din perspectiva proprie diversele metode şi tehnici folosite de sistemele actuale de indexare după conţinut a imaginilor, a sunetului, a secvenţelor de imagini, precum şi video. Prezentarea se va focaliza pe tehnicile de analiză şi adnotare a conţinutului secvenţelor de imagini, ce fac subiectul acestei cărţi.*

Dacă în urmă cu aproximativ un deceniu, cantitatea de informaţie multimedia disponibilă era foarte redusă, iar accesarea acesteia se realiza dificil şi ineficient, în zilele noastre putem vorbi despre o *explozie informaţională*. Accesul la date, fie că este vorba de imagini, sunet, text sau video, a devenit strict necesar şi face parte integrantă din viaţa noastră de zi cu zi.

Evoluţia tehnologică a dispozitivelor de achiziţie şi prelucrare a datelor (calculatoare personale, medii de stocare, dispozitive de redare şi captură audio-video) cât şi a infrastructurii de transmisie de date (protocoale de transmisie fără fir: WiFi, BlueTooth, reţele LAN de mare viteză, telefonia

multimedia 3G) au dus la creşterea exponenţială a volumului informaţional prin simplificarea stocării şi prelucrării acestuia. Astfel *că problema cu care ne confruntăm acum, nu este lipsa de informaţie, ci, dimpotrivă imposibilitatea de a selecţiona din volumul informaţional imens disponibil, informaţia utilă* căutată.

Pentru a răspunde acestei noi provocări, cercetările actuale din domeniile prelucrării de semnal şi a vederii asistate de calculator, au dus la elaborarea a ceea ce numim *sisteme de indexare după conţinut* sau CBRS ("Content-Based Retrieval Systems").

1.1 Definirea conceptului de indexare

Conceptul de indexare este definit ca fiind **procesul de adnotare** a informaţiei existente într-o colecţie de date, prin adăugarea de informaţii suplimentare relative la conţinutul acesteia [Kyungpook 06], informaţii numite şi *indici* de conţinut. Această etapă este necesară accesării colecţiei de date, deoarece permite catalogarea automată în funcţie de conţinut a datelor.

Într-o colecţie de date suficient de vastă, putem spune că datele care nu au fost adnotate sunt practic inexistente pentru utilizator. Un exemplu simplu de sistem de indexare este însuşi sistemul de fişiere al oricărui calculator personal. Acesta ne furnizează datele aflate pe diversele medii de stocare (disc dur, memorie externă, etc.) sub formă de fişiere ce sunt indexate după nume, extensie, dată, etc. Să ne imaginăm situaţia în care un fişier a fost omis din această listă de indici, cu toate că el este prezent fizic pe suportul de stocare, acesta va fi invizibil şi inaccesibil pentru utilizatorul de rând.

Procesul de adnotare a datelor este văzut din două perspective: pe de-o parte există *adnotarea manuală*, iar pe de altă parte *adnotarea automată*. Gradul de complexitate al adnotării este direct proporţional cu nivelul de detaliu dorit pentru accesarea datelor. Dacă se doreşte ca utilizatorul să poata accesa datele folosind criterii mai complexe, ca de exemplu căutarea unei anumite secvenţe video pentru care nu se cunoaşte nici numele, nici extensia fişierului, dar totuşi utilizatorul dispune de informaţii referitoare la conţinutul vizual al acesteia, în această situaţie, procesul de indexare va fi mult mai complex, necesitând întelegerea de către calculator a conţinutului datelor.

Astfel, în cazul unei indexări după criterii complexe de conţinut, adnotarea manuală este foarte dificil de realizat, deoarece necesită un număr important de operatori umani. Aceştia ar trebui să "răsfoiască" manual întregul conţinut al bazei de date pentru definirea indicilor de conţinut. Luând în calcul că o astfel de colecţie poate conţine milioane de înregistrări, timpul

necesar indexării manuale devine semnificativ. În acest moment, cercetările existente în domeniu se focalizează pe dezvoltarea de algoritmi de adnotare automată a conţinutului, mai ales în cazul datelor ce necesită un timp important pentru vizualizare, ca de exemplu documentele video.

Cu toate că adnotarea conţinutului datelor este solutia optimală pentru a accesa informaţia utilă dintr-o vastă colecţie de date, aceasta nu este şi suficientă. Adnotarea în sine nu oferă decat o serie de date suplimentare, putem spune, de nivel scăzut, care deseori sunt inaccesibile utilizatorului neavizat.

Pentru a accesa baza de date, utilizatorul trebuie să dispună de o interfaţă grafică software prin care să poată accesa sau vizualiza uşor datele, fie pe baza indicilor, fie în mod direct. Aceasta trebuie să aibă o funcţionalitate naturală şi intuitivă. Sistemul care permite utilizatorului să vizualizeze conţinutul bazei de date poartă numele de **sistem de navigare**.

Pe de altă parte, accesul la date presupune un proces de căutare. Utilizatorul trebuie să mai dispună, pe lângă sistemul de navigare, de utilitare software care să-i permită căutarea informaţiilor dorite în baza de date. Căutarea se realizează prin formularea de cereri de căutare sau "queries". Pentru uşurinţă, o astfel de cerere trebuie să fie exprimată într-un limbaj natural, apropiat de limbajul uman, cum ar fi de exemplu "caută filmele de acţiune" sau "caută imaginile ce conţin peisaje". Sistemul care răspunde acestor cerinţe poartă numele de **sistem de căutare**.

În Figura 1.1 am prezentat diagrama simplificată de funcţionare a unui sistem de indexare.

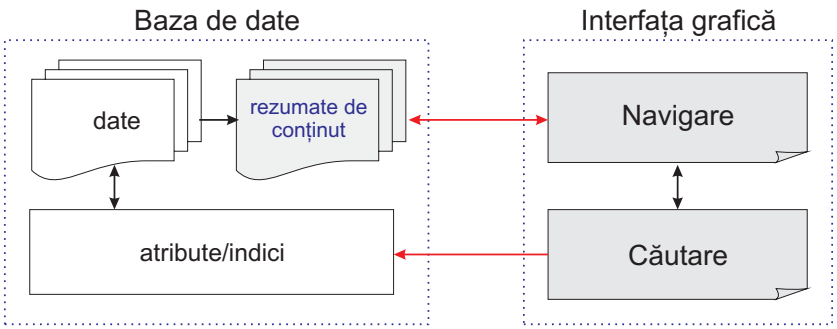


Figura 1.1: Principiul de funcţionare al unui sistem de indexare după conţinut.

Astfel, pentru realizarea indexării după conţinut, în primă etapă sistemul analizează datele din baza de date în vederea generării indicilor/atributelor de conţinut. Aceştia pot fi: fie date de *nivel semantic scăzut*, precum măsuri

statistice, parametri numerici (de exemplu în cazul documentelor video: histograme de culoare, câmpuri vectoriale de mişcare, etc.), fie date simbolice de *nivel semantic ridicat* (de exemplu în cazul imaginilor statice: obiecte de interes, percepţia culorilor, recunoaştere text "încrustat" în imagine, etc.). În final, fiecare înregistrare va avea asociată o colecţie de astfel de atribute ce vor caracteriza conţinutul acesteia.

Pentru etapa de adnotare a bazei de date, timpul disponibil nu este critic, în sensul că aceasta este efectuată o singură dată în mod "offline", în momentul creării bazei de date (utilizatorul nu este conectat la baza de date). Tot în această etapă, opţional, se pot genera *descrieri compacte ale conţinutului* datelor, precum scurte rezumate pentru secvenţele video sau pasaje de text reprezentative pentru documentele textuale. Aceste descrieri vor fi folosite ulterior de sistemul de navigare pentru a facilita accesul la date, utilizatorul putând apela la aceste informaţii pentru a evita accesarea integrală a conţinutului acestora. Totuşi, la cererea utilizatorului, sistemul de navigare poate furniza un acces direct la baza de date.

Sistemul de căutare va permite utilizatorului să localizeze datele dorite prin formularea de cereri de căutare. Acestea, după cum am menţionat anterior, trebuiesc formulate într-un limbaj natural apropiat de modul de percepţie umană. Pentru a fi înţelese de sistem, cererile de căutare sunt mai întâi convertite în indici folosind acelaşi mecanism ca şi în cazul adnotării iniţiale a bazei de date. Mai departe, căutarea propriu-zisă se efectuează prin compararea indicilor de căutare cu cei deja existenţi în baza de date. Rezultatele căutării vor fi acele date ale căror indici sunt cei mai apropiaţi din punct de vedere al unuia sau a mai multor *criterii de similaritate* de indicii de căutare (de exemplu, pentru indici numerici este vorba de măsuri de distanţă, baze de reguli, etc.).

Rezultatele sunt puse la dispoziţia utilizatorului prin intermediul sistemului de navigare. Frecvent, în această etapă, pentru a ameliora performanţele sistemului, se permite interacţia cu utilizatorul prin ceea ce numim "feedback"[1] al sistemului. Astfel, utilizatorul işi va exprima gradul de satisfacţie faţă de rezultatele obţinute, sistemul autoinstruindu-se în funcţie de preferinţele acestuia, pentru a furniza datele cele mai relevante pentru căutare.

În concluzie, sistemele de indexare actuale sunt rezultatul necesităţii accesării după conţinut a datelor multimedia, ce sunt din ce în ce mai diverse. Tehnicile existente au fost adaptate la tipul de date ce trebuiesc indexate, astfel întâlnim:

[1] conceptul de feedback este definit în general ca fiind procesul prin care o proporţie din semnalul de ieşire al unui sistem este trecut la intrarea sistemului. În contextul indexării după conţinut, "feedback"-ul sistemului constă în informaţia furnizată de utilizator pentru ameliorarea performanţelor acestuia.

- sisteme **CBIR** ("Content-Based Image Retrieval"): dezvoltate pentru accesarea bazelor de imagini statice (imagini medicale, fotografii, picturi, etc.);
- sisteme **CBAR** ("Content-Based Audio Retrieval"): dezvoltate pentru accesarea bazelor de înregistrări audio (muzică, voce, sunete, etc.);
- sisteme **CBISR** ("Content-Based Image Sequence Retrieval"): dezvoltate pentru indexarea secvenţelor de imagini, înţelegând prin aceasta, orice înşiruire temporală de imagini statice sau orice document video în absenţa informaţiei sonore;
- sisteme **CBVR** ("Content-Based Video Retrieval"): dezvoltate pentru indexarea bazelor de documente audio-vizuale (ştiri, sport, filme, etc.).

Pentru a înţelege mai bine conceptul de indexare a secvenţelor video precum şi a tehnicilor folosite de sistemele existente, tematică ce face subiectul acestei cărti, în cele ce urmează vom prezenta succint caracteristicile generale ale sistemelor CBIR şi CBAR. Problematica adnotării conţinutului de imagine şi respectiv sunet influenţează în mod direct tehnicile de indexare a documentelor video, acestea din urmă fiind informaţii audio-vizuale.

1.2 Sistemele de indexare de imagini

Unele dintre primele sisteme de indexare au fost sistemele de indexare a bazelor de date de imagini statice sau CBIRS ("Content-Based Image Retrieval Systems"). Necesitatea acestora a fost dată de creşterea semnificativă a numărului de imagini ce trebuiau accesate şi stocate, în special în domeniile prioritare precum medicină (baze de date de imagini medicale: radiografii, tomografii, etc.) sau domeniul militar (baze de imagini satelitare), dar şi de interese pur comerciale, cum ar fi comercializarea bazelor de date de imagini destinate domeniului Prepress[2].

Global tehnicile de adnotare a conţinutului imaginilor sunt orientate către trei axe principale de analiză, şi anume: *analiza de culoare*, *analiza formelor* şi *analiza de textură* [Smeulders 00].

[2]termenul Prepress este folosit în industria tipografică şi de publicare pentru a desemna procesele şi procedurile software (procesare imagini, text) cât şi hardware (creare film tipografic, pregătire tipar) necesare pregătirii manuscrisul sau a creaţiei artistice pentru imprimarea finală pe suportul fizic (hârtie).

Analiza de culoare este una dintre tehnicile de adnotare a imaginilor cel mai frecvent folosită, deoarece însuşi sistemul vizual uman este bazat pe prelucrarea informaţiei de culoare (unde luminoase de diverse frecvenţe). Astfel, sistemele de indexare a imaginilor ce folosesc exclusiv analiza culorilor sunt folosite în aplicaţii practice în care distribuţia de culoare este trăsătura fundamentală a imagini, precum indexarea picturilor [Lay 04] sau a fotografiilor [Flickner 95]. Culorile sunt analizate folosind diverse spaţii de culoare, de la cele clasice, precum sistemul RGB (Roşu-Verde-Albastru), până la sisteme perceptuale în care culorile sunt structurate pentru a fi în concordanţă cu percepţia vizuală umană, precum sistemul HSV (Tentă-Saturaţie-Intensitate) sau Lab (Luminozitate şi diferenţe cromatice) [Mojsilovic 00] (un studiu detaliat al spaţiilor de culoare este prezentat în Secţiunea 4.1).

Analiza formelor se foloseşte de proprietăţile geometrice ale obiectelor conţinute în imagine pentru a caracteriza scena. Această analiză presupune detecţia în prealabil a acestora folosind tehnici de segmentare bazate pe contur sau pe regiuni de pixeli. Succesul adnotării este astfel direct condiţionat de calitatea segmentării în obiecte a imaginii. Mai mult, caracteristicile obiectelor din scenă, obţinute în urma analizei formelor, nu trebuie să fie dependente de unghiul sub care a fost prelevată imaginea. Un anumit obiect trebuie caracterizat în acelaşi fel chiar dacă a fost imortalizat din unghiuri diferite. Metodele de adnotare existente propun pentru a soluţiona această problemă folosirea de descriptori de formă invarianţi la transformările geometrice ce pot surveni în imagine [Rivlin 95]. De asemenea, se încearcă şi soluţionarea problemei suprapunerii diferitelor obiecte de interes, suprapunere survenită în urma proiecţiei scenei reale 3D, în spaţiul 2D al imaginii. Aceasta poate duce la recunoaşterea eronată a obiectului de interes [Schmid 97].

Analiza de textură este de asemenea des folosită, deoarece permite caracterizarea proprietăţilor materialelor prezente în imagine, ca de exemplu caracterizarea texturii pielii pentru detecţia persoanelor sau a feţelor prezente în imagine. Metodele existente de analiză de textură folosesc, fie metode clasice de caracterizare a texturii: matrice de co-ocurenţă, parametri fractali, etc. [Gimel'farb 96], fie abordări mai complexe, cum ar fi analiza Markoviană [Choi 98].

Tendinţa actuală a sistemelor de indexare de imagini constă în dezvoltarea de metode mixte ce folosesc colaborarea celor trei modalităţi de analiză, şi anume, culoare-formă-textură, profitând astfel de avantajele oferite de fiecare dintre cele trei surse de informaţie. Ne limităm în această lucrare la

prezentarea a doar câtorva generalităti ale sistemele CBIR. Pentru un studiu bibliografic detaliat al tehnicilor de indexare de imagini existente, cititorul se poate raporta la lucrarea de sinteză propusă în [Smeulders 00].

Dezvoltarea sistemelor de indexare după conţinut a imaginilor a dus la apariţia a două noi provocări în cercetarea ştiinţifică din domeniu. Prima dintre ele este cunoscută sub numele de **paradigma semantică**[3] ("semantic gap") ce este enunţată ca fiind: *discrepanţa dintre informaţiile extrase în mod automat din imagine şi semnificaţia semantică pe care le-o putem atribuii acestora*. Astfel, tehnicile de indexare de imagini duc deseori la rezultate corecte din punct de vedere al algoritmilor de calcul, dar care au o semnificaţie semantică redusă pentru utilizator. A doua problemă este **paradigma senzorială** ("sensor gap") ce este enunţată ca fiind: *discrepanţa care există între informaţiile prezente în scena reală 3D şi informaţiile furnizate de imagine*, imagine ce reprezintă o proiecţie discretă 2D obţinută în momentul înregistrării scenei [Smeulders 00].

În concluzie, datorită acestor limitări informaţionale şi tehnologice, sistemele de indexare de imagini existente au tendinţa să furnizeze rezultate ce nu sunt întotdeauna în conformitate cu realitatea, prezentă în scena reală, a cărei reprezentare o constituie imaginea.

1.3 Sistemele de indexare a sunetului

O altă categorie de sisteme de indexare sunt sistemele de indexare a sunetului sau CBARS ("Content-Based Audio Retrieval Systems"). În acest caz, datele indexate sunt documentele audio, precum înregistrările de voce sau înregistrările muzicale.

În general, pentru adnotarea conţinutului audio, datele sunt analizate temporal folosind două niveluri de detaliu. Un prim nivel de analiză este *nivelul cadrelor*. Un cadru audio este definit ca fiind o fereastră temporală de analiză de durată redusă. Un al doilea nivel de analiză este *nivelul segment*, un segment audio fiind o fereastră temporală de lungă durată sau chiar întreaga secvenţă audio ("long-term clip level"). Atributele de conţinut specifice datelor audio sunt calculate atât în domeniul *temporal* cât şi *frecvenţial*. Metodele folosite sunt în mare parte metode clasice ale domeniului de prelucrare de semnal sau de prelucrare a vocii. Proprietăţile audio sunt exprimate cu parametri de nivel scăzut specifici, precum: volum, număr de treceri prin zero ale semnalului, tonalitate, parametri spectrali etc. [Wang 00].

[3]termenul de semantică este definit global ca fiind relaţia între semne şi lucrurile la care acestea se referă sau înţelesul lor ("denotata").

O primă direcţie larg abordată de sistemele de indexare audio o constituie indexarea vocii [Naphade 02]. În acest caz, pentru a uşura analiza, metodele de adnotare a conţinutului sunt aplicate în condiţii simplificate, cum ar fi de exemplu absenţa zgomotului de fond sau folosirea unui dicţionar de termeni predefinit [Research 05].

O altă direcţie de studiu este clasificarea automată a sunetelor în categorii predefinite, precum: voce, muzică, scene de violenţă, etc. Metodele existente folosesc cu predilecţie două abordări: analiza bazată pe reguli şi analiza bazată pe modele [Naphade 02].

De mare interes comercial s-a dovedit a fi clasificarea automată a genurilor muzicale, folosită pentru căutarea în bazele de documente muzicale disponibile în format electronic, precum mp3, ogg[4], etc. Acestea sunt de regulă indexate după gen (blues, clasic, jazz), numele artistului şi titlul piesei muzicale. Metodele de indexare folosite utilizează în general algoritmi de clasificare, precum metoda K-means, reţelele neuronale sau sistemele expert (pentru o descriere detaliată a tehnicilor de clasificare a datelor, vezi Capitolul 7). Clasificarea este efectuată în acest caz după parametri specifici, precum: timbrul muzical, armonicitate, melodicitate şi ritm [Scaringella 06].

Interesul pentru sistemele de indexare audio, putem spune că este în principal unul comercial, datorat creşterii exponenţiale a volumului de documente muzicale "tranzacţionate" în format electronic. Pe de altă parte, tehnicile avansate de adnotare semantică a conţinutului audio, precum recunoaşterea vocii sau a vorbitorului, îşi au aplicaţie în principal în sistemele de indexare multimodală complexe (imagine-sunet-text), precum sistemele de indexare video, fiind rar dezvoltate individual.

1.4 Sistemele de indexare a secvenţelor de imagini

Înlocuirea suportului magnetic de stocare a secvenţelor video cu stocarea în format digital, a orientat sistemele de indexare a imaginilor către secvenţele de imagini. Sistemele de indexare a secvenţelor de imagini sau CBISRS ("Content-Based Image Sequence Retrieval Systems") sunt la origine extensia temporală a sistemelor de indexare a imaginilor CBIR. În cazul CBISR, datele prelucrate nu sunt imagini statice independente, ci serii temporale de

[4]Ogg Vorbis este un format de compresie audio comparabil cu celelalte formate existente pentru stocarea şi redarea audio digitală. Diferenţa faţă de acestea constă în faptul ca nu este patentat, putând fi astfel folosit liber, fără a necesita drept de autor. Acesta poate încapsula atât informaţie video cât şi text.

imagini sau pe scurt *imagini în mişcare* (dinamice). Deseori se întâmplă ca secvenţele de imagini să fie numite artificial filme sau documente video. Diferenţa dintre acestea constă în faptul că documentele video conţin în plus informaţia audio. În acest context, o secvenţă de imagini poate fi definită ca fiind informaţia spaţio-temporală a unui document video.

În cazul sistemelor de indexare a secvenţelor de imagini, paradigma senzorială (vezi Secţiunea 1.2) este mai puţin sesizabilă datorită informaţiilor suplimentare furnizate de acestea, dintre care cea mai importantă este **informaţia temporală**. Pe de altă parte, particularităţile sistemelor de indexare a secvenţelor de imagini, vor aduce o serie de noi dificultăţi şi cerinţe de prelucrare.

O primă problemă este *cantitatea importantă de date* ce trebuiesc prelucrate. La o frecvenţă de 25 de imagini pe secundă (de exemplu standardul PAL[5]) o secvenţă de 10 minute conţine nu mai puţin de 15000 de imagini statice, numite şi cadre ("frames"). Astfel, o singură secvenţă de imagini poate fi echivalentul mai multor baze de date de imagini statice. În cazul unei baze de secvenţe de imagini, volumul de date este absolut impresionant, deoarece, în mod uzual, aceasta poate conţine mii de secvenţe. Volumul important de date antrenează după sine dificultatea accesului şi a prelucrării acestora.

Pe de altă parte, pe lângă informaţia spaţială furnizată de imagini, secvenţele de imagini mai conţin *informaţie temporală*. Dacă într-un sistem de indexare a imaginilor, două imagini ce conţin aceleaşi obiecte, pot fi considerate ca fiind similare ca şi conţinut, într-un sistem de indexare a secvenţelor de imagini, două secvenţe ce conţin aceleaşi obiecte pot avea un conţinut complet diferit, acest lucru datorându-se aspectului dinamic sau temporal. Astfel, comportamentul obiectelor şi evoluţia temporală a acestora în scenă sunt informaţii esenţiale pentru înţelegerea conţinutului unei sevenţe de imagini şi drept urmare, pentru procesul de indexare a acestora.

O altă specificitate a secvenţelor de imagini este *structura ierarhică* a informaţiei. Într-o secvenţă, imaginile sunt grupate în ceea ce numim *plane video*. Acestea constituie unitatea sintactică de bază a secvenţei şi sunt definite ca fiind grupuri de imagini ce au proprietatea de unitate temporală, spaţială şi de acţiune. La un nivel semantic superior, conţinutul secvenţei este structurat în unităţi semantice, precum scenele sau episoadele (structura temporală a secvenţelor de imagini este detaliată pe larg în Capitolul 2).

Tinând cont de diversitatea informaţională furnizată de secvenţele de imagini, tehnicile existente de prelucrare au trebuit să se adapteze pentru a ţine cont de aspectul temporal şi structural al datelor.

[5]PAL este prescurtarea pentru Phase Alternating Line ce reprezintă un standard de condare a semnalului TV color, pe 625 de linii, la frecvenţa de 50Hz.

În cele ce urmează, vom prezenta pe larg metodele de prelucrare folosite de sistemele de indexare a secvenţelor de imagini existente, evidenţiind cele trei părţi componente de prelucrare ale acestora şi anume: *sistemul de adnotare a conţinutului*, *sistemul de navigare* în baza de date şi *sistemul de căutare*.

1.4.1 Principiul adnotării de conţinut

După cum am menţionat în capitolul introductiv, procesul de adnotare al conţinutului datelor constă în *crearea atributelor* sau a indicilor ce permit sistemului înţelegerea automată a conţinutului bazei de date. Aceste informaţii sunt în general proprietăţi ale secvenţei extrase la nivel de pixel, la nivel de regiuni de pixeli, la nivel de imagine sau grup de imagini.

Indexarea de imagini, fiind un caz particular, simplificat, al indexării secvenţelor de imagini, face ca metodele specifice de prelucrare a secvenţelor să aibă la bază informaţiile folosite la analiza imaginilor statice, şi anume: analiza de culoare, textură şi formă. Acestea caracterizează proprietăţile spaţiale ale imaginilor.

Specificitatea secvenţelor de imagini constă pe de-o parte în analiza temporală a *evoluţiei atributelor* extrase la nivel de imagine, precum şi în analiza *structurii temporale* a secvenţei şi a *informaţiei dinamice de miscare*. În Figura 1.2 am sintetizat diferitele surse de informaţii ce intervin în procesul de adnotare al conţinutului secvenţelor de imagini.

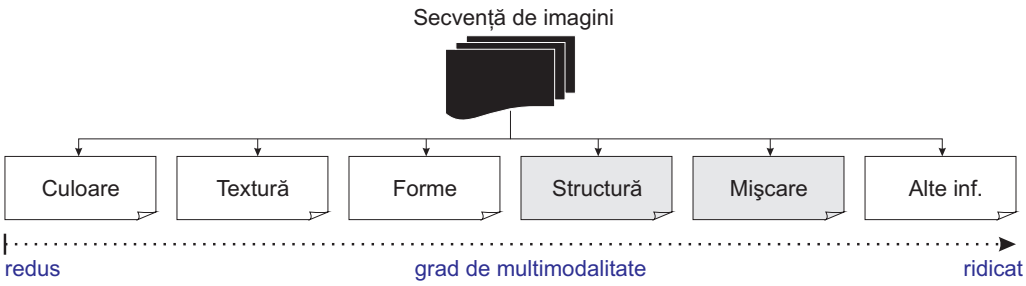


Figura 1.2: Sursele de informaţii exploatate de sistemele de indexare a secvenţelor de imagini (elementele marcate cu culoarea gri sunt specifice secvenţelor).

În cele ce urmează, vom face o trecere în revistă a metodelor şi a tehnicilor de indexare ce folosesc fiecare dintre modalităţile menţionate anterior.

Culoarea este unul dintre atributele cele mai importante pentru caracterizarea conţinutului unei imagini. Analiza conţinutului de culoare permite

găsirea similarităţilor vizuale dintre secvenţele de imagini. În particular, pentru a cuantifica similaritatea locală sau globală, marea parte a metodelor existente folosesc măsuri statistice, precum *histogramele color*. O astfel de abordare multi-rezoluţie este propusă în [Calic 02b] unde pentru a realiza indexarea după conţinut, histogramele color sunt calculate pe imagini cu diverse rezoluţii. Sistemul propune astfel utilizatorului un număr variabil de niveluri de detaliu, iar o măsură de pertinenţă este calculată pentru a controla gradul de degradare al imaginilor.

Totuşi, măsurile statistice bazate pe histogramă nu conţin informaţie spaţială. Pentru a rezolva această problemă, s-au propus diferite soluţii. De exemplu, [Chen 99] foloseşte histograme augmentate în care fiecărui pixel i se adaugă informaţii statistice suplimentare referitoare la similaritatea faţă de pixelii vecini (medie statistică, entropie, varianţă, etc.).

Metodele bazate pe histogramă nu permit nici analiza evoluţiei temporale, specifică secvenţelor de imagini. Pentru a depăşii aceste inconveniente, unele metode propun caracterizarea conţinutului secvenţei pe baza studiului evoluţiei temporale a unor vectori de caracteristici locale de culoare. O astfel de abordare este propusă în [Zhong 97], unde secvenţele de imagini sunt caracterizate prin proprietăţi specifice obiectelor de interes: culori specifice, dimensiuni, poziţie şi traiectorie de deplasare.

Drept alte metode de analiză a conţinutului de culoare în vederea indexării, putem menţiona metodele ce folosesc arbori de decizie fuzzy pentru extragerea regulilor de indexare [Detyniecki 03], sau modele ale distribuţiei de rapoarte de culoare ("color ratio models") ce sunt calculate folosind informaţia de contur [Adjeroh 01] (o descriere detaliată a metodelor de analiză a culorii în contextul indexării după conţinut este prezentată în Capitolul 4).

Textura în cazul secvenţelor de imagini este folosită cu precădere pentru caracterizarea proprietăţilor materialelor prezente în scenă sau a obiectelor de interes. Metodele existente folosesc în general aceiaşi parametri de textură ca în cazul indexării imaginilor statice [Vertan 04]. În literatura de specialitate întâlnim puţine tehnici de adnotare a conţinutului ce se folosesc exclusiv de analiza de textură, marea parte a metodelor existente folosind abordări mixte în care se regăseşte şi informaţia de textură. Cu toate acestea, ca exemple, putem menţiona analiza de textură folosită în [Chang 99] pentru segmentarea cadrelor, sau metoda propusă în [Bouthemy 98] unde câmpul vectorial de mişcare al secvenţei este văzut din prisma unei texturi, iar proprietăţile temporale ale acestuia sunt caracterizate cu ajutorul parametrilor specifici, precum matricea de co-ocurenţă.

Formele. Parametrii ce caracterizează forma obiectelor de interes, ca şi în

cazul sistemelor de indexare a imaginilor statice, sunt analizaţi în domeniul spaţial al imaginii. Specificitatea secvenţelor de imagini constă în faptul că obiectele se deplasează în scenă, deplasare ce se traduce în spaţiul imaginii prin transformări geometrice progresive ale obiectului vizat. Astfel, *invarianţa la transformări geometrice* este una dintre proprietăţile fundamentale ce trebuiesc îndeplinite de descriptorii de formă ce vor servi la indexarea conţinutului secvenţei. Descriptorii de formă cel mai frecvent folosiţi sunt momentele invariante şi descriptorii Fourier. În [Mehtre 97] eficienţa descriptorilor bazaţi pe contur (Fourier, UFF - "UNL Fourier Features", etc.) este comparată cu cea a descriptorilor bazaţi pe regiuni de pixeli (momente invariante, momente Zernike). Invarianţa acestora este imbunătăţită prin propunerea unor abordări mixte între diversele tipuri de descriptori: momente invariante şi descriptori Fourier, sau momente invariante şi caracteristici UFF. O altă direcţie de studiu este analiza evoluţiei temporale a formelor obiectelor, des întâlnită şi în metodele de urmărire temporală a obiectelor ("object tracking"), cum ar fi de exemplu metoda bazată pe modele multi-rezoluţie a contururilor active propusă în [Mazière 00].

Structura temporală. Dacă analiza de culoare, textură şi formă sunt specifice atât imaginilor statice cât şi secvenţelor de imagini, structura temporală este o informaţie specifică doar secvenţelor de imagini şi documentelor video. În etapa de montaj a secvenţei, planele video sunt concatenate pentru a defini ceea ce are să fie evoluţia temporală a evenimentelor secvenţei (vezi Capitolul 2). Metodele de adnotare a conţinutului folosite de sistemele de indexare a secvenţelor de imagini sunt bazate pe *segmentarea temporală* în plane, pe *extragerea de imagini cheie* ("key frames"[6]) şi pe *analiza similarităţii* între unităţile temporale ale secvenţei: scene, episoade, etc. Modalitatea structurală în care a fost constituită secvenţa oferă informaţii preţioase relative la conţinutul semantic al acesteia. Astfel, o secvenţă de acţiune va avea o densitate ridicată de plane video de scurtă durată, în timp ce o secvenţă a unui reportaj TV, este foarte probabil să conţină doar câteva plane video. Ca metode existente de analiză a modului în care este structurată secvenţa, putem menţiona folosirea de modele probabiliste precum lanţuri Markov ascunse [Ferman 99] sau adnotarea conţinutului pe baza analizei şi evaluării scenelor video [Vendriga 01].

Mişcarea. Analiza de mişcare este o etapă de analiză naturală în cazul secvenţelor de imagini, deoarece mişcarea reprezintă însăşi esenţa unei sec-

[6] o imagine cheie este o imagine considerată ca fiind reprezentativă pentru conţinutul unităţii structurale din care face parte (plan, scenă, segment, etc.).

venţe. În sistemele de indexare a secvenţelor de imagini, caracterizarea conţinutului de mişcare este realizată de regulă pe baza estimării câmpului vectorial de mişcare. Vectorii de deplasare sunt fie estimaţi la nivel de imagine, fie recuperaţi direct din fluxul video comprimat MPEG (vezi Secţiunea 3.1.5). O primă direcţie de studiu o constituie adnotarea *spaţio-temporală*. Aceasta include segmentarea, urmărirea de obiecte şi caracterizarea mişcării în cadrul anumitor pasaje de interes ale secvenţei. Ca exemplu putem menţiona sistemul propus în [Dagtas 00], unde mişcarea spaţio-temporală a obiectelor este folosită pentru a determina şi caracteriza evenimentele importante ale secvenţei; sau sistemul VideoQ propus în [Chang 98], dedicat exclusiv caracterizării globale a mişcării obiectelor.

O altă direcţie de studiu pentru adnotarea conţinutului de mişcare o constituie analiza mişcării camerei video. Metodele existente clasifică diversele mişcări globale în: mişcări translaţionale, de rotaţie, mărire sau micşorare ("zoom in", "zoom out"), basculare, etc. (vezi Secţiunea 3.2). În această categorie putem menţiona metoda propusă în [Lee 01] unde diversele mişcări ale camerei video sunt clasate folosind modele de mişcare predefinite şi reţele neuronale, sau abordarea din [Fablet 02] ce foloseşte modele Gibbs pentru a reprezenta derivatele semnalului spaţio-temporal al imaginii.

În ceea ce priveşte alte categorii de abordări, putem menţiona caracterizarea informaţiei de mişcare folosind traiectoria obiectelor sau a regiunilor de interes din imagine [Hsu 02], sau metoda inedită prezentată în [Zeng 02] ce propune transpunerea informaţiei temporale în domeniul spaţial al imaginii prin constituirea de hărţi de mişcare.

Alte direcţii de studiu folosesc diferite surse de informaţie obţinute în mod indirect. Una dintre abordările cel mai frecvent folosite pentru procesul de adnotare al conţinutului este detecţia şi analiza prezenţei persoanelor în scenă. Aceasta este realizată pe baza detecţiei caracteristicilor specifice acestora, precum: prezenţa culorii pielii în imagine, prezenţa feţei, a ochilor, etc. [Acosta 02]. Metodele de localizare a feţei sunt bazate pe tehnici de clasificare supervizată, precum reţelele neuronale sau modelele Markov ascunse [Ben-Yacoub 99]. În aceeaşi categorie putem include şi abordările bazate pe detecţia anumitor obiecte de interes din scenă, ce sunt considerate ca fiind reprezentative pentru conţinutul secventei, cum ar fi detecţia prezenţei maşinilor propusă în [Schneiderman 00].

O altă informaţie preţioasă furnizată de secvenţele de imagini este textul "încrustat" în imagine, text ce poate corespunde adnotărilor textuale, genericului, subtitrărilor, scorului în secvenţele sportive, numărului de înmatriculare al maşinilor, diverselor indicatoare, etc. Metodele de adnotare a conţinutului ce folosesc această informaţie sunt bazate pe recunoaşterea au-

tomată a caracterelor sau OCR ("Optical Character Recognition"[7]). Un exemplu este sistemul propus în [Kim 00b] unde regiunile din imagine ce conţin text încrustat sunt mai întâi izolate folosind o clasificare pe bază de reţele neuronale, iar mai departe literele sunt segmentate şi identificate.

1.4.2 Adnotarea semantică a conţinutului

Metodele existente de adnotare a conţinutului secvenţelor de imagini se împart în două mari categorii:

- **metode de adnotare sintactică**, ce sunt utilizate de prima generaţie de sisteme de indexare, precum cele enumerate în paragrafele anterioare,
- **metode de adnotare semantică**, ce reprezintă noua direcţie de analiză folosită de marea parte a sistemelor de indexare actuale.

Adnotarea sintactică este definită generic ca fiind adnotarea ce se referă la *relaţiile dintre unităţile de nivel scăzut constituente ale secvenţei şi modul de constituire a structurii acesteia.* Aceasta se poate realiza pe baza atributelor de nivel scăzut extrase din secvenţă, precum parametri statistici calculaţi la nivel de pixel sau regiuni de pixeli, proprietăţi geometrice ale obiectelor, structura temporală a secvenţei sau vectori de mişcare. De regulă, indicii obţinuţi în urma procesului de adnotare sunt valori numerice ce descriu atributele enumerate mai sus dar şi relaţiile sintactice ce pot exista între acestea. Extraşi la acest nivel de percepţie, indicii sintactici sunt accesibili doar pentru utilizatorul avizat. De exemplu, căutarea unei secvenţe de imagini care să conţină 30% mişcare de translaţie şi 20% mişcare de rotaţie, nu constituie o formulare prea relevantă pentru utilizator.

În contrast cu adnotarea sintactică, adnotarea semantică a conţinutului propune o descriere perceptuală la un nivel similar cu nivelul de percepţie uman. Informaţiile de nivel scăzut obţinute în urma analizei sintactice pot fi convertite în concepte lingvistice folosind informaţii "a priori" despre conţinutul secvenţei. Totuşi, obţinerea unei descrieri semantice de conţinut necesită înţelegerea completă a conţinutului secvenţei, astfel că pentru aceasta se preferă o abordare multimodală (imagine-sunet-text).

Un sistem semantic este definit generic ca fiind *orice sistem ce implică o colecţie de simboluri (vocabularul sistemului), reguli ce permit constituirea de propoziţii, reguli de desemnare şi reguli de validare.* În cazul sistemelor

[7] recunoaşterea automată a caracterelor reprezintă procesul mecanic sau electronic de traducere a imaginilor ce conţin scris de mână, scris de maşină sau text imprimat (de regulă rezultate în urma procesului de scanare) în text editabil de către calculator.

de indexare, termenul de "semantic" îşi conservă acest sens. Acesta se traduce prin *codarea interpretării datelor pentru a servi unei aplicaţii specifice* [Smeulders 00]. Astfel, sistemele de indexare semantică implică existenţa unui *set de simboluri* şi *reguli* ce permit interpretarea lingvistică a anumitor evenimente sau proprietăţi ale secvenţelor de imagini.

Adnotarea semantică a conţinutului a fost abordată pentru prima oară în sistemele de indexare a imaginilor, dar aceasta era greu de realizat deoarece proprietăţile semantice ale scenei sunt dificil de extras dintr-o simplă imagine statică. Datorită informaţiilor suplimentare furnizate de secvenţele de imagini (informaţia spaţio-temporală şi de mişcare), analiza semantică devine mai naturală în acest caz. De exemplu, dacă luăm cazul unei imagini ce surprinde un jucător de fotbal, singurele caracteristici ce reies din analiza imaginii sunt fizionomia acestuia şi prezenţa sa în scenă. Pe de altă parte, dacă dispunem de secvenţa ce îl surprinde pe jucător, putem determina dacă acesta va marca golul, modul în care acesta joacă, despre ce meci este vorba, etc., informaţii semantice esenţiale pentru înţelegerea conţinutului secvenţei.

Pentru a înţelege mai bine diferenţa dintre cele două categorii de adnotări, în Figura 1.3 am ilustrat un exemplu concret de adnotare sintactică şi respectiv semantică în cazul unei secvenţe de fotbal (axa orizontală reprezintă axa temporală, secvenţa este rezumată în doar câteva imagini reprezentative).

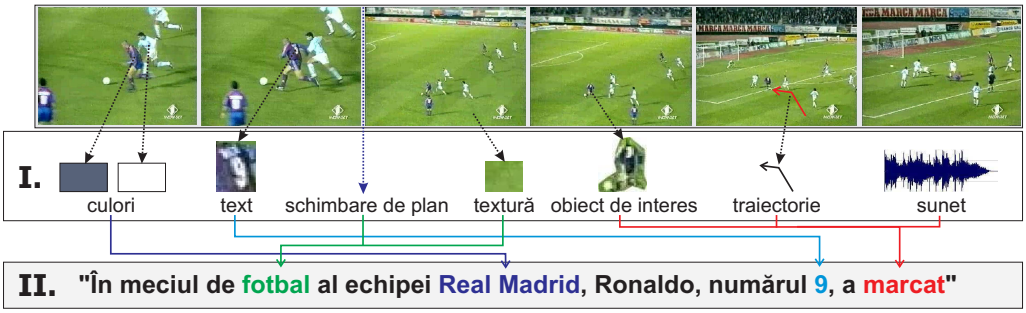


Figura 1.3: Diferenţa dintre adnotarea sintactică (punctul I.) şi semantică (punctul II.). Săgeţile colorate indică gradul de implicare al parametrilor de nivel scăzut în construirea descrierii semantice.

Astfel, în acest caz, adnotarea sintactică ne va furniza doar informaţii relative la scenă şi la proprietăţile acesteia, precum culoare, prezenţă text, textură, traiectoria obiectelor în mişcare, ritmul de desfăşurare al acţiunii, etc. Pe de altă parte, adnotarea semantică va da sens acestor informaţii: în mod ideal textura verde va indica că este vorba de un meci de fotbal, culorile jucătorilor (obiecte în mişcare) vor dezvălui echipele, recunoaşterea

numerelor de pe tricou va identifica jucătorii, segmentarea obiectului de interes, urmărirea acestuia şi prezenţa zgomotului specific vor indica marcarea golului. Punând cap la cap toate informaţiile, sistemul va "înţelege" că este vorba despre un meci de fotbal al echipei Real Madrid în care jucătorul cu numărul 9, Ronaldo, marchează.

Paradigma senzorială enunţată în cazul sistemelor de indexare a imaginilor este mai puţin pronunţată în cazul sistemelor de indexare a secvenţelor de imagini, acest lucru datorându-se în principal informaţiilor suplimentare ce facilitează înţelegerea conţinutului secvenţei. Cu toate acestea, paradigma semantică, de asemenea prezentă în sistemele de indexare a imaginilor, ia amploare în cazul secvenţelor de imagini datorită *lipsei de corelaţie dintre informaţia pe care o putem recupera din conţinutul datelor şi interpretarea care i-o atribuim* [Smeulders 00].

Astfel că, un sistem de indexare semantică eficient, trebuie să reunească următoarele trăsături importante [Naphade 02]:

- în primul rând este *capacitatea de analiză semantică* pe baza cererilor de căutare formulate de utilizator (vezi Secţiunea 1.4.4),
- un sistem eficient trebuie să fie *multimodal*, reunind şi armonizând metode de analiză ce folosesc diversele modalităţi ale secvenţei: imagine, text, sunet, etc.,
- relaţiile existente între atributele de nivel scăzut şi percepţia lor semantică trebuie *rezumate în mod eficient* pentru ca sistemul să fie capabil să ofere utilizatorului o descriere semantică coerentă.

Tendinţa actuală a sistemelor de indexare a secvenţelor de imagini către analiza semantică a fost motivată şi de atenţia acordată relativ noului standard de compresie video şi anume standardul MPEG-7[8] [Wang 00]. Noul standard video încearcă să introducă în procesul de codare, direct în fluxul de date, informaţii semantice referitoare la conţinutul secvenţei. Astfel, în momentul indexării, acestea vor putea fi recuperate direct din fluxul MPEG, eliminând astfel procesul de adnotare.

Pentru o descriere mai amănunţită a sistemelor de indexare semantică, cititorul se poate raporta la studiile prezentate în [Naphade 02] şi [Snoek 05b]. În cele ce urmează vom prezenta obiectivele sistemelor de indexare semantică precum şi dificultăţile impuse de analiza semantică a conţinutului.

[8]standardul MPEG-7 este un standard de descriere a conţinutului multimedia. Acesta foloseşte descrieri suplimentare ataşate conţinutului video clasic MPEG, pentru a facilita indexarea automată după conţinut. Standardul MPEG-7 este denumit formal şi Interfaţă de Descriere a Conţinutului Multimedia.

A. Obiectivele sistemelor de indexare semantică

Global, obiectivele sistemelor de indexare semantică a secvenţelor de imagini pot fi structurate pe patru direcţii de studiu [Naphade 02], şi anume:

- analiza **structurilor de nivel înalt** ale secvenţei,
- clasificarea după **gen**,
- analiza **dependentă de domeniul** de aplicaţie,
- analiza **independentă de domeniul** de aplicaţie.

Analiza structurilor de nivel înalt prezente în secvenţă presupune detecţia şi analiza diferitelor pasaje cu semnificaţie semantică, precum scenele de dialog, spoturile publicitare, diverse evenimente etc. Un exemplu este metoda propusă în [Hauptmann 98] unde frecvenţa schimbărilor de plan şi prezenţa imaginilor constante, negre, sunt folosite pentru a detecta pasajele publicitare în secvenţele de ştiri. O abordare similară, dar de această dată aplicată pentru detecţia pasajelor publicitare în filme, foloseşte viteza de schimbare a contururilor şi analiza amplitudinii vectorilor de mişcare pentru a detecta scenele de acţiune [Lienhart 97]. Un alt exemplu este metoda propusă în [Alatan 01] ce detectează scenele de dialog folosind analiza sunetului şi detecţia şi localizarea prezenţei feţelor în imagine. La cel mai jos nivel semantic de analiză se găsesc metodele de detecţie a evenimentelor de interes, precum metoda multimodală propusă în [Haering 00]. Aceasta se foloseşte de analiza de culoare, textură şi mişcare într-un clasificator pe bază de reţele neuronale, pentru a identifica pasajele de vânătoare în secvenţele documentare.

Clasificarea după gen. Un alt obiectiv al sistemelor de indexare semantică este clasarea secvenţelor după gen sau tip. Datorită interesului comercial prezentat de acestea, genurile cel mai des vizate sunt ştirile, reportajele sportive, filmele şi spoturile publicitare. Un exemplu este metoda propusă în [Truong 00a] ce foloseşte ca atribute durata planelor video, procentul de apariţie al diferitelor tranziţii video precum şi parametri de culoare, pentru a clasa secvenţele de imagini în genurile: animaţie, publicitate, videoclipuri, ştiri şi sport. O metodă generică, aplicabilă pentru orice tip de secvenţe, este propusă în [Kobla 00]. Aceasta foloseşte ca informaţie gradul de repetitivitate al mişcării, prezenţa textului şi prezenţa mişcărilor specifice ale camerei video sau de obiecte. Genul secvenţei este mai departe determinat folosind arbori de decizie. O altă abordare interesantă este prezentată în [Colombo 99] unde secvenţele publicitare sunt clasate în funcţie de percepţia conţinutului,

în patru sub-genuri: genul practic, critic, utopic şi animat. Informaţiile folosite pentru aceasta sunt: saturaţia culorilor, prezenţa în imagine a liniilor orizontale, mişcarea şi respectiv statistica tranziţiilor video.

Analiza dependentă de domeniul de aplicaţie. În acest caz, metodologia de analiză este adaptată domeniului de aplicaţie. Astfel, informaţiile sunt extrase pe baza expertizei domeniului vizat şi sunt specifice tipului de secvenţe analizate (vezi studiul bibliografic prezentat în [Snoek 05b]). Fiecare gen de secvenţă prezintă o serie de particularităţi ce pot face identificarea şi analiza conţinutului acestora mai uşoare. De exemplu, întotdeauna un desen animat va avea culori pastelate şi nenaturale, sau într-o secvenţă a unui meci de biliard, culoarea predominantă va fi culoarea verde. Dezavantajul acestui tip de abordare este dat de faptul că nu va putea fi aplicată în alte domenii decât cel vizat. Ca exemplu de analiză dependentă de domeniul de aplicaţie, putem menţiona analiza secvenţelor sportive de basket din [Saur 97], a secvenţelor medicale din [Fan 04] sau a filmelor de animaţie din [Ionescu 08].

Analiza independentă de domeniul de aplicaţie. Aceasta este direcţia de studiu cea mai dificilă abordată de sistemele de indexare semantică actuale. În prezent nu există multe sisteme care să fie independente de domeniul de aplicaţie. Cercetările existente încearcă să dezvolte metode de adnotare şi clasificare automată a conţinutului secvenţelor generice, fără a avea cunoştinţă de provenienţa acestora şi fără a folosi cunoştinţe "a priori" despre conţinutul acestora. Ca exemplu de astfel de sisteme putem menţiona folosirea de reprezentări probabiliste ale conţinutului multimedia propusă în [Naphade 01a], sau folosirea abordărilor de clasificare semantică [Qian 99]. Această direcţie rămâne totuşi o provocare pentru sistemele de indexare actuale, progresul ştiinţific actual din domeniu neputând oferi o soluţie definitivă.

B. Dificultăţile ridicate de analiza semantică

Sistemele de indexare semantică existente răspund mai mult sau mai puţin exigenţelor actuale de indexare a secvenţelor de imagini. Succesul acestora depinde în principal de modalitatea în care se depăşesc problemele impuse de adnotarea semantică. În cele ce urmează vom prezenta diversele puncte critice ale adnotării semantice [Fan 04].

Problema concordanţei. O primă dificultate a adnotării semantice este concordanţa dintre analiza de nivel scăzut şi descrierea semantică. Aceasta

depinde de eficienţa şi de calitatea parametrilor numerici folosiţi pentru inferenţa semantică. Pentru a putea face distincţia între diversele concepte semantice ce pot fi atribuite datelor analizate, diversitatea parametrilor numerici utilizaţi trebuie să fie suficient de vastă. Marea parte a sistemelor existente de indexare a secvenţelor de imagini folosesc ca informaţie de plecare pentru extragerea atributelor de conţinut, descompunerea secvenţei în plane video sau în obiecte semantice (de exemplu scene, pasaje de interes, etc.) [Fan 01]. Pe de altă parte, acest tip de atribute de nivel scăzut sunt dificil de asociat conceptelor semantice, în acest caz o implementare complet automată fiind greu realizabilă [Erol 00].

Problema modelării. O a doua dificultate a adnotării semantice este modelarea conceptelor semantice. Datorită paradigmei semantice (vezi Secţiunea 1.2), sistemele actuale nu sunt capabile să furnizeze un acces la baza de date care să fie în totalitate semantic. Diferite soluţii au fost totuşi propuse pentru a soluţiona sau reduce această problemă. Putem menţiona metodologia dezvoltată în [Adames 02] ce foloseşte informaţii "a priori" din domeniul de aplicaţie pentru a genera regulile perceptuale aferente descrierii semantice. De asemenea, o altă soluţie propusă este folosirea conceptului de "feedback" al sistemului pentru a exploata interacţia utilizator-sistem în vederea ameliorării performanţelor indexării [Cox 00]. Tot aici putem menţiona metodele de tip "machine learning"[9] ce exploatează corelaţiile ascunse ce există între datele multimodale [Barnard 03] (pentru o descriere detaliată a metodelor de clasificare supervizată, vezi Secţiunea 7.2).

Problema clasificării. Clasificarea semantică a datelor este de asemenea una dintre dificultăţile întâlnite în sistemele de indexare semantică. Metodele existente se împart în metode bazate pe sisteme de reguli constituite pe baza expertizei domeniului de aplicaţie [Alatan 01] şi metode statistice [Wang 01]. Prima categorie de metode se limitează la a folosi doar reguli perceptuale, fără a lua în calcul relaţiile existente între informaţiile furnizate de diversele modalităţi ale secvenţei. În contrast, metodele statistice permit exploatarea relaţiilor ascunse dintre date, dar performanţele acestora sunt dependente de eficacitatea parametrilor aleşi, precum şi de etapa de antrenare a clasificatorilor folosiţi.

Problema selecţiei. O altă dificultate este selecţia atributelor şi reduce-

[9] "machine learning" este un sub-domeniu al inteligenţei artificiale ce se ocupă cu proiectarea şi dezvoltarea de algoritmi şi tehnici ce permit calculatorului să simuleze procesul de învăţare.

rea dimensiunii spaţiului parametrilor. Intuitiv, putem spune că folosirea unui număr cât mai mare de atribute de nivel scăzut implică o mai bună putere de discriminare, şi prin urmare, o adnotare semantică mai eficientă. Totuşi, creşterea numărului de atribute implică creşterea radicală a timpului de antrenare al clasificatorilor, precum şi a redundaţei datelor. Astfel că este foarte important ca din "marea de atribute" disponibile să alegem doar pe acelea care sunt cele mai eficiente pentru indexare.

Problema organizării. Organizarea bazei de date şi accesul la conţinut sunt două aspecte ale sistemelor de indexare cel puţin la fel de importante ca cele enumerate anterior. Din păcate, domeniul bazelor de date şi cel al vederii asistate de calculator, nu interacţionează încă suficient pentru a propune structurarea bazei de date în funcţie de necesităţile de indexare semantică [Smeulders 00]. În cazul ideal, baza de date ar trebuii concepută încă din momentul constituirii în aşa fel încât să permită ca accesul la date să se realizeze în mod intuitiv, folosind criterii apropiate de modul de percepţie umană [Benitez 01].

1.4.3 Sistemul de navigare în baza de date

Accesul la conţinutul unei baze de secvenţe de imagini se realizează dificil, în primul rând datorită volumului important de informaţii conţinute de aceasta (număr foarte mare de imagini ce trebuiesc vizualizate). Vizualizarea în parte a fiecărei secvenţe este aproape imposibilă. O astfel de bază poate conţine milioane de secvenţe, astfel că timpul necesar vizualizării poate fi de ordinul anilor. Pentru a facilita accesul la informaţie, sistemele de indexare a secvenţelor de imagini[10] se folosesc de reprezentări compacte de conţinut (rezumate). Rezumatele sunt puse la dispoziţia utilizatorului prin intermediul unui pachet de utilitare software ce au ca rol principal să permită utilizatorului vizualizarea rapidă şi eficientă a conţinutului bazei de date. Acestea constituie **sistemul de navigare** în baza de date.

O primă modalitate de construcţie a rezumatelor constă în rezumarea conţinutului secvenţei pe baza structurii temporale (de exemplu, folosind descompunerea în plane). În acest caz, este posibilă construcţia a două tipuri diferite de rezumat, şi anume: *rezumatul în imagini* (static), care la bază este definit ca fiind o colecţie de imagini reprezentative pentru conţinutul secvenţei, şi *rezumatul în mişcare* (dinamic), ce este definit ca fiind o colecţie de pasaje reprezentative ale secvenţei (pentru o descriere detaliată a tehni-

[10] şi nu numai, aceasta fiind valabilă şi pentru sistemele de indexare video şi respectiv audio.

cilor de rezumare automată de conţinut, cititorul se poate raporta la Capitolul 5).

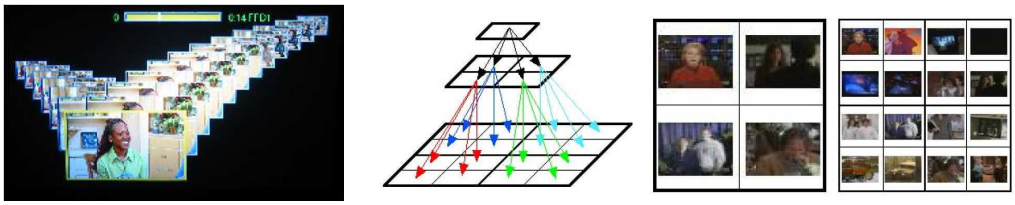
Rezumatele de conţinut permit utilizatorului să-şi facă rapid o idee globală asupra conţinutului secvenţei. Astfel, rezumatul static permite reprezentarea conţinutului vizual al secvenţei în doar câteva imagini, ce sunt uşor accesibile utilizatorului prin sistemul de navigare (timpul de vizualizare fiind neglijabil). Pe de altă parte, rezumatul dinamic aduce un plus de informaţie la nivelul acţiunii prezente în secvenţă, informaţie ce nu este disponibilă în rezumatul static. Totuşi, fiind el însuşi o secvenţă, în funcţie de nivelul de detaliu furnizat, timpul necesar vizualizării acestuia este mai ridicat decât în cazul rezumatului static, dar net inferior timpului de vizualizare integrală a secvenţei.

Ca exemple de sisteme de indexare a secvenţelor de imagini ce folosesc un sistem de navigare bazat pe rezumate, putem menţiona sistemul propus în [Zhu 05], unde conţinutul secvenţelor este rezumat folosind "imagini cheie" ("key-frames"[11]). Imaginile astfel reţinute, sunt prezentate utilizatorului sub forma unei broşuri. Un alt exemplu este sistemul propus în [Houten 03], unde conţinutul secvenţelor este prezentat sub formă de colecţii de fragmente ("patches"), unde fragmentele sunt definite ca fiind pasaje ale secvenţei de aceeaşi natură semantică (de exemplu scene de dialog, interviuri, etc.).

Un caz particular de rezumare a conţinutului o constituie vizualizarea imaginilor reprezentative ale secvenţei într-un spaţiu 3D, în care primele două dimensiuni sunt date de spaţiul imaginii, iar a treia dimensiune o constituie axa temporală a secvenţei. De exemplu, în [Vogl 99] secvenţele sunt rezumate cu serii temporale de "imagini cheie" pe care utilizatorul le poate vizualiza într-un mediu virtual interactiv. Un sistem similar, numit tunelul temporal ("time tunnel") este prezentat în [Electric 05]. "Imaginile cheie" ale secvenţei sunt vizualizate stratificat în funcţie de evoluţia temporală a cadrelor secvenţei (vezi Figura 1.4.a).

O altă modalitate de vizualizare compactă a conţinutului secvenţelor de imagini este pe baza reprezentării acesteia ca o structură ierarhică. De exemplu, sistemul propus în [Eidenberger 04] foloseşte un sistem de navigare interactivă constituit pe baza descriptorilor de conţinut furnizaţi de standardul video MPEG-7 [Jeannin 01]. Astfel, acesta pune la dispoziţia utilizatorului două modalităţi de reprezentare arborescentă a conţinutului secvenţei, şi anume: o reprezentare a conţinutului de plane video şi o reprezentare a structurii temporale. Pe fiecare nivel ierarhic, datele sunt structurate sub formă de hărţi cu auto-organizare ("Self-Organizing Maps"). Un alt exemplu de reprezentare ierarhică este cea folosită de sistemul ViBE [Chen 06]. În

[11] vezi explicaţia de la pagina 12.



Sistemul "Timetunnel" Sistemul ViBE

Figura 1.4: Exemple de modalităţi de vizualizare a conţinutului secvenţelor de imagini, folosite de sistemul de navigare în baza de date.

aceasta, planele video sunt vizualizate sub formă de structuri arborescente de "imagini cheie", ce sunt clasate într-o serie de categorii preudo-semantice determinate "a priori". Datele sunt prezentate utilizatorului sub formă piramidală, pe mai multe niveluri de detaliu (vezi Figura 1.4.b).

O categorie aparte de sisteme de navigare sunt sistemele disponibile "online" pe Internet. Acestea se folosesc de interfaţa grafică a programului de navigare sau "Internet browser". Ca exemple putem menţiona sistemul Vimix propus în [Yao 01] în care conţinutului secvenţelor este ierarhizat pe baza limbajului XML[12]. Un alt sistem este sistemul BIBS [Rowe 01] ce propune o organizare ierarhică liniară a secvenţei precum şi sincronizarea vizuală a pasajelor secvenţei cu adnotări textuale de conţinut.

1.4.4 Sistemul de căutare în baza de date

Scopul unui sistem de indexare după conţinut este de a furniza utilizatorului posibilitatea de *a căuta* şi *accesa* simplu şi eficient informaţiile prezente în baza de date. Pentru aceasta, sistemul foloseşte mai multe etape de prelucrare. În primă fază sunt create *adnotările de conţinut* ale datelor, ce permit gruparea acestora în funcţie de similaritatea conţinutului (vezi Secţiunea 1.4.1). În continuare, *accesul la baza* de date este efectuat cu ajutorul sistemului de navigare (vezi Secţiunea anterioară).

Datorită numărului important de informaţii disponibile în baza de date, aceste două etape nu sunt încă suficiente pentru a accesa în mod eficient conţinutul bazei. O ultimă cerinţă este posibilitatea de căutare a datelor dorite. Sistemul de căutare permite utilizatorului să localizeze datele prin

[12]XML este acronimul pentru "Extensible Markup Language" şi reprezintă un limbaj de uz general folosit în principal pentru a furniza informaţii suplimentare la conţinutul datelor prin crearea unui limbaj de tip "markup". Acesta este considerat ca fiind un limbaj extensibil, deoarece permite utilizatorului să-şi definească propriile elemente.

interogarea sistemului. Această interogare se face pe bază de cereri de căutare sau "queries". Pentru o căutare optimală şi eficientă, sistemul trebuie să permită formularea cererilor de căutare într-un limbaj intuitiv şi natural, pe de-o parte accesibil pentru utilizatorul neavizat, dar şi inteligibil pentru sistem.

În general, sistemul de căutare funcţionează în felul următor:

- **formularea cererii**: mai întâi utilizatorul concepe cererea de căutare. Aceasta poate fi formulată, fie pe baza unui exemplu a ceea ce caută, fie folosind o descriere textuală a conţinutului datelor căutate, sau pe baza unei descrieri grafice schematice a proprietăţilor datelor căutate.

- **conversia în descriptori sintactici sau semantici**: mai departe sistemul de căutare traduce cererea utilizatorului în atribute sintactice de nivel scăzut sau în atribute semantice de conţinut (în funcţie de tipul sistemului de indexare). Mecanismul folosit este similar cu cel folosit de sistem în etapa de adnotare a conţinutului bazei de date.

- **căutarea propriu-zisă**: căutarea se realizează prin compararea atributelor cererii de căutare cu cele deja stocate în baza de date. Folosind diverse măsuri de distanţă şi similaritate între atribute, sistemul va căuta datele ce sunt cele mai apropiate de criterile formulate (de exemplu, distanţa minimă în cazul parametrilor numerici). Rezultatele obţinute vor fi vizualizate în sistemul de navigare.

- **interacţia cu utilizatorul**: în mod opţional, sistemul poate interacţiona cu utilizatorul ("feedback") pentru a-şi ameliora algoritmii de căutare. În acest caz, utilizatorul este încurajat să-şi exprime gradul de satisfacţie cu privire la rezultatele furnizate de căutare. Aceste date vor servi ca exemple pentru antrenarea ulterioară a sistemului.

Calitatea unui sistem de căutare depinde de mai mulţi factori. Mai întâi este vorba de calitatea atributelor ce caracterizează datele şi de puterea discriminatorie a măsurilor de similaritate folosite pentru compararea acestora. Totuşi, rezultatele căutării sunt dependente în mare măsură de modul în care a fost formulată cererea de căutare. Dacă sistemul nu este capabil să înţeleagă criteriile utilizatorului, căutarea eşuează încă din faza incipientă.

Astfel, calitatea căutării este dependentă mai întâi de nivelul de cunoaştere de către utilizator a datelor căutate. În acest caz întâlnim mai multe situaţii posibile [Maillet 03]:

- utilizatorul ştie cu siguranţă că secvenţa căutată se află în baza de date. În acest caz, ţinta este unică iar utilizatorul va fi capabil să formuleze

eficient cererea de căutare. Căutarea se va repeta până în momentul în care secvenţa va fi găsită.

- utilizatorul caută o anumită secvenţă, dar nu este sigur că aceasta este prezentă în baza de date. În acest caz, sistemul de indexare trebuie să îi pună la dispoziţie algoritmi de căutare precişi şi eficienţi, pentru ca acesta să decidă rapid dacă secvenţa este cu adevărat prezentă în baza de date.

- utilizatorul caută o secvenţă folosind un exemplu sau pe baza unuia sau a mai multor evenimente de interes prezente în aceasta. În această situaţie, sistemul de indexare trebuie să îl ghideze pe utilizator pe tot parcursul căutării. Sistemul trebuie să interacţioneze cu utilizatorul pentru a filtra rezultatele obţinute ("feedback"). De asemenea, sistemul trebuie să prezinte eficient rezultatele căutării, de regulă folosind reprezentări şi descrieri compacte de conţinut, pentru ca utilizatorul să se decidă rapid dacă este vorba de secvenţa căutată, sau în caz contrar, să repete căutarea cu alte criterii.

În al doilea rând, calitatea căutării este condiţionată şi de modalitatea în care utilizatorul formulează cererea de căutare [Fan 04]:

- **folosirea unui exemplu**: în acest caz, cererea este formulată folosind un model al datelor [Tong 01]. De exemplu, utilizatorul caută toate secvenţele ce sunt asemănătoare cu o anumită secvenţă de care acesta dispune. Similaritatea va fi tradusă pe baza conţinutului multimodal al secvenţei în: similaritate de culoare, de tehnici de mişcare, a acţiunii, a obiectelor de interes prezente, etc. Această modalitate de căutare se dovedeşte a fi mai puţin eficientă în situaţia în care utilizatorul nu cunoaşte domeniul respectiv, caz în care acesta nu va fi capabil să indice un bun exemplu pentru căutare.

- **folosirea reprezentărilor textuale**: în acest caz, cererea de căutare este formulată prin sintetizarea caracteristicilor datelor dorite sub formă textuală [Benitez 01]. Acest tip de formulare a căutării, fiind textuală, este apropiată de modalitatea de percepţie umană. De exemplu, utilizatorul caută toate "serialele TV", sau la un nivel semantic superior toate "filmele dramatice". Principalul inconvenient la acest tip de căutare este lipsa de sens a anumitor adnotări textuale automate, generate de sistemul de indexare, ce pot conduce la rezultate eronate ale căutării.

- **folosirea sistemului de navigare**: în acest caz, utilizatorul poate folosi direct sistemul de navigare în baza de date pentru a efectua căutarea [Smith 99]. Acest tip de căutare este adecvat utilizatorilor neavizaţi, ce nu au cunoştinţă asupra conţinutului bazei de date şi nici asupra criterilor de căutare. Totuşi, principala constrângere a acestei modalităţi de căutare o constituie faptul că în general, datele din baza de date nu sunt structurate pe criterii semantice de conţinut. Astfel că timpul de localizare a informaţiei dorite poate deveni important.

În cele ce urmează vom prezenta câteva dintre sistemele de căutare folosite de sistemele de indexare a secvenţelor de imagini actuale. Marea parte a acestor sisteme, analizează proprietăţile conţinutului secvenţelor folosindu-se de relaţiile temporale existente între atributele de conţinut, relaţii puse în evidenţă pentru prima oară în [Allen 83]. Ca exemple de astfel de sisteme putem menţiona sistemul SMOOTH [Kosch 01], GOALGLE şi News RePortal [Snoek 05a], în care căutarea este efectuată pe bază de criterii semantice sau temporale (vezi Figura 1.5).

Sistemul "Goalgle" Sistemul "News RePortal"

Figura 1.5: Exemple de motoare de căutare: sistemul "Goalgle" de căutare a scenelor de gol în secvenţele de fotbal şi sistemul "News RePortal" de căutare a ştirilor.

Un alt exemplu este sistemul SoccerQ [Chen 05] ce permite căutarea după criterii semantice a secvenţelor sportive în funcţie de prezenţa anumitor evenimente de interes. Căutarea este realizată pe trei niveluri structurale: la nivel de secvenţă, la nivel de segment şi la nivel de variabilă (prin variabilă înţelegând proprietăţi ale obiectelor, ca de exemplu numele echipelor). Cererile de căutare sunt formulate într-un limbaj natural, precum: "caută toate secvenţele ce conţin offside". Alte abordări folosesc predefinirea limbajului de formulare a cererii de căutare, precum sistemul pro-

pus în [Donderler 04]. O abordare diferită este prezentată în [Liu 02a], unde relaţiile spaţiale şi temporale existente între obiectele din scenă sunt modelate prin relaţii textuale între simboluri, folosind şirurile 3D ("3D-strings"). Astfel, problema căutării este transformată într-o problemă de similaritate între simboluri textuale.

În concluzie, sistemele de căutare în baza de date sunt dependente de domeniul de aplicaţie (sport, film, etc.), fiind adaptate acestuia. Criterile de căutare sunt construite pe baza expertizei domeniului respectiv. O posibilă soluţie pentru constituirea unui sistem general valabil, aplicabil în cazul oricărei categorii de secvenţe de imagini, constă în reunirea "experienţei" sistemelor de căutare dezvoltate pentru fiecare domeniu aplicativ în parte, într-un sistem global. În acest fel, se va profita de metodologiile de căutare cele mai performante ale fiecărei categorii de secvenţe.

1.5 Sistemele de indexare video

Sistemele de indexare video după conţinut sau CBVRS ("Content-Based Video Retrieval Systems") sunt extensia naturală a sistemelor de indexare a secvenţelor de imagini şi a sistemelor de indexare a sunetului, deoarece datele prelucrate sunt în acest caz *documente audio-vizuale*. Un document video este definit ca fiind o secvenţă de imagini care conţine şi informaţie audio (coloana sonoră). Din această cauză, frecvent în literatura de specialitate, cuvântul video este utilizat în mod abuziv pentru a desemna secvenţele de imagini. Pe de altă parte, definirea unei frontiere precise între cele două tipuri de sisteme, de indexare de secvenţe şi respectiv de indexare video, nu este întotdeauna un lucru uşor, deoarece deseori acestea se întrepătrund. În cele ce urmează, ne vom referi prin sistem de indexare video la orice sistem de indexare ce ia în calcul imaginea şi sunetul.

În absenţa informaţiei audio, sistemele de indexare a secvenţelor de imagini sunt un caz particular al sistemelor de indexare video. Astfel, toate metodele utilizate de acestea sunt aplicabile şi în cazul sistemelor de indexare video. Descrierea sistemelor de indexare video din acest subcapitol nu va relua în discuţie metodele deja prezentate în Sectiunea 1.4, în cadrul sistemelor de indexare a secvenţelor de imagini, ci ne vom focaliza atenţia doar asupra metodelor specifice prelucrării informaţiei audio-vizuale.

Majoritatea sistemelor de indexare video nu folosesc abordări cu adevărat multimodale, în care informaţia vizuală şi audio este prelucrată simultan, în strânsă corelaţie. În realitate, cele două modalităţi sunt mai întâi prelucrate în mod separat, iar rezultatele sunt apoi fuzionate pentru a obţine descrierea conţinutului audio-vizual [Naphade 02]. De exemplu, analiza de imagine

poate fi folosită pentru a realiza segmentarea în plane a secvenţei iar sunetul poate fi folosit ulterior pentru a cataloga conţinutul video, cum este cazul sistemului propus în [Wang 00].

Astfel, sistemele ce folosesc integrarea analizei multimodale, imagine-sunet, sunt de două tipuri [Snoek 05a]:

- sisteme ce folosesc o abordare pe *bază de reguli* definite prin expertiza domeniului de aplicaţie [Babaguchi 02],
- sisteme ce folosesc *abordări statistice* pe bază de algoritmi de clasificare a datelor [Han 02b].

Prima categorie de sisteme analizează independent fiecare modalitate video, iar rezultatele sunt fuzionate la final folosind o clasificare pe bază de reguli. În această categorie putem include sistemul propus în [Babaguchi 02], în care sunetul este folosit mai întâi pentru a identifica o serie de cuvinte specifice anumitor evenimente de interes din înregistrările meciurilor de fotbal (de exemplu: momentul golului, aclamaţiile publicului, etc.), iar informaţia vizuală este folosită pentru catalogarea conţinutului scenei. Pentru a doua categorie de sisteme, printre abordările statistice cel mai des folosite, putem menţiona: metodele ce folosesc reţele Bayes dinamice [Naphade 01b], arbori de decizie [Zhou 02] sau clasificări de tip "Support Vector Machines" [Lin 02] (vezi Secţiunea 7.2.2).

Analiza multimodală folosită de sistemele de indexare video necesită metode eficiente de fuziune între diferitele surse de informaţie. Dificultatea unei astfel de prelucrări constă în următoarele aspecte:

- **sincronizarea datelor**: este necesară pentru a omogeniza informaţiile provenite de la diferitele surse de informaţie. Soluţia cea mai frecvent adoptată constă în conversia tuturor datelor la un sistem unic de referinţă, cum ar fi axa temporală. Datele sunt astfel sincronizate în funcţie de momentul temporal de producere al acestora [Snoek 05b].
- **alegerea modelului adecvat**: constă în introducerea de informaţii suplimentare, ce nu sunt disponibile în momentul exact al producerii evenimentului semantic analizat, acestea fiind prelevate din documentul video, fie dinaintea acestuia (sens negativ al timpului) sau de după producerea lui (sens pozitiv al timpului).
- **redundanţa parametrilor folosiţi**: adnotarea multimodală a conţinutului foloseşte o serie de parametri de conţinut ce sunt calculaţi pentru fiecare modalitate a documentului video. Problema care apare este corelaţia puternică dintre aceştia, astfel că o etapă de decorelare şi selecţie este strict necesară [Wang 00].

În concluzie, sistemele actuale de indexare a documentelor video folosesc fuziunea diferitelor tipuri de atribute extrase la nivel de imagine, sunet şi text încrustat în imagine. Scopul final este adnotarea conţinutului în vederea accesării ulterioare a acestuia la un nivel semantic apropiat de percepţia umană.

1.6 Concluzii

În acest capitol am prezentat conceptul de indexare după conţinut a datelor multimedia. Astfel am făcut o trecere în revistă a metodelor şi tehnicilor folosite în sistemele actuale de indexare a imaginilor (CBIR), a sunetului (CBAR), a secvenţelor de imagini (CBIRS) precum şi a documentelor video (CBVR).

Global, analizând tendinţele cercetării din acest domeniu, putem spune că există două direcţii distincte. Pe de-o parte sunt sistemele ce folosesc **analiza sintactică a conţinutului**. În acest caz, adnotarea bazei de date este realizată cu descriptori statistici de nivel scăzut ce sunt calculaţi folosind informaţii precum: culoare, formă, textură, sunet, mişcare, text, etc. Aceşti descriptori sunt de regulă măsuri numerice complexe, dificil accesibile pentru un utilizator neavizat în domeniu.

Pe de altă parte, întâlnim sistemele ce folosesc **analiza semantică**. Acestea propun descriptori perceptuali ai conţinutului datelor, de regulă obţinuţi pe baza descrierilor sintactice, prin adăugarea expertizei manuale a domeniului vizat. Descriptorii semantici sunt exprimaţi cu ajutorul unui dicţionar de simboluri similar vocabularului folosit de limbajul uman.

Sistemele actuale de indexare după conţinut tind să evolueze exclusiv spre *descrierea semantică automată* a conţinutului datelor în încercarea de simplificare a problematicii accesării bazelor de date multimedia. Totuşi pentru a atinge acest nivel de descriere a datelor, trebuiesc depăşite o serie de probleme, de la problema selecţiei atributelor cele mai reprezentative până la paradigma semantică (lipsa de corespondenţă între parametrii matematici de care dispunem şi interpretarea lor semantică).

Sistemele actuale nu au reuşit încă să adopte o soluţie definitivă pentru aceste probleme, soluţia provizorie fiind simplificarea prelucrării prin adaptarea metodelor la domeniul de aplicaţie. Majoritatea sistemelor de indexare semantică folosesc astfel informaţii "a priori" despre conţinutul datelor, fapt ce a dus la adaptarea metodelor la diversele tipuri de date.

CAPITOLUL 2

Segmentarea temporală

Rezumat: *Structura temporală a unei secvenţe de imagini este similară cu modul în care este structurată o carte, unde diversele capitole sunt înlănţuite pentru a constitui naraţiunea. În acest capitol vom discuta problematica descompunerii temporale a secvenţei în plane video, descompunere ce permite înţelegerea structurală a conţinutului acesteia. Astfel, vom analiza diversele metode de detecţie a tranziţiilor video, atât abrupte, cât şi graduale. De asemenea, vom prezenta şi problematica descompunerii secvenţei în unităţi structurale de nivel semantic superior planelor, precum scenele video. Dacă planele video pot fi considerate ca fiind unităţile sintactice de bază ale secvenţei, atunci scenele pot fi văzute ca fiind unităţile semantice, acestea permiţând o înţelegere mai profundă a conţinutului secvenţei.*

Segmentarea temporală a unei secvenţe de imagini este definită ca fiind procesul de divizare al acesteia în *unităţi structurale* de bază, numite şi **plane video**. Fiind o etapă premergătoare înţelegerii structurale a secvenţei, marea majoritate a tehnicilor existente de analiză a conţinutului secvenţelor de imagini, folosesc ca punct de plecare segmentarea în plane [Lienhart 01b].

Procesul de decupare în plane poate fi văzut şi din prisma procesului de editare al secvenţei ce are loc în studio, la momentul montajului. Planele video sunt concatenate folosind diverse tehnici specifice sau efecte speciale (de exemplu tranziţii video) pentru a da naştere secvenţei finale, proces ce este

numit în literatura de specialitate şi "final cut". În acest sens, segmentarea temporală poate fi percepută ca fiind procesul invers editării secvenţei ce are loc în studio.

2.1 Structura temporală a unei secvenţe

Din punct de vedere al structurii temporale, o secvenţă de imagini poate fi reprezentată pe mai multe niveluri ierarhice. Acestea sunt ilustrate în Figura 2.1, astfel:

- **nivelul imagine**: reprezintă nivelul structural cu gradul de granularitate cel mai mare (cel mai detaliat nivel) şi este reprezentat de toate imaginile conţinute în secvenţă.
- **nivelul planelor video**: corespunde imaginilor secvenţei ce au fost filmate între două porniri consecutive ale camerei video. Secvenţa de imagini astfel obţinută are proprietatea de continuitate vizuală (vezi [Corridoni 95]).

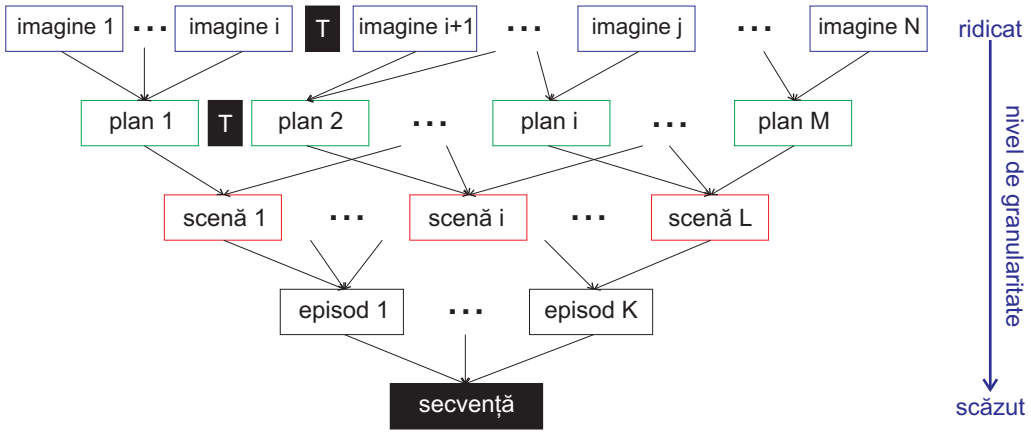


Figura 2.1: Structura ierarhică a unei secvenţe de imagini (T reprezintă o tranziţie video).

- **nivelul scenelor**: corespunde grupurilor de plane video ce sunt corelate din punct de vedere al conţinutului semantic. Acestea trebuie să respecte regula celor trei unităţi: unitate de loc, unitate de timp şi unitate de acţiune [Corridoni 95].

- **nivelul episoadelor**: corespunde grupurilor de scene ce sunt similare din punct de vedere al acţiunii globale (de exemplu, episoadele unei serii TV) [Bimbo 99].

- **nivelul secvenţei**: este nivelul structural cu gradul de granularitate cel mai mic şi este reprezentat de secvenţa însăşi.

Marea parte a metodelor de analiză a secvenţelor de imagini prelucrează secvenţa la nivel de plan video. Celelalte niveluri ierarhice, precum scenele sau episoadele, sunt folosite cu predilecţie de sistemele de indexare semantică, deoarece detecţia acestora presupune o analiză perceptuală de conţinut.

Într-o secvenţă, planele video sunt concatenate pe baza **tranziţiilor video** (vezi Figura 2.1). O tranziţie video este un efect vizual folosit pentru a lega imaginea de sfârşit a unui plan, de imaginea de început a planului următor. În funcţie de tipul transformărilor 2D aplicate imaginilor, tranziţiile video existente se împart în cinci clase:

- **clasa de identitate**: tranziţiile din această categorie nu modifică imaginile planelor video şi nici nu adaugă imagini suplimentare (vezi [Lienhart 01b]). În această categorie se află doar tranzitiile de tip "cut", numite şi tranziţii abrupte. Un "cut" produce o discontinuitate vizuală în secvenţă, deoarece planele vecine sunt alipite în mod direct (vezi Figura 2.2).

- **clasa spaţială**: din această categorie fac parte tranziţiile ce aplică imaginilor planelor transformări spaţiale [Lienhart 01b] (vezi Figura 2.2). Ca exemple putem menţiona efectele de tip "wipes" în care o imagine este înlocuită progresiv de o alta folosind o margine de o anumită formă, efectele de tip "mattes" care de regulă sunt folosite pentru a combina imaginea din planul principal cu imaginea de fundal sau efectele de tip "page turns" în care noua imagine este descoperită simulând răsfoirea paginii unei carţi.

- **clasa cromatică**: în acest caz, imaginile planelor video sunt modificate prin transformări de culoare [Lienhart 01b]. Ca exemple putem menţiona tranziţiile de tip "fade" şi "dissolve" (vezi Figura 2.2). Un "fade" este o tranziţie ce permite, fie dizolvarea progresivă a unei anumite imagini într-o imagine constantă, de regulă neagră, ceea ce numim "fade-out", fie apariţia progresivă a unei imagini dintr-o imagine constantă, proces numit "fade-in". O tranziţie de tip "dissolve" este în general definită de superpoziţia unui efect "fade-out" peste un efect "fade-in", suprapunere ce are ca efect vizual dizolvarea unei imagini în alta.

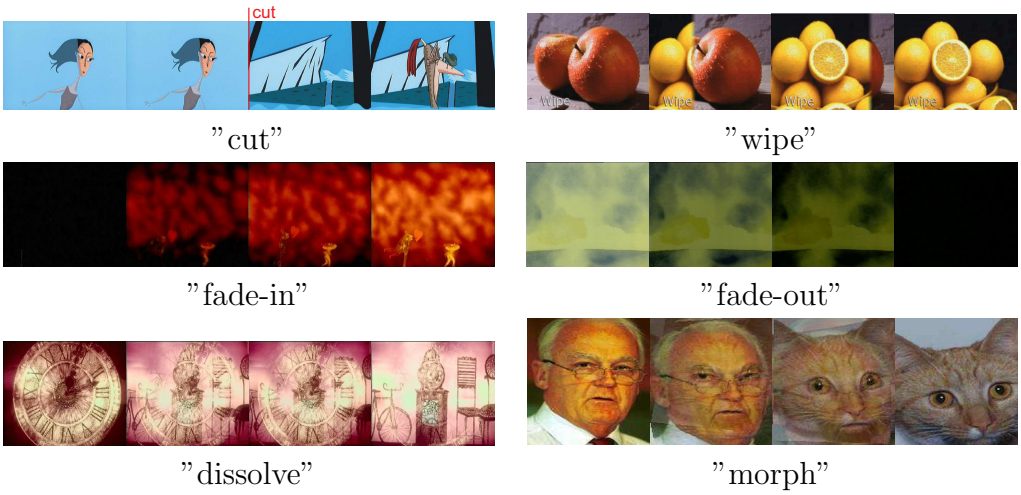


Figura 2.2: Exemple de tranziţii video (pentru fiecare tranziţie au fost prezentate doar câteva imagini reprezentative, axa orizontală reprezentând axa temporală, sursă imagini [Folimage 06] [Wikipedia 08] [Morphing 08]).

- **clasa spaţio-cromatică**: tranziţiile video din această categorie sunt o combinaţie a clasei spaţiale şi cromatice, imaginile planelor fiind modificate atât prin transformări spaţiale, cât şi cromatice [Lienhart 01b]. În această categorie se regăsesc toate efectele de tip "morphing"[1] (vezi Figura 2.2). Cu toate acestea, anumite transformări din clasa cromatică pot fi încadrate şi în această categorie, un exemplu fiind transformările de tip "dissolve" ce înglobează mişcări ale camerei video.

- **clasa temporală**: reprezintă o categorie aparte de tranziţii video. În anumite situaţii, tranziţia de la un plan video la altul se face temporal folosind o mişcare 3D a camerei video (vezi Secţiunea 3.2). De exemplu, camera video filmează un obiect de interes, iar apoi se translatează şi se focalizează pe un punct de interes îndepărtat, din fundalul imaginii. Astfel, anumite mişcări 3D ale camerei video, cu toate că nu sunt tranziţii video propriu-zise în sensul definiţiei enunţate anterior, au rolul de a face legătura între două momente distincte ale secvenţei (două plane diferite), putând astfel fi considerate drept tranziţii.

Din punct de vedere al duratei, tranziţiile video se împart în două categorii, astfel întâlnim *tranziţii abrupte* sau "cuts" şi *tranziţii graduale*, precum

[1] "morphing" este un efect special ce presupune metamorfozarea unei imagini în alta prin tranziţii uniforme şi constante.

"fade", "dissolve", "matte" etc. Dintre toate tranziţiile existente, cel mai frecvent folosite sunt tranziţiile abrupte de tip "cut", deoarece sunt simplu de utilizat şi nu introduc întârzieri în derularea conţinutului secvenţei. Cu o frecvenţă de apariţie cu cel puţin un ordin de măsură mai mic sunt tranziţiile graduale, dintre care, cele mai întâlnite sunt de tip "fade" şi "dissolve".

Folosirea unui anumit tip de tranziţie pentru a face legătura între două plane nu este aleatorie. Fiecare tranziţie are o semnificaţie semantică ce este adaptată conţinutului secvenţei. De exemplu, folosirea frecventă a tranziţiilor de tip "cut" are ca efect creşterea dinamismul secvenţei [Colombo 99]. Pe de altă parte, tranziţiile de tip "dissolve" şi "fade" sunt folosite frecvent pentru a schimba timpul sau locul acţiunii [Lienhart 01b], în timp ce folosirea unei tranziţii "fade-out" urmată de o tranziţie "fade-in", introduce un moment de pauză în derularea secvenţei şi este folosită de regulă pentru a trece la un alt capitol al acţiunii.

2.2 Descompunerea în plane video

În cele ce urmează, ne vom limita la trecerea în revistă a algoritmilor şi tehnicilor de detecţie a tranziţiilor cel mai frecvent analizate de sistemele de indexare semantică a secvenţelor de imagini. Astfel, vom prezenta detecţia tranziţiilor de tip "cut", "fade-in", "fade-out" şi respectiv "dissolve".

Pentru mai multe informaţii cu privire la algoritmii folosiţi pentru detectarea altor tipuri de tranziţii video, cititorul se poate raporta la lucrările [Bimbo 99] ("wipes" şi "mattes"), [Song 02] ("wipes"), [Ren 03] (tranziţii temporale şi alte tipuri de tranziţii) sau [Hanjalic 02] (abordare generică a problematicii detecţiei tranziţiilor video folosind metode statistice).

2.2.1 Detecţia de "cuts"

După cum am precizat şi în paragrafele anterioare, tranziţiile de tip "cut" sunt tranziţiile cel mai frecvent folosite în secvenţele de imagini. Acestea sunt definite ca fiind concatenarea directă a două plane video adiacente din punct de vedere temporal, $P_1(x, y, t)$ şi $P_2(x, y, t)$, unde (x, y) reprezintă coordonatele spaţiale ale imaginii iar t coordonata temporală.

Din punct de vedere matematic, secvenţa ce rezultă în urma concatenării, $S(x, y, t)$, este dată de relaţia următoare [Lienhart 01b]:

$$S(x, y, t) = (1 - u(t - t_{cut})) \cdot P_1(x, y, t) + u(t - t_{cut}) \cdot P_2(x, y, t) \qquad (2.1)$$

unde t_{cut} reprezintă momentul de timp ce corespunde primei imagini de după

tranziţia "cut" iar $u()$ este funcţia treaptă unitate:

$$u(t) = \begin{cases} 1 & \text{dacă } t \geq 0 \\ 0 & \text{altfel} \end{cases} \tag{2.2}$$

Definit în acest fel, un "cut" are proprietatea de a introduce o *discontinuitate vizuală* în fluxul secvenţei. Metodele existente de detecţie folosesc diverse abordări pentru a măsura tocmai această discontinuitate. Cu toate acestea, în general, se pot identifica anumite etape comune de prelucrare.

O primă etapă constă în parametrizarea cadrelor secvenţei prin extragerea de *parametri specifici* tranziţiilor de tip "cut", ca de exemplu histograme de culoare, medii spaţiale, etc. Mai departe, variaţia temporală a parametrilor extraşi este evaluată între imaginile la momentele k şi $k + l$, unde $k = 0, ..., N_{sec}$ reprezintă indicele cadrului curent, N_{sec} este numărul total de imagini al secvenţei iar $l \geq 1$ reprezintă pasul temporal de analiză[2]. Pentru aceasta se folosesc *măsuri de distanţă*, sau mai general *măsuri de similaritate*.

Valorile de discontinuitate astfel obţinute sunt folosite pentru a localiza tranziţia prin compararea acestora cu un anumit prag T, operaţie numită şi "thresholding". Dacă valoarea discontinuităţii se dovedeşte a fi superioară pragului T, atunci este foarte probabil ca un "cut" să fi avut loc între imaginile k şi $k + l$.

Principalele puncte critice ale unei astfel de abordări au fost bine evidenţiate în [Hanjalic 02], astfel:

- o primă problemă este **puterea discriminatorie a parametrilor**. Performanţa algoritmului de detecţie este total dependentă de puterea de discriminare a parametrilor aleşi. În cazul în care parametrii nu sunt reprezentativi pentru conţinut, atunci valorile funcţiei de discontinuitate pentru tranziţia de tip "cut" vor fi similare cu cele obţinute pentru celelalte cadre ale secvenţei. Separarea acestora va fi astfel imposibilă.

- o a doua problemă este alegerea **măsurii de distanţă**. Aceasta trebuie aleasă astfel încât să furnizeze valori neglijabile pentru imaginile similare, imagini ce aparţin aceluiaşi plan, şi respectiv valori importante pentru imaginile foarte diferite ce separă o tranziţie de tip "cut".

- în ultimul rând este dificultatea alegerii **pragului de detecţie**. Astfel, alegerea unui prag T prea mic va duce la mărirea numărului de false detecţii, deoarece pe lângă valorile de discontinuitate specifice unui

[2]metodele de analiză şi prelucrare a secvenţelor de imagini se raportează de regulă la numărul total de cadre al secvenţei şi mai puţin la durata totală a acesteia. Astfel că, indicele unei imagini va reprezenta numărul cadrului în secvenţă şi nu indicele temporal.

"cut", va lăsa să treacă şi alte astfel de variaţii mai mici datorate altor factori (de exemplu, mişcării camerei video, mişcării de obiecte). Pe de altă parte, un prag prea ridicat va mări numărul de tranziţii ce vor trece nedetectate.

O problemă stringentă ce este responsabilă pentru marea parte a falselor detecţii, este disimilaritatea imaginilor aceluiaşi plan video. Aceasta este cauzată în principal de trei situaţii: mişcarea camerei video, mişcarea obiectelor din scenă sau fluctuaţiile intensităţii luminoase a imaginii. Pentru a compensa aceste situaţii, o soluţie constă în folosirea şi a altor surse de informaţii, pe lângă valorile de discontinuitate. De exemplu, compensarea mişcării poate fi folosită pentru a anula efectul mişcării globale a camerei video [Marichal 98], divizarea imaginii în mai multe regiuni şi medierea funcţiei de discontinuitate în acestea poate fi folosită pentru a reduce influenţa mişcării obiectelor [Ionescu 06a], detecţia bliţului aparatului foto poate fi folosită pentru a elimina falsele detecţii provocate de acesta [Heng 99] (îndeosebi în secvenţele de ştiri), etc.

În literatura de specialitate există o vastă diversitate de metode de detecţie de "cut" ce ameliorează sau corectează influenţa factorilor enumeraţi mai sus, precum metodele propuse în [Bimbo 99], [Lienhart 01b], [Fernando 01], [Hanjalic 02] sau [Ren 03]. În funcţie de natura parametrilor folosiţi pentru a măsura discontinuitatea vizuală specifică acestui tip de tranziţie, metodele existente se împart în cinci categorii principale, astfel:

- metode bazate pe *analiza intensităţii pixelilor* din imagine,
- metode bazate pe *analiza contururilor* din imagine,
- metode bazate pe *analiza de mişcare*,
- metode ce analizează conţinutul secvenţei în formatul original al datelor video, cum ar fi *fluxul comprimat* MPEG.
- alte metode ce folosesc *alte surse de informaţie*.

În cele ce urmează, vom face o trecere în revistă a tehnicilor şi algoritmilor de detecţie folosiţi de metodele din fiecare dintre cele cinci categorii enumerate.

Metode bazate pe analiza intensităţii pixelilor

Metoda cea mai simplă de evaluare a discontinuităţii vizuale produsă de o tranziţie de tip "cut", constă în folosirea diferenţei, la nivel de intensitate a pixelilor, dintre imaginile consecutive la momentele k şi $k + l$, unde l

reprezintă pasul de analiză. De exemplu, în [Otsuji 91] un "cut" este detectat în momentul în care numărul de pixeli, N_{pixels}, ce se schimbă de la o imagine la alta depăşeşte un anumit prag T, astfel $N_{pixels} \geq T$, unde N_{pixels} este dat de relaţia următoare:

$$N_{pixels} = \frac{1}{NM} \sum_{x=1}^{X} \sum_{y=1}^{Y} D_{k,k+l}(x, y) \tag{2.3}$$

unde $X \cdot Y$ reprezintă dimensiunea imaginii iar $D_{k,k+l}(x, y)$ este definit ca fiind:

$$D_{k,k+l}(x, y) = \begin{cases} 1 & \text{dacă } |I_k(x, y) - I_{k+l}(x, y)| > T_1 \\ 0 & \text{altfel} \end{cases} \tag{2.4}$$

unde $I_k(x, y)$ reprezintă imaginea la momentul k iar T_1 este un alt prag estimat în funcţie de gradul de schimbare al intensităţii la nivel de pixel.

Principala problemă a acestui gen de abordare constă în sensibilitatea preciziei detecţiei la prezenţa zgomotului în imagine sau a mişcărilor globale ale camerei video, ce au ca efect obţinerea de valori semnificative pentru N_{pixels}. Dintre metodele ce folosesc tehnici similare, putem enumera:

- [Zhang 93] ce propune reducerea zgomotului prin filtrarea mediană a imaginilor înaintea efectuării diferenţei dintre acestea,
- [Boreczky 98] ce clasifică distanţele dintre pixeli folosind modele Markov ascunse,
- [Kobla 99] ce foloseşte ca măsură de disimilaritate distanţa Euclidiană calculată în spaţiul de culoare YUV şi RGB.

O altă cauză ce influenţează valoarea distanţei dintre imagini, sunt transformările geometrice ce pot surveni în imagine. Pentru a diminua influenţa acestora, unele metode de detecţie apelează la calculul histogramelor de intensitate a pixelilor. Acestea sunt calculate, fie pe imaginea de nivele de gri, fie folosind informaţia de culoare.

Metoda cea mai des întâlnită constă în măsurarea discontinuităţii vizuale produse între imaginile la momentele k şi $k+l$ pe baza sumei distanţei dintre binii[3] histogramelor celor două imagini [Yeo 95]. Pentru a reduce influenţa schimbărilor de intesitate luminoasă ce pot surveni în imagine, [Furht 95] propune folosirea de histograme calculate în spaţiul de culoare HSV (H-nuanţă, S-saturaţie şi V-intensitate). Astfel, intensitatea luminoasă este separată de

[3]histogramele sunt funcţii ce contabilizează valorile din anumite grupuri de valori. Acestea sunt numite şi clase, dar în contextul histogramei sunt cunoscute sub numele de "bini" (containeri ce acumulează valori la aceeaşi rată cu frecvenţa clasei).

informaţia de culoare. [Arman 93a] propune calcularea histogramelor folosind doar componentele H şi S ce formează suprafaţa 2D, HC. Aceasta este folosită mai departe la estimarea discontinuităţii vizuale astfel:

$$D(k,k+l)=\sum_{x=1}^{X}\sum_{y=1}^{Y}|d_{k,k+l}(x,y)|\times\Delta_H\times\Delta_C \tag{2.5}$$

unde $d_{k,k+l}(x,y)$ reprezintă diferenţa dintre binii de coordonate (x,y) (aparţinând suprafeţei HC) pentru imaginile la momentele k şi $k+l$ iar Δ_H, Δ_C sunt paşii de cuantizare a componentelor H şi S.

Alte abordări bazate pe calculul histogramelor încearcă să îmbunătăţească invarianţa detecţiei folosind diverse măsuri de distanţă ce sunt calculate în spaţii de culoare precum: HSV, YIQ, Lab, Luv, etc. [Lienhart 01b] (un studiu detaliat al spaţiilor de culoare existente este prezentat în Secţiunea 4.1). De exemplu, [Shen 97] propune folosirea distanţei Hausdorff[4] multinivel pentru a calcula distanţa între histograme, [Drew 00] propune ca măsură de similaritate intersecţia histogramelor calculată folosind distanţa dintre culorile proiectate în spaţiul CbCr şi rb, [Kim 02] propune calculul histogramelor în spaţiul de culoare YUV, [Ma 01] foloseşte intersecţia între histograme şi diferenţa între culorile medii ale blocurilor de pixeli, etc.

Pentru a reduce influenţa mişcării obiectelor sau a zgomotului prezent în imagine, alte metode calculează histogramele la nivel de bloc de pixeli şi nu pentru întreaga imagine. În [Nagasaka 92], imaginile la momentele k şi $k+l$ sunt divizate în 16 blocuri de pixeli iar histogramele, $H_{k,i}$, şi respectiv $H_{k+l,i}$, sunt calculate pentru blocurile de pixeli $b_i(k)$ şi $b_i(k+l)$, unde $i=1,...,16$ reprezintă indicele blocului. Pentru a compara histogramele astfel obţinute se foloseşte testul χ^2, astfel:

$$D(i)=\sum_{j=0}^{63}\frac{[H_{k,i}(j)-H_{k+l,i}(j)]^2}{H_{k+l,i}(j)} \tag{2.6}$$

unde j reprezintă indicele unui bin al histogramei, în acest caz $j=0,...,63$.

Un alt exemplu este metoda propusă în [Ionescu 07c], unde imaginile sunt divizate în patru regiuni egale (cadrane) în scopul reducerii influenţei mişcării obiectelor care intră sau ies din scenă, asupra valorilor histogramelor. Pentru fiecare cadran al imagini este calculată o histogramă color. Funcţia de discontinuitate vizuală este calculată ca fiind media distanţelor Euclidiene obţinute între cele patru histograme ale imaginilor la momentele k şi respectiv

[4]vezi explicaţia de la pagina 172.

$k+2$, astfel:

$$D(k,k+2) = \frac{1}{4}\sum_{q=1}^{4} d_E^q(k) \tag{2.7}$$

unde q reprezintă indicele cadranului, $q \in \{1,2,3,4\}$, iar $d_E^q(k)$ este distanţa Euclidiană:

$$d_E^q{}^2(k) = \sum_{j=1}^{216} [H_{k+2,q}(j) - H_{k,q}(j)]^2 \tag{2.8}$$

unde j reprezintă indicele culorii iar $H_{k,q}(j)$ reprezintă histograma color a cadranului q din imaginea la momentul k.

Mai mult, valorile funcţiei de discontinuitate astfel obţinute sunt derivate pentru a reduce fluctuaţiile cauzate de mişcările globale ale camerei video sau de diferenţe succesive de culoare. Principiul de funcţionare este ilustrat în Figura 2.3. Un "cut" produce o discontinuitate între două imagini succesive, discontinuitate ce se traduce printr-o valoare semnificativă a funcţiei de discontinuitate. Pe de altă parte, imaginile ce preced şi respectiv, ce succed tranziţia, sunt foarte similare, valorile funcţiei de discontinuitate fiind în acest caz reduse. Astfel, o tranziţie de tip "cut" are o semnătură particulară a valorilor funcţiei de discontinuitate, şi anume o succesiune de valori de tip SHS (S-valoare redusă, H-valoare ridicată) ce poate fi pusă în evidenţă cu ajutorul derivatei temporale.

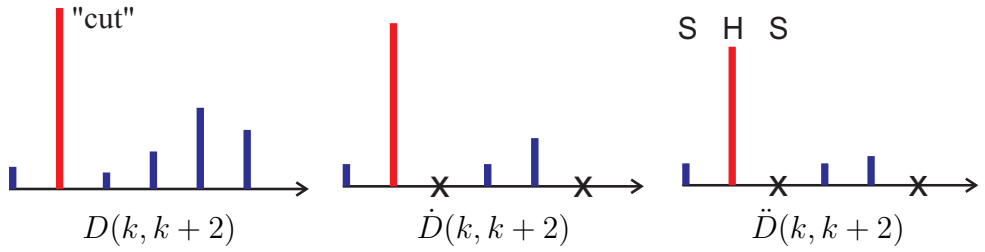


Figura 2.3: Folosirea derivatei temporale pentru accentuarea tranziţiilor de tip "cut" (axa oX corespunde axei temporale, valorile negative ale derivatei, marcate cu simbolul $\times$, sunt anulate deoarece conţin informaţie redundantă pentru detecţie).

Un studiu comparativ al diferitelor metode de detecţie de "cut" bazate pe calculul histogramelor color, pe fluxul video MPEG şi respectiv pe estimarea mişcării, este propus în [Gargi 00]. Astfel, în contextul prezentat, cea mai eficientă metodă de detecţie se dovedeşte a fi o metodă bazată pe histogramă, ce foloseşte intersecţia histogramelor în spaţiul de culoare al lui

Munsell (MTM, vezi Secţiunea 4.1). Din punct de vedere al complexităţii de calcul, aceasta are însă o complexitate moderată, raportat la celelalte metode testate. Acest lucru este totuşi valabil şi în general, metodele bazate pe analiza intensităţii pixelilor se dovedesc încă a fi metodele cele mai robuste de detecţie.

Metode bazate pe analiza de contur

Acestea folosesc pentru detecţie analiza contururile obiectelor prezente în imagine. Un "cut" produce, pe lângă discontinuitatea vizuală a fluxului secvenţei, şi o discontinuitate structurală la nivel de imagine. Astfel, contururile obiectelor prezente în imaginea ce precede un "cut" nu se vor mai regăsi în imaginea ce succede tranziţia. Metodele existente exploatează tocmai această proprietate.

Algoritmii propuşi în [Zabih 95] şi [Zabih 99] folosesc pentru detecţie calculul raportului de schimbare a conturului, ECR ("Edge Change Ratio"). Acesta este definit pentru imaginile la momentele k şi $k+l$ în felul următor:

$$ECR_{k+l} = max\left(\frac{X_k^{out}}{\sigma_k}, \frac{X_{k+l}^{in}}{\sigma_{k+l}}\right) \tag{2.9}$$

unde σ_k reprezintă numărul de puncte de contur existente în imaginea la momentul k iar X_k^{out} şi X_{k+l}^{in} reprezintă numărul de puncte de contur ce au dispărut din imaginea la momentul k, şi respectiv, numărul de puncte de contur ce au apărut în imaginea la momentul $k+l$.

Pentru a îmbunătăţii invarianţa raportului ECR la prezenţa mişcării, [Zabih 95] propune folosirea compensării de mişcare ce este calculată folosind distanţa Hausdorff[5]. Punctele de contur analizate în imaginea curentă la momentul k, ce sunt apropiate de punctele de contur din imaginea următoare analizată la momentul $k+l$, nu vor fi luate în calcul decât dacă distanţa Hausdorff depăşeşte 6 pixeli. Un alt exemplu este metoda propusă în [Kim 02] ce foloseşte pentru detecţie histograme color calculate în spaţiul de culoare YUV, precum şi evaluarea raportului de potrivire al contururilor, EMR ("Edge Matching Rate"). În [Lienhart 00] sunt propuse mai multe metode pentru detecţia tranziţiilor video, metode ce folosesc diverse informaţii, precum informaţii de contur, histograme şi analiza de mişcare.

Un studiu comparativ al performanţelor metodelor bazate pe analiza de contur precum şi a celor bazate pe histogramă este propus în [Lienhart 01a] şi [Lupatini 98]. Astfel, testele efectuate arată că metodele bazate pe analiza contururilor sunt mai puţin eficiente şi necesită un timp de calcul mult mai

[5]vezi explicaţia de la pagina 172.

ridicat decât metodele bazate pe histogramă. Totuşi, metodele bazate pe contur au avantajul de a putea fi utilizate în acelaşi timp şi pentru detecţia tranziţiilor video graduale, precum tranziţiile de tip "fade" sau "dissolve" [Lienhart 01b].

Metode bazate pe analiza de mişcare

Pentru detecţie, metodele bazate pe mişcare pornesc de la ipoteza că o tranziţie de tip "cut" produce în secvenţă o discontinuitate a mişcării. Detecţia este astfel bazată pe estimarea câmpului vectorial de mişcare (un studiu detaliat al tehnicilor de estimare a mişcării este prezentat în Secţiunea 3.1). Estimarea mişcării este în general realizată folosind metode bazate pe blocuri de pixeli ("block-based"), în principal datorită faptului că acestea oferă un bun compromis între complexitatea de calcul şi precizia rezultatelor. Procedeul constă mai întâi în împărţirea imaginilor analizate în blocuri disjuncte de pixeli. Mai departe, pentru fiecare bloc de pixeli, $b_i(k)$, de indice i, din imaginea la momentul k, se caută blocul cel mai similar, $b_{i,j}(k+l)$, de indice j, din imaginea la momentul următor $k+l$ (l este pasul de analiză), astfel $b_i(k) \approxeq b_{i,j}(k+l)$.

Similaritatea între blocurile de pixeli se traduce prin minimizarea unei funcţii de cost, $F_c()$, ce poate fi distanţa Euclidiană între pixeli, eroarea pătratica medie, eroarea absolută, etc. Astfel, eroarea minimală este dată de ecuaţia următoare:

$$\tilde{D}_{k,k+l}(i) = min|_{j=1,..,N_{cand}} F_c(b_i(k), b_{i,j}(k+l)) \tag{2.10}$$

unde $\tilde{D}_{k,k+l}(i)$ reprezintă valoarea minimă a funcţiei de cost, N_{cand} reprezintă numărul de blocuri $b_{i,j}(k+l)$ din imaginea la momentul $k+l$ ce sunt "canditate" pentru a fi similare cu blocul curent, $b_i(k)$. Acestea definesc ceea ce se numeşte fereastră de căutare.

Dacă imaginile la momentele k şi $k+l$ sunt imagini vecine în interiorul unui plan video, atunci valorile funcţiei de cost sunt mici sau chiar apropiate de zero. Pe de altă parte, dacă imaginea la momentul k este imaginea ce precede un "cut", valorile funcţiei de cost vor fi importante. Aceasta se datorează faptului că între cele două imagini există o diferenţă vizuală semnificativă, iar blocurile de pixeli ale imaginii k nu vor fi regăsite în imaginea $k+l$.

De exemplu, în [Shahraray 95] imaginile sunt divizate în 12 blocuri disjuncte de pixeli iar compensarea mişcării este realizată folosind ca funcţie de cost diferenţa între intensităţile pixelilor. Valorile $\tilde{D}_{k,k+l}()$ astfel obţinute sunt ordonate şi normalizate între 0 şi 1 fiind notate cu $d^s_{k,k+l}()$. Măsura de

discontinuitate între imaginile la momentele k şi $k+l$ este calculată pe baza unui anumit set de ponderi, c_i, în felul următor:

$$DSC(k, k+l) = \sum_{i=1}^{12} c_i \cdot d^s_{k,k+l}(i) \tag{2.11}$$

Mai departe, tranziţiile de tip "cut" sunt detectate prin filtrarea valorilor lui $DSC(k, k+l)$ cu un prag predefinit.

O altă abordare similară este metoda propusă în [Porter 00] ce foloseşte ca funcţie de cost corelaţia dintre blocurile de pixeli, calculată de această dată în domeniul frecvenţial.

Metoda propusă în [Hanjalic 02] foloseşte pentru detecţie informaţii precum statistica duratei tranziţiilor, compensarea mişcării şi amplitudinea diferenţelor temporale ce survin în secvenţă. Alte abordări cu o complexitate de calcul mai ridicată, folosesc estimarea fluxului optic iar măsurile de similaritate între imagini sunt calculate pe baza vectorilor de mişcare şi a deplasărilor survenite în imagine [Zhong 96] [Lupatini 98].

Din punct de vedere global, metodele de detecţie a tranziţiilor de tip "cut" bazate pe analiza mişcării sunt mai puţin eficiente şi precise decât metodele bazate pe histogramă [Gargi 00]. Estimarea mişcării este o operaţie cu o complexitate de calcul ce poate fi importantă, fiind mult mai complexă decât evaluarea unor histograme color [Lienhart 01b]. Cu toate acestea, metodele bazate pe mişcare rămân o soluţie pentru secvenţele ce au un conţinut preponderent dinamic, caz în care metodele bazate pe contur şi respectiv pe analiza intensităţii pixelilor, tind să obţină un număr ridicat de false detecţii, datorat în principal diferenţelor succesive dintre imaginile în mişcare.

Metode de detecţie în domeniul comprimat

Metodele de detecţie din această categorie exploatează direct informaţia din domeniul comprimat al fluxului video MPEG, ca de exemplu coeficienţii transformatei Cosinus Discrete (DCT - "Discrete Cosine Transform"[6]).

De exemplu, [Fernando 01] propune detecţia tranziţiilor de tip "cut" în secvenţele codate MPEG-2. Metoda propusă foloseşte analiza numărului de predicţii ale macro-blocurilor în cadrele de tip B (compresie bidirecţională ce foloseşte informaţie atât din imaginile anterioare, cât şi din cele ulterioare imaginii curente).

[6] "Discrete Cosine Transform" sau DCT reprezintă transformata cosinus discretă. Aceasta permite reprezentarea unei mulţimi finite de valori într-un spaţiu de reprezentare frecvenţial, sub forma unei sume de funcţii cosinus ce oscilează pe diferite frecvenţe.

Metoda propusă în [Arman 93b] foloseşte pentru detecţie sub-blocuri de pixeli, de dimensiune 8×8, codate DCT, ce sunt alese din n regiuni conexe ale imaginii curente k. Pentru toate aceste blocuri sunt reţinuţi, în mod aleator, doar 64 din coeficienţii DCT ai componentei alternative. Astfel, fiecare imagine va fi reprezentată în domeniul comprimat de un vector de coeficienţi, $V_k = (c_1, c_2, ..., c_{64})$, unde c_i reprezintă coeficientul de indice i, $i = 1, ...64$. Similaritatea imaginilor la momentele k şi $k + l$ este calculată pe baza produsului scalar normalizat al vectorilor V_k, dat de relaţia:

$$\Psi_{k,k+l} = \frac{V_k \cdot V_{k+l}}{|V_k| \cdot |V_{k+l}|} \tag{2.12}$$

Un "cut" este detectat în momentul în care $1 - |\Psi| > T$, unde T reprezintă pragul de discontinuitate. Metode similare ce folosesc fluxul MPEG sunt propuse şi în [Zhang 94] sau [Meng 95].

Metodele de detecţie ce folosesc direct informaţiile furnizate de fluxul MPEG nu necesită o etapă prealabilă de decompresie a datelor, etapă ce este necesară tuturor celorlalte metode ce folosesc o analiză la nivel de imagine. Coeficienţii fluxului MPEG conţin suficientă informaţie pentru a detecta discontinuităţile produse de tranziţiile de tip "cut". Complexitatea de calcul în acest caz este foarte redusă, lucru ce facilitează implementarea în timp real[7] a algoritmilor de detecţie. Totuşi, deseori precizia metodelor bazate pe fluxul MPEG este inferioară celorlalte metode datorită incoerenţei vectorilor de mişcare (vezi Secţiunea 3.1.5). Soluţia de compromis constă în decompresia datelor până la un anumit nivel de detaliu.

Ca tendinţă generală, este posibil ca noile standarde de codare, precum standardul de compresie video MPEG-7, să înglobeze informaţii referitoare la structura temporală a secvenţei, precum distribuţia de plane, scene, etc., lucru ce va face ca etapa de segmentare temporală să nu mai fie necesară [Wang 00].

Alte metode

În această categorie putem încadra metodele ce folosesc pentru detecţie alte surse de informaţie decât cele enumerate anterior. De exemplu, metoda propusă în [Boreczky 98] transformă problema detecţiei într-o problemă de clasificare. Astfel, aceasta propune segmentarea temporală în plane video pe baza modelelor Markov ascunse (HMM - "Hidden Markov Models"). Diferitele stări ale HMM sunt folosite pentru a modeliza diferitele tipuri de segmente ale secvenţei. Pentru mai multe detalii referitoare la folosirea modelelor Markov

[7] vezi explicaţia de la pagina 102.

în segmentarea temporală, cititorul se poate raporta la studiul [Wang 00]. O altă metodă ce tratează problematica segmentării, generic, independent de conţinutul secvenţei este propusă în [Hanjalic 02]. Aceasta este o abordare statistică ce foloseşte pentru segmentare minimizarea probabilităţii erorii medii de detecţie.

O abordare inedită a problemei detecţiei tranziţiilor de tip "cut" este propusă în [Guimaraes 03]. În prima fază, fiecare imagine a secvenţei este rezumată cu o singură linie de pixeli, şi anume, diagonala principală a imaginii. Liniile de pixeli astfel obţinute pentru întreaga secvenţă sunt juxtapuse pentru a forma o singură imagine, numită şi ritm vizual al secvenţei. Tranziţiile de tip "cut" apar în ritmul vizual sub formă de tranziţii verticale. Pentru detecţia acestora sunt folosite metode de detecţie de contur şi de morfologie matematică.

Aceste alte tipuri de abordări, relativ particulare sau inedite, au fost în general testate şi aplicate în cazuri particulare. Din această cauză, nu dispunem de suficiente teste comparative pentru a trage o concluzie relativă la performanţa acestora în comparaţie cu celelalte metode existente. Totuşi, prin aceste abordări se încearcă valorificarea şi a altor surse de informaţie decât cele clasice: imagine-contur-mişcare.

Problematica estimării pragului de detecţie

Marea majoritate a metodelor de măsurare a discontinuităţilor produse de tranziţiile de tip "cut" folosesc unul sau mai multe praguri pentru detecţie. Noţiunea de *similaritate* între două imagini se rezumă în final la compararea unei măsuri de distanţă cu un anumit prag, calculat sau fixat "a priori". Dacă valoarea acesteia depăşeşte valoarea pragului, atunci vorbim despre *imagini disimilare*, în caz contrar, imaginile sunt considerate ca fiind similare. Această operaţie este cunoscută în literatura de specialitate sub numele de "thresholding".

Astfel, alegerea adecvată a pragurilor este esenţială pentru precizia detecţiei. După cum am menţionat la începutul acestui capitol, un prag ales prea mic va avea ca efect creşterea numărului de false detecţii, iar un prag prea ridicat va conduce la un număr mare de tranziţii ce vor trece nedetectate. Pentru un studiu bibliografic complet relativ la metodele de estimare a pragului de detecţie, cititorul se poate raporta la lucrările [Lienhart 01b] şi [Hanjalic 02].

Primele abordări de estimare a pragurilor de detecţie erau bazate pe *metode euristice*, ca cele propuse în: [Otsuji 91], [Nagasaka 92], [Arman 93b]. Astfel, pragurile erau alese "a priori" ca rezultat al expertizei manuale a datelor folosite. Alte abordări propuneau o *analiză statistică* a distribuţiei valo-

rilor funcţiei de discontinuitate folosită la detecţie. De exemplu, în [Zhang 93] este propusă modelizarea acestei distribuţii cu funcţii Gaussiene de medie μ şi varianţă σ^2. Astfel, pragul de detecţie, T, era definit ca fiind:

$$T = \mu + r \cdot \sigma \tag{2.13}$$

unde parametrul r corespunde unei probabilităţi de falsă detecţie ce este fixată "a priori".

Un alt exemplu este abordarea propusă în [Ionescu 06a]. Aceasta porneşte de la observaţia că discontinuitatea produsă de tranziţiile de tip "cut" este mai puţin frecventă decât alte discontinuităţi produse de variaţii de culoare sau de mişcare. Astfel, în această ipoteză, pragul T definit în ecuaţia 2.13 va avea o valoare prea scăzută ce va duce la un număr important de false detecţii. Pentru a corecta acest lucru, metoda propusă calculează pragul în două etape. În prima etapă sunt selectate maximele locale ale funcţiei de discontinuitate ce sunt superioare mediei globale a acesteia. Pragul de detecţie este calculat în a doua etapă ca fiind valoarea medie a acestor maxime pentru întreaga secvenţă, asigurând astfel o valoare apropiată de valoarea optimă.

Aceste două tipuri de abordări, euristică şi statistică, au ca rezultat determinarea unui prag global de detecţie ce este acelaşi pentru întreaga secvenţă.

O altă metodă constă în calcularea pragurilor în mod *adaptiv*. [Yeo 95] propune calcularea pragului de detecţie, T, în funcţie de informaţia temporală a secvenţei. Sevenţa este împărţită în ferestre temporale de analiză de dimensiune N. Astfel, un "cut" este detectat în mijlocul ferestrei curente de analiză, dacă funcţia de discontinuitate, $D(k, k+1)$, calculată între imaginile la momentele k şi respectiv $k+1$, îndeplineşte următoarele condiţii:

$$D(k, k+1) = max|_{i=-\frac{N}{2},\ldots,\frac{N}{2}}\{D(k+i, k+1+i)\} \tag{2.14}$$

$$D(k, k+1) \geq \alpha \cdot D_{smax} \tag{2.15}$$

unde D_{smax} reprezintă a doua valoare maximală a funcţiei de discontinuitate obţinută pentru fereastra temporală curentă de dimensiune N, iar α este un parametru determinat în funcţie de forma funcţiei de discontinuitate. Pentru mai multe detalii referitor la metodele de calcul adaptiv al pragului de detecţie, cititorul se poate raporta la [Gargi 00] şi [Truong 00b].

O altă abordare a problemei estimării pragului de detecţie sunt *metodele mixte* ce combină metodele adaptive cu abordările statistice. De exemplu, [Hanjalic 97] propune modelarea cu funcţii Gausiene a distribuţiei valorilor de discontinuitate în ferestre temporale de analiză. Parametrul α, din ecuaţia 2.15, este determinat pe baza analizei probabilităţii "a priori" de falsă detecţie propusă în [Zhang 93].

Calculul *pragului optimal* pentru detecţie reprezintă o altă direcţie de studiu. Metodele din această categorie sunt inspirate din teoria detecţiei statistice. Acestea folosesc informaţii statistice despre distribuţia tranziţiilor de tip "cut" obţinute în urma expertizei manuale a unui număr semnificativ de secvenţe de imagini. Regula de detecţie a discontinuităţii unui "cut" este calculată prin minimizarea erorii de detecţie [Vasconcelos 00].

2.2.2 Detecţia de "fades"

După cum am menţionat în partea introductivă a acestui capitol, o tranziţie de tip "fade" este o tranziţie video graduală ce corespunde efectului optic prin care imaginea de interes apare progresiv dintr-o imagine constantă, de regulă neagră. Acest efect este numit şi "fade-in". Procesul invers, respectiv de dispariţie progresivă a imaginii curente către un fond constant, este numit "fade-out" (vezi Figura 2.2). Deseori, cele două tipuri de "fade" sunt folosite împreună, unul după altul, pentru a forma o secvenţă de tip "fade-out" - "fade-in". În acest caz, cele două tranziţii se comportă ca o singură tranziţie globală numită şi grup de "fade".

Secvenţa de imagini ce constituie o tranziţie de tip "fade" de durată T, notată $F(x, y, t)$, în care (x, y) reprezintă spaţiul imaginii iar t dimensiunea temporală, este definită matematic ca fiind transformarea intensităţilor pixelilor secvenţei $S_1(x, y, t)$ printr-o funcţie monotonă $f(t)$ [Lienhart 01b]. Astfel, aceasta este dată de relaţia:

$$F(x, y, t) = f(t) \cdot S_1(x, y, t) \tag{2.16}$$

unde $0 \leq t \leq T$.

Tranziţiile de tip "fade-in" folosesc funcţii monotone, $f(t)$, cu proprietatea că $f(0) = 0$ şi $f(T) = 1$. Similar, pentru o tranziţie de tip "fade-out", funcţia $f(t)$ trebuie să îndeplinească condiţiile $f(0) = 1$ şi respectiv $f(T) = 0$. În cele mai multe cazuri, funcţia $f(t)$ este aleasă ca fiind liniară şi este definită în felul următor:

$$f_{fade-in}(t) = \frac{t}{T} \tag{2.17}$$

$$f_{fade-out}(t) = 1 - \frac{t}{T} \tag{2.18}$$

Comparate cu varietatea de metode disponibile pentru detecţia tranziţiilor abrupte de tip "cut", metodele existente pentru detecţia de "fade" sunt mai puţin numeroase. Acest lucru se datorează pe de-o parte complexităţii algoritmilor de detecţie precum şi a faptului că frecvenţa acestora în secvenţă este net inferioară tranziţiilor de tip "cut". Mai mult, tranziţiile graduale

pot fi aproximate fără pierderi semnificative pentru segmentarea temporală, cu tranziţii abrupte, pe când reciproca nu este valabilă.

Metodele de detecţie de "fade" existente sunt orientate către trei axe principale de studiu, şi anume:

- metode bazate pe *analiza intensităţii pixelilor*,
- metode bazate pe *analiza contururilor*,
- alte metode ce folosesc *alte surse de informaţie* decât cele clasice enumerate anterior.

Pentru un studiu bibliografic complet al literaturii de specialitate, cititorul se poate raporta la lucrările [Lienhart 01b], [Hanjalic 02] sau [Ren 03]. În cele ce urmează, ne vom limita la o prezentare succintă a particularităţilor metodelor din fiecare categorie.

Metode bazate pe analiza intensităţii pixelilor

Una dintre primele metode de detecţie a fost propusă în [Zhang 93] şi se baza pe folosirea a două praguri. Aceasta modifică o metodă de detecţie a tranziţiilor de tip "cut" bazată pe analiza distanţei între histograme. Astfel, un "cut" era detectat dacă disimilaritatea între două imagini succesive era superioară unui prag τ_{start}. Dacă la un moment k, disimilaritatea între două imagini succesive era superioară unui al doilea prag, τ_{cand} ($\tau_{cand} < \tau_{start}$), dar totodată inferioară pragului iniţial τ_{start}, atunci imaginea k era considerată ca o potenţială imagine de început a unei tranziţii graduale. Mai departe, această imagine era comparată cu imaginile următoare, iar distanţa între acestea acumulată. În momentul în care cumulul distanţelor depăşea pragul τ_{start}, dar în acelaşi timp diferenţele individuale dintre imaginile succesive rămâneau inferioare pragului τ_{cand}, atunci era detectată o tranziţie de tip "fade".

Pe durata unei tranziţii de tip "fade", una dintre informaţiile caracteristice acesteia o reprezintă schimbarea intensităţii luminoase a imaginii. Astfel, alte abordări studiază evoluţia dispersiei intensităţii luminoase globale a imaginii. În condiţii de ergodicitate[8], dispersia secvenţei unei tranziţii de tip "fade" poate fi exprimată ca:

$$\sigma(F(x,y,t)) = f(t) \cdot \sigma(S_1(x,y)) \tag{2.19}$$

[8]un proces aleator este considerat ca fiind ergodic dacă momentele statistice sunt egale cu momentele temporale.

unde $F(.)$, $f(.)$ şi $S_1(.)$ au aceeaşi semnificaţie ca în ecuaţia 2.16 iar $\sigma()$ reprezintă dispersia. Se poate observa că dispersia secvenţei unui "fade" respectă monotonia funcţiei $f(t)$.

Metoda de detecţie propusă în [Lienhart 99a] localizează mai întâi în secvenţă imaginile monocromatice pentru care valoarea varianţei intensităţii luminoase este apropiată de zero. Aceste imagini particulare sunt posibile imagini de început sau finale ale unor tranziţii de tip "fade-in" şi respectiv "fade-out". Mai departe, detecţia este realizată prin analiza creşterii progresive a intensităţii luminoase şi a dispersiei acesteia în sensul pozitiv al axei temporale. Liniaritatea este verificată prin evaluarea erorii de aproximare a valorilor acestora cu dreptele de regresie ce le modelează.

O abordare similară, bazată pe analiza varianţei intensităţii luminoase, este propusă în [Alattar 97]. Aceasta propune o detecţie preliminară pe baza analizei punctelor de extrem negativ ale derivatei a doua a valorilor varianţei intensităţii luminoase a imaginii. Detecţia unui "fade" este ulterior confirmată dacă valoarea derivatei întâi a intensităţii luminoase medii este constantă între două puncte de extrem negativ.

Metoda propusă în [Truong 00a] vine să combine metodele propuse în [Lienhart 99a] şi [Alattar 97]. Astfel, în primă fază a detecţiei vor fi selectate imaginile monocromatice. Dintre acestea sunt reţinute doar imaginile ce furnizează valori apropiate de extremele negative ale derivatei secunde a varianţei intensităţii luminoase din imagine. Un "fade" este mai departe detectat dacă sunt satisfăcute criteriile următoare:

- derivata întâi a valorilor intensităţii medii rămâne constantă şi nu îşi schimbă semnul,
- valoarea medie a pantei derivatei întâi trebuie să fie superioară unui anumit prag,
- valoarea varianţei intensităţii luminoase pentru prima şi ultima imagine a tranziţiei trebuie de asemenea să fie superioare unui anumit prag.

Una dintre principalele surse de erori în cazul metodelor de detecţie de "fade" este prezenţa mişcării. Aceasta face ca intensitatea luminoasă să aibă variaţii neliniare ce nu mai respectă monotonia funcţiei $f(t)$ (vezi ecuaţia 2.16). În [Fernando 99], pentru a reduce influenţa mişcării, detecţia este realizată folosind atât măsuri statistice ale intensităţii pixelilor, cât şi măsuri statistice ale semnalului de crominanţă. Analiza este efectuată în spaţiul de culoare YCbCr (Y-luminanţă şi Cb, Cr-diferenţe cromatice). Media semnalului de crominanţă, $C = \frac{C_b+C_r}{2}$, se dovedeşte a fi mai puţin sensibilă la prezenţa mişcării decât media semnalului de luminanţă Y. Aceste două

informaţii sunt folosite pentru a defini parametrul $R(k)$ ce reprezintă raportul de schimbare incrementală a mediei semnalului de luminanţă relativ la semnalul de crominanţă la momentul k:

$$R(k) = \begin{cases} \frac{\Delta_k^Y}{\Delta_k^C} & \text{dacă } k < L_1 \text{ sau } k \geq (L_1 + T) \\ \frac{|C_0 - m_{k+1}^Y + (L_1 - k) \cdot \Delta_k^Y|}{|C_0 - m_{k+1}^C + (L_1 - k) \cdot \Delta_k^C|} & \text{dacă } L_1 \leq k < (L_1 + T) \end{cases} \tag{2.20}$$

unde Δ_k^Y şi Δ_k^C reprezintă schimbările incrementale ale mediei semnalului Y şi respectiv ale semnalului C, m_{k+1}^Y şi m_{k+1}^C sunt valorile medii ale lui Y şi C calculate la momentul $k+1$, L_1 reprezintă momentul de început al tranziţiei, T este durata totală a acesteia, iar C_0 reprezintă nivelul semnalului video pentru începutul tranziţiei.

Detecţia este realizată mai departe pe baza analizei valorilor lui $R(k)$. Pe durata unui "fade", acestea trebuie să rămână aproximativ constante, ceea ce face ca valorile diferenţei $|R(k) - R(k-1)|$ să fie apropiate de zero.

O abordare diferită, cu o complexitate de calcul mai redusă, a principiului metodei din [Fernando 99], este propusă în [Ionescu 05a] unde detecţia este realizată pe baza analizei colaborative a infomaţiei de culoare şi a intensităţii luminoase. Astfel, detecţia este realizată prin analiza evoluţiei temporale a trei parametrii, şi anume:

- valoarea medie a componentei de intensitate luminoasă a imaginii, $\overline{Y}$, calculată în spaţiul de culoare $YCbCr$,
- varianţa globală în imagine a componentei Y, $\sigma^2(Y)$,
- valoarea absolută a diferenţei dintre valorile medii pe imagine ale componentelor Cb şi Cr, $|\overline{Cb} - \overline{Cr}|$.

Varianţa $\sigma^2(Y)$ este folosită pentru a detecta începutul unei tranziţii de tip "fade-in" şi respectiv sfârşitul unei tranziţii de tip "fade-out", deoarece în acest caz, varianţa are valori apropiate de zero (fiind vorba de imagini constante). Detecţia propriu-zisă este realizată prin localizarea evoluţiei crescătoare ("fade-in") şi respectiv descrescătoare ("fade-out") a parametrilor $\overline{Y}$ şi $|\overline{Cb} - \overline{Cr}|$. Parametrul $|\overline{Cb} - \overline{Cr}|$ se dovedeşte a fi mai puţin sensibil la prezenţa mişcării decat $\overline{Y}$, care la rândul său se dovedeşte a fi mai eficient în absenţa mişcării, astfel că cei doi parametri sunt folosiţi disjunctiv.

Metode bazate pe analiza de contur

O altă direcţie de studiu este analiza bazată pe informaţia de contur. Metodele din această categorie se folosesc de ipoteza că pe durata unei tranziţii

de tip "fade", contururile obiectelor din imagine fie dispar ("fade-out"), fie apar gradual ("fade-in").

O măsură cantitativă a schimbării de contur este dată de raportul ECR ("Edge Change Ratio") propus în [Zabih 95], măsură folosită şi la detecţia tranziţiilor de tip "cut" (vezi Secţiunea 2.2.1). Pe durata unei tranziţii de tip "fade-in" numărul de pixeli de contur ce apar în imagine, notat ECR_{in}, este superior numărului de pixeli de contur ce dispar, notat ECR_{out}, astfel $ECR_{in} > ECR_{out}$. Similar, pentru un "fade-out" întâlnim situaţia inversă, şi anume $ECR_{in} < ECR_{out}$.

Dacă tranziţiile de tip "cut" corespundeau valorilor maxime ale funcţiei ECR, tranziţiile de tip "fade" precum şi alte tranziţii graduale sunt caracterizate de intervale de valori semnificative ale ECR [Zabih 99]. Marea majoritate a metodelor de detecţie existente bazate pe analiza de contur folosesc abordări similare, ce se bazează pe contabilizarea punctelor de contur ce apar sau dispar din scenă [Yu 97] [Lupatini 98].

Global, metodele bazate pe analiza conturului sunt mai puţin eficiente decât cele bazate pe analiza intensităţii pixelilor [Lienhart 01b]. Acest lucru se datorează în mare parte sensibilităţii ridicate a detecţiei la prezenţa mişcării sau a diverselor efecte de culoare. Acestea conduc la schimbarea substanţială a distribuţiei conturilor din imagine şi astfel la eşuarea detecţiei.

Alte metode

În această categorie putem menţiona mai întâi *abordările mixte*, ce fuzionează diferitele surse de informaţie disponibile, precum metodele ce folosesc informaţii statistice ale intensităţii pixelilor, dar care realizează analiza în domeniul comprimat al coeficienţilor DCT [Bimbo 99]. Pe de altă parte putem menţiona abordările inovantive ce încearcă exploatarea de noi surse de informaţie pentru a ameliora punctele slabe ale metodelor clasice existente. Astfel putem menţiona:

- utilizarea informaţiei de mişcare pentru a reduce influenţa acesteia şi astfel numărul de false detecţii [Porter 01],
- utilizarea ritmului vizual bazat pe histogramă pentru a reduce influenţa zgomotului din imagine [Guimaraes 03],
- exploatarea de parametri calculaţi în domeniul frecvenţial al coeficienţilor FFT ("Fast Fourier Transform") ai imaginii [Miene 01],
- utilizarea de abordări statistice pentru detecţia generică a tranziţiilor graduale [Heng 01].

În general, fiecare metodă propusă prezintă o serie de avantaje cât şi de inconveniente, care în anumite situaţii vor permite ameliorarea detecţiei raportat la metodele clasice, dar vor da greş în altele. De exemplu, metoda propusă în [Guimaraes 03], bazată pe ritmul vizual, detectează poziţia în secvenţă cât şi durata tranziţiilor de tip "fade" cu o precizie de o imagine, pe când în majoritatea abordărilor clasice, respectarea duratei tranziţiei nu este un criteriu de calitate. Pe de altă parte, metoda propusă eşuează în cazul în care tranziţiile prezintă diferenţe de luminozitate între imaginea de început şi cea de final.

2.2.3 Detecţia de "dissolves"

Un alt tip de tranziţie graduală des întâlnită în secvenţele de imagini sunt tranziţiile de tip "dissolve". Un "dissolve" reprezintă efectul obţinut prin transformarea graduală progresivă a unei imagini în alta. Transformarea este realizată la nivel de intensitate a pixelilor şi nu prin transformări geometrice (vezi Figura 2.2).

Din punct de vedere matematic, secvenţa unei tranziţii de tip "dissolve", $D(x, y, t)$, de durată T, unde (x, y) reprezintă spaţiul imaginii iar t este dimensiunea temporală, este definită pe baza secvenţelor $S_1(x, y, t)$ şi $S_2(x, y, t)$ în felul următor:

$$D(x, y, t) = f_1(t) \cdot S_1(x, y, t) + f_2(t) \cdot S_2(x, y, t) \tag{2.21}$$

unde $0 \leq t \leq T$. În funcţie de forma funcţiilor $f_1(t)$ şi $f_2(t)$ folosite, întâlnim mai multe tipuri de "dissolve" [Lienhart 01b].

Cel mai frecvent folosite sunt tranziţiile de tip "cross-dissolve" (vezi Figura 2.4). Acestea sunt construite ca suprapunerea unei tranziţii de tip "fade-out" cu o tranziţie de tip "fade-in", astfel funcţiile $f_1(t)$ şi $f_2(t)$ fiind cele utilizate de tranziţiile de tip "fade":

$$f_1(t) = 1 - \frac{t}{T}, \quad f_2(t) = \frac{t}{T} \tag{2.22}$$

unde $0 \leq t \leq T$.

O altă categorie de "dissolve" mai puţin utilizată sunt tranziţiile "dissolve" aditive ("additive dissolve", vezi Figura 2.4). Acestea pot fi văzute ca fiind suma unei tranziţii de tip "fade-out" cu o tranziţie de tip "fade-in". În acest caz, cele două funcţii $f_1(t)$ şi $f_2(t)$ sunt definite astfel:

$$f_1(t) = \begin{cases} 1 & \text{dacă } t \leq c_1 \\ \frac{T-t}{T-c_1} & \text{altfel} \end{cases} \tag{2.23}$$

$$f_2(t) = \begin{cases} \frac{t}{c_2} & \text{dacă } t \leq c_2 \\ 1 & \text{altfel} \end{cases} \tag{2.24}$$

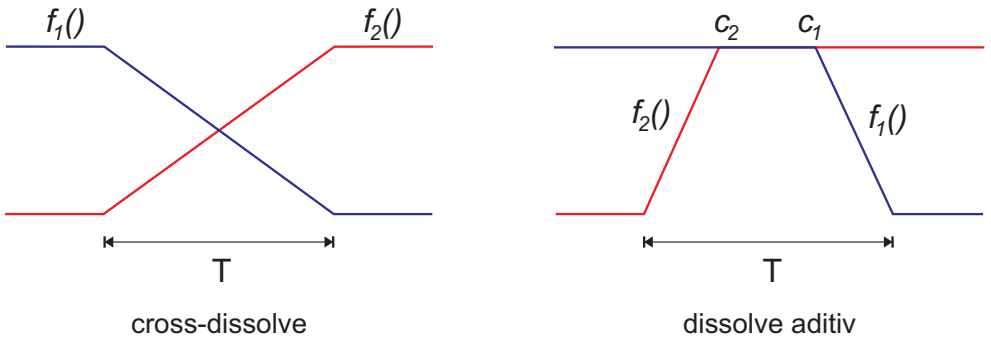


Figura 2.4: Funcţiile de scalare folosite de tranziţiile de tip "dissolve" (axa orizontală corespunde axei temporale).

unde $c_1, c_2 \in (0; T)$, $c_2 < c_1$ şi $0 \leq t \leq T$.

Din punct de vedere al diferenţei vizuale dintre secvenţele $S_1(x, y, t)$ şi respectiv $S_2(x, y, t)$, ce constituie tranziţia "dissolve", acestea se împart în trei categorii [Lienhart 01b], astfel:

- secvenţele S_1 şi S_2 au *distribuţii de culoare suficient de diferite* astfel încât tranziţia "dissolve" rezultată să fie confundată cu una sau mai multe tranziţii de tip "cut",
- secvenţele S_1 şi S_2 au *distribuţii de culoare similare* ce nu sunt detectabile folosind metode de detecţie a discontinuităţii vizuale bazate pe histogramă. Pe de altă parte, diferenţele la nivel structural ale obiectelor prezente în scenă sunt importante, fiind detectabile cu metode bazate pe analiza contururilor,
- secvenţele S_1 şi S_2 au *distribuţia de culoare şi structura obiectelor din scenă similare*. Acest tip de "dissolve" este de fapt un caz particular al unui efect de "morphing"[9].

Similar cu metodele de detecţie a tranziţiilor de tip "fade", metodele existente de detecţie a tranziţiilor de tip "dissolve" se împart în trei categorii [Lienhart 01b]:

- metode bazate pe *analiza intensităţii pixelilor* din imagine,
- metode bazate pe *analiza contururilor*,
- alte metode ce folosesc *alte surse de informaţie* decât cele menţionate anterior.

[9] vezi explicaţia de la pagina 32.

Pentru un studiu bibliografic complet al literaturii de specialitate, cititorul se poate raporta la lucrările [Bimbo 99] [Lienhart 01b], [Hanjalic 02], [Ren 03] sau [Su 05a]. În cele ce urmează, vom face o trecere în revistă a tehnicilor de detecţie reprezentative pentru fiecare categorie de metode.

Metode bazate pe analiza intensităţii pixelilor

Dacă pe durata unei tranziţii de tip "cross-dissolve", mişcarea obiectelor sau a camerei video este neglijabilă, atunci pornind de la ecuaţia 2.19 putem obţine următoarea relaţie:

$$\frac{\partial D(x,y,t)}{\partial t} = \frac{S_2(x,y) - S_1(x,y)}{T} \tag{2.25}$$

unde $0 \leq t \leq T$ iar T reprezintă durata tranziţiei. Astfel, o posibilă metodă de detecţie o reprezintă localizarea în secvenţă a tuturor schimbărilor liniare ale intensităţii pixelilor, metodă propusă în [Hampapur 95]. Această abordare, conform chiar ipotezei de plecare, este foarte sensibilă la prezenţa zgomotului în imagine cât şi la prezenţa mişcării.

Metoda propusă în [Gu 97] vine cu o serie de îmbunătăţiri menite să amelioreze invarianţa detecţiei. Detecţia este realizată de această dată în domeniul comprimat al fluxului MPEG pe baza coeficienţior DCT ai informaţiei de luminanţă. Pentru imagini succesive, este calculat procentul de blocuri de pixeli pentru care diferenţa absolută între coeficienţii DCT are o evoluţie specifică unei tranziţii de tip "dissolve". Un "dissolve" va fi detectat în cazul în care valorile astfel obţinute rămân superioare unui anumit prag pe durata a 10 până la 60 de imagini (durata medie maximală a unui "dissolve" calculată la o frecvenţă de cadre de 30 imagini/s).

O abordare ce nu foloseşte direct modelul matematic al unui "dissolve" o constituie analiza comportamentului temporal al pixelilor din imagine într-un anumit spaţiu de caracteristici. De exemplu, [Nam 00] aproximează evoluţia temporală a fiecărui coeficient DCT într-o fereastră de durată L, cu funcţii "B-spline"[10]. Dacă pe durata unui "dissolve" valoarea dispersiei funcţiei temporale de evoluţie a intensităţii pixelilor prezintă valori semnificative, eroarea de aproximare a evoluţiei coeficienţilor DCT cu funcţii "B-spline", trebuie să prezinte valori reduse. Aceasta se datorează efectului "sintetic" pe care il are tranziţia "dissolve" în evoluţia temporală a secvenţei. [Nam 00]

[10] în analiza numerică, o funcţie "spline" este o funcţie polinomială pe porţiuni folosită la interpolarea datelor. "B-spline" este un caz particular, fiind o funcţie "spline" care pentru anumite valori date ale gradului polinomial, gradului de uniformitate şi ale domeniului de definiţie, oferă un suport minimal al funcţiei.

defineşte eroarea de aproximare ca fiind:

$$e(x,y,t) = \frac{1}{L+1} \sum_{i=t-L/2}^{t+L/2} [D(x,y,i) - D_{spline}(x,y,i)]^2 \tag{2.26}$$

unde $D_{spline}(x,y,t)$ este aproximarea de tip "B-spline" iar $L = 31$. Comportamentul acestei funcţii va fi diferit pentru diversele tipuri de pasaje ale secvenţei. Astfel, pe durata unui "dissolve", $e(x,y,t)$ prezintă valori mici dar varianţa dintre imagini este semnificativă. Pe de altă parte, un pasaj ce conţine mişcării de obiecte sau mişcări ale camerei video, va conduce atât la o valoare semnificativă a varianţei cât şi a erorii.

Dacă presupunem că cele două secvenţe, $S_1(x,y,t)$ şi $S_2(x,y,t)$, ce constituie tranziţia de tip "dissolve" sunt procese aleatoare ergodice statistic independente, atunci varianţa secvenţei tranziţiei poate fi exprimată astfel:

$$Var\{D(x,y,t)\} = f_1^2(t) \cdot Var\{S_1(x,y)\} + f_2^2(t) \cdot Var\{S_2(x,y)\} \tag{2.27}$$

unde varianţele secvenţelor S_1 şi S_2 sunt independente de timp.

În cazul unui "cross-dissolve" (vezi ecuaţia 2.22), ecuaţia anterioară devine:

$$Var\{D(x,y,t)\} = \frac{(T-t)^2}{T^2} \cdot Var\{S_1(x,y)\} + \frac{t^2}{T^2} \cdot Var\{S_2(x,y)\} \tag{2.28}$$

şi mai departe:

$$Var\{D(x,y,t)\} = c \cdot (t-a)^2 - b \tag{2.29}$$

unde

$$a = \frac{T \cdot Var\{S_1(x,y)\}}{Var\{S_1(x,y)\} + Var\{S_2(x,y)\}} \tag{2.30}$$

$$b = \frac{Var\{S_1(x,y)\} \cdot Var\{S_2(x,y)\}}{Var\{S_1(x,y)\} + Var\{S_2(x,y)\}} \tag{2.31}$$

$$c = \frac{Var\{S_1(x,y)\} + Var\{S_2(x,y)\}}{T^2} \tag{2.32}$$

Calculând mai departe derivata întâi şi secundă a varianţei vom obţine relaţiile următoare:

$$\frac{\partial Var\{D(x,y,t)\}}{\partial t} = 2 \cdot c \cdot (t-a) \tag{2.33}$$

$$\frac{\partial^2 Var\{D(x,y,t)\}}{\partial t^2} = 2 \cdot c \tag{2.34}$$

Astfel, se observă că pe parcursul unei tranziţii de tip "dissolve", evoluţia temporală a varianţei intensităţii pixelilor are un comportament parabolic. Acest lucru are ca implicaţie valori apropiate de zero ale derivatei secunde, calculată înainte şi după "dissolve", şi valori constante şi pozitive pe durata tranziţiei.

Această ipoteză a fost folosită pentru prima oară în metoda propusă în [Alattar 93] ce exploata prezenţa a două puncte de extrem negativ în derivata secundă, puncte ce corespundeau începutului şi respectiv sfârşitului tranziţiei. Alte exemple sunt metodele propuse în [Fernando 99], [Gu 97] sau [Truong 00b]. De exemplu, metoda din [Truong 00b] foloseşte pentru detecţie următoarele consideraţii:

- derivata întâi a varianţei intensităţii pixelilor pe parcursul unui "cross-dissolve" trebuie să fie o funcţie monoton crescătoare ce porneşte cu o valoare negativă şi se opreşte într-o valoare pozitivă,
- varianţa intensităţii pixelilor pentru cele două secvenţe S_1 şi S_2 trebuie să fie superioară unui anumit prag τ_{min},
- durata unui "dissolve" nu trebuie să depăşească un anumit interval de valori, $[T_{min}, T_{max}]$, determinat experimental.

O abordare diferită este propusă în [Su 05b]. În prima fază, pentru fiecare imagine analizată la momentul k se constituie o nouă imagine pseudo-binară, $B_k^L(x, y)$, unde $L + 1$ reprezintă durata minimă a unei tranziţii de tip "dissolve". Pixelii din imagine ce prezintă o creştere sau respectiv o diminuare progresivă a valorii intensităţii, în intervalul temporal $[k-L; k]$, sunt marcaţi în imaginea $B_k^L(x, y)$ cu valoarea 1, fiind pixeli activi. În mod similar, ceilalţi pixeli sunt marcaţi cu 0 fiind pixeli inactivi. Pentru detecţie se foloseşte un indicator al cantităţii de pixeli activi din imaginea pseudo-binară. Dacă acesta depăşeşte valoarea unui anumit prag statistic, atunci detecţia este validată.

Metode bazate pe analiza de contur

O altă categorie de metode o reprezintă metodele ce folosesc pentru detecţie informaţia de contur. Acestea se folosesc de ipoteza conform căreia pe durata unei tranziţii de tip "dissolve", contururile obiectelor din imaginea de început a tranziţiei dispar progresiv, în timp ce contururile obiectelor din imaginea finală a tranziţiei apar progresiv. Astfel, contrastul imaginii va avea o evoluţie descrescătoare odată cu apropierea de mijlocul tranziţiei.

O măsură cantitativă a gradului de schimbare al contururilor din imagine este raportul ECR definit în [Zabih 95] pentru a servi la detecţia tranziţiilor

de tip "cut" şi "fade" (vezi ecuaţia 2.9). Unele metode existente de detecţie, precum metoda propusă în [Zabih 99], folosesc măsuri similare pentru a contabiliza numărul de pixeli de contur ce apar şi respectiv ce dispar din imagine. Totuşi aceste abordări sunt foarte sensibile la prezenţa mişcării ce duce la modificarea contururilor şi astfel la creşterea numărului de false detecţii.

Metoda din [Lienhart 99a] propune pentru detecţie o nouă mărime numită contrastul de contur sau EC ("Edge-Based Contrast"). Aceasta este o măsură matematică ce pune în evidenţă punctele de contur ce contrastează legătura între punctele de "contur slabe" şi "cele semnificative". Dacă presupunem că $K(x, y, t)$ reprezintă harta de contururi a imaginii curente analizate la momentul t, iar pragurile τ_w şi respectiv τ_s definesc punctele de contur slabe ($K(x, y, t) < \tau_w$) şi respectiv semnificative ($K(x, y, t) > \tau_s$), atunci raportul EC este dat de relaţia următoare:

$$EC(K) = 1 + \frac{s(K) - w(K) - 1}{s(K) + w(K) + 1} \tag{2.35}$$

unde $s(K)$ şi $w(K)$ reprezintă numărul total de puncte de contur semnificative şi respectiv slabe prezente în $K(x, y, t)$. Astfel, tranziţiile de tip "dissolve" au o semnătură particulară în evoluţia temporală a valorilor EC şi anume prezenţa unui minim local mărginit de variaţii abrupte ale valorilor EC.

Global, metodele bazate pe analiza de contur sunt mai puţin eficiente decât abordările ce folosesc analiza intensităţii pixelilor, acestea fiind mai vulnerabile la prezenţa mişcării sau a zgomotului în imagine.

Alte metode

În această categorie putem menţiona metoda propusă în [Lienhart 01a] ce vine cu o abordare diferită a problematicii detecţiei tranziţiilor de tip "dissolve". Astfel, în loc să localizeze tranziţiile pe baza analizei une funcţii de similaritate între cadre, detecţia se realizează pe bază de comparaţii cu modele de "dissolve" predefinite. Pentru aceasta, un sintetizor automat de "dissolve" generează un număr foarte mare de astfel de tranziţii folosind ca parametri durata tranziţiei precum şi poziţia imaginii centrale. Tranziţiile obţinute sunt folosite pentru a antrena în mod iterativ, pe baza metodei "bootstrap"[11], un clasificator optimal precum reţelele neuronale sau "sup-

[11] "bootstrap" este o metodă de inferenţă statistică. Aceasta estimează proprietăţile unui estimator prin măsurarea acestora pe baza mostrelor unei distribuţii aproximative ale acestuia.

port vector machines"[12], ce vor servi ulterior la detectarea tranziţiilor din secvenţă.

Un alt exemplu este metoda propusă în [Boccignone 00] ce efectuează detecţia direct în fluxul comprimat MPEG. Pentru aceasta sunt propuşi doi parametri, şi anume: parametrul D ce reprezintă procentul de blocuri de pixeli ce prezintă o diferenţă importantă între coeficienţii DCT, şi respectiv parametrul σ_{mv} ce caracterizează gradul de aleatoricitate al vectorilor de mişcare. Detecţia este realizată prin analiza evoluţiei temporale a celor două mărimi. Un "dissolve" este caracterizat de valori ale lui D şi σ_{mv} ce depăşesc un anumit prag determinat în mod automat.

Alte abordări ale detecţiei de "dissolves" se folosesc de modele Markov ascunse [Boreczky 98], de analiza similarităţii cadrelor în domeniul frecvenţial al coeficienţilor FFT [Miene 01] sau de estimarea de mişcare cu metode bazate pe blocuri de pixeli [Porter 01].

2.2.4 Evaluarea detecţiei tranziţiilor video

Am vorbit în capitolele anterioare de metodele de detecţie a tranziţiilor video, precum şi de performanţele acestora. La complexitatea procesului de detecţie al algoritmilor existenţi se adaugă o problemă conexă, ce nu este imediat evidentă. Aceasta o reprezintă **evaluarea** preciziei procesului de detecţie.

Evaluarea detecţiei constă în principiu în calcularea a o serie de *erori de detecţie* pe baza a ceea ce numim "realitate de teren" ("groundtruth"[13]). O "realitate de teren" în acest caz, este o *segmentare de referinţă* a secvenţei ce presupune marcarea poziţiilor reale ale tranziţilor video şi etichetarea acestora prin analiza manuală, cadru cu cadru, a secvenţei. Acestea constituie datele de referinţă pentru validare. Tranziţiile video rezultate din procesul de detecţie sunt ulterior comparate cu "realitatea de teren" pentru a evalua erorile de detecţie şi astfel precizia algoritmului.

Procesul de evaluare este o problemă complexă datorită volumului mare de date conţinut într-o secvenţă de imagini, date ce trebuiesc analizate manual pentru constituirea segmentării de referinţă. Mai mult, anumite tranziţii video în practică pot fi interpretabile, lucru ce duce la existenţa mai multor segmentări de referinţă, în funcţie de modul de percepţie al persoanei care le-a realizat.

În ceea ce priveşte procesul de validare, sunt vizate două situaţii de eroare ale algoritmului de detecţie, şi anume:

[12]"Support Vector Machine" sau SVM reprezintă o colecţie de metode "înrudite" de învăţare supervizată, ce sunt folosite pentru clasificarea datelor. Acestea fac parte din categoria clasificatorilor liniari generalizaţi (vezi Secţiunea 7.2.2).

[13]vezi explicatia de la pagina 170.

- algoritmul nu a detectat anumite tranziţii prezente în secvenţă, această situaţie este o **eroare de nedetectare**,
- algoritmul a detectat în mod eronat anumite pasaje ale secvenţei drept tranziţii, această situaţie fiind o **eroare de falsă detecţie**.

Pe baza acestor două situaţii posibile, metodele de evaluare existente propun diverse măsuri de eroare. Astfel, o primă abordare constă în calcularea erorii de detecţie, E_D, ce corespunde procentului de tranziţii nedetectate, precum şi a erorii de falsă detecţie, E_{FD}, ce corespunde procentului de false detecţii. Acestea sunt date de relaţiile următoare:

$$E_D = \frac{N_t - GD}{N_t} \tag{2.36}$$

$$E_{FD} = \frac{FD}{N_t} \tag{2.37}$$

unde N_t reprezintă numărul total de tranziţii prezente în secvenţă, GD reprezintă numărul de tranziţii ce au fost detectate corect iar FD reprezintă numărul de false detecţii. Folosind cele două măsuri, E_D şi E_{FD}, *performanţa algoritmului de detecţie este maximală când acestea sunt minimale.*

Abordarea cea mai frecvent întâlnită, constă în calculul erorilor de tip "precision" şi "recall". Acestea sunt date de relaţiile următoare:

$$Precision = \frac{GD}{GD + FD} \tag{2.38}$$

$$Recall = \frac{GD}{N_t} \tag{2.39}$$

unde N_t, GD şi FD au aceeaşi semnificaţie ca mai sus.

Eroarea de tip "precision" este într-un fel o măsură cantitativă a numărului de false detecţii. Aceasta este maximală (valoare 1) pentru $FD = 0$ şi deci în cazul în care nu au avut loc false detecţii. Pe de altă parte, eroarea de tip "recall" este o măsură a numărului de detecţii corecte, aceasta având valoarea maximală pentru $GD = N_t$ (adică toate tranziţiile prezente au fost detectate). Folosind cele două măsuri, "precision" şi "recall", *performanţa algoritmului de detecţie este maximală atunci când cele două mărimi au valori maxime.*

Indiferent de metoda adoptată, validarea detecţiei trebuie efectuată pe baza evaluării atât a detecţiilor corecte cât şi a falselor detecţii. Un algoritm ce detectează toate tranziţiile prezente în secvenţă nu este neapărat un algoritm eficient, acesta putând furniza adiţional un număr foarte mare de false detecţii, şi vice-versa.

O soluţie pentru a reprezenta simultan cele două erori o constituie prezentarea rezultatelor sub formă de curbe de precizie. Pentru aceasta, algoritmul de detecţie este rulat de mai multe ori pentru valori diferite ale parametrilor de reglaj (de exemplu, valori diferite ale pasului de analiză, valori diferite ale pragului de detecţie, etc.). Erorile de detecţie astfel obţinute, "precision" şi "recall", sunt reprezentate grafic ca puncte în spaţiul bidimensional format de acestea. Procesul se poate repeta similar şi pentru alţi algoritmi de detecţie. În final, metoda cea mai eficientă va fi metoda ce va furniza punctul cel mai apropiat de colţul din dreapta sus al graficului, prezentând valorile cele mai ridicate ale "precision" şi "recall".

În Figura 2.5 am ilustrat un exemplu de curbe de precizie obţinute pentru mai mulţi algoritmi de detecţie de "cut" (marcaţi cu simboluri diferite), precum şi pentru un acelaşi algoritm executat cu parametri de reglaj diferiţi (marcat cu simboluri de aceeaşi culoare). Astfel, pe baza curbei de precizie obţinute putem decide care sunt algoritmii de detecţie cei mai precişi, precum şi care sunt seturile de parametri optimali ce conduc la rezultatele cele mai performante (marcaţi în Figura 2.5 cu cercul roşu).

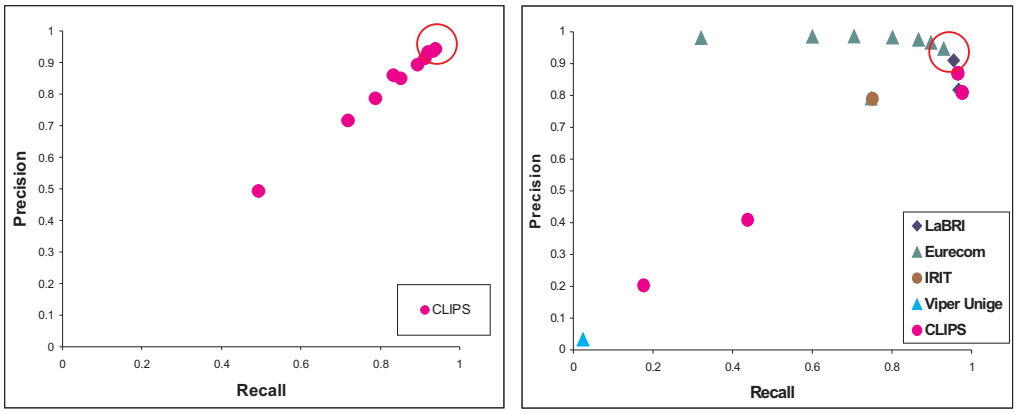


Figura 2.5: Exemplu de curbe de precizie (sursă [ARGOS 06]): (a) o singură metodă rulată pentru mai multe valori ale parametrilor de reglaj, (b) mai multe metode şi reglaje diferite ale parametrilor.

În concluzie, punctul cheie al procesului de evaluare îl constituie disponibilitatea unei segmentări de referinţă. Datorită vastei diversităţi de secvenţe de imagini disponibile, este aproape imposibilă constituirea manuală a unei referinţe pentru fiecare secvenţă în parte. Mai mult, precizia algoritmului de detecţie este dependentă de genul secvenţei folosite, astfel, o secvenţă cu un conţinut bogat în efecte vizuale va fi mai probabil să declanşeze un număr

mai mare de false detecţii decât o secvenţă cvasi-uniformă. Soluţia constă în validarea detecţiei folosind o bază de secvenţe de test, special constituită, ce dispune de "realitate de teren". Din păcate, datorită constrângerilor de "drept de autor" cu care se confruntă fiecare secvenţă, o astfel de bază nu este încă disponibilă pentru publicul larg, aproape fiecare metodă existentă fiind testată pe o bază de secvenţe particulară.

Totuşi, notabile sunt eforturile depuse de o serie de campanii de evaluare a algoritmilor de prelucrare video, precum campania ARGOS - "Campagne d'Evaluation d'Outils de Surveillance de Contenus Vidéo" [ARGOS 06] sau campania TRECVID - "Video Retrieval Evaluation" [Trecvid 08], ce au ca scop standardizarea procesului de evaluare prin constituirea unei baze unice de test cât şi a "realităţii de teren" aferente.

2.2.5 Constituirea planelor video

După cum am menţionat în partea introductivă a acestui capitol, o secvenţă de imagini este constituită prin concatenarea planelor video pe baza tranziţiilor video. Astfel, planele video constituie *unitatea structurală de bază* a secvenţei, descompunerea în plane stând la baza marii majorităţi a metodelor existente de analiză şi prelucrare a secvenţelor de imagini.

Până în acest punct al lucrării, în scopul realizării segmentării temporale a secvenţei, am prezentat diversele tehnici de detecţie existente a tranziţiilor video cel mai frecvent folosite, şi anume *tranziţiile abrupte* de tip "cut" şi *tranziţiile graduale* de tip "fade" şi "dissolve". Localizarea individuală a tranziţiilor în secvenţă nu este însă suficientă pentru a obţine descompunerea în plane, pentru aceasta trebuind luate în calcul o serie de aspecte practice [Ionescu 05a].

Detecţia tranziţiilor de tip "cut", raportată la detecţia celorlalte tranziţii, este cea mai probabilă să returneze un număr important de false detecţii. Acest lucru se datorează faptului că toate celelalte tranziţii graduale sunt surse de discontinuitate ale fluxului vizual, fiind de regulă confundate cu una sau mai multe tranziţii de tip "cut". Anumite efecte sau schimbări de culoare, precum bliţul camerei foto, produc de asemenea o schimbare vizuală importantă. Acestea nu sunt schimbări de plan dar este foarte probabil să fie detectate drept schimbări abrupte de tip "cut". Evitarea acestor situaţii constă în folosirea adiţională de algoritmi de detecţie specifici acestor efecte.

Trebuie ţinut cont şi de faptul că diversele tranziţii video detectate nu sunt sincronizate între ele, algoritmii de detecţie fiind în general independenţi unul de altul. Astfel, pentru a constitui segmentarea temporală în plane video a secvenţei o primă etapă constă în sincronizarea temporală a tranziţiilor detectate prin ordonarea cronologică a acestora. Planele video vor fi definite

astfel ca fiind intervalele continue de imagini cuprinse între două tranziţii succesive ce ţin cont de următoarele posibile reguli [Ionescu 05a] (vezi Figura 2.6):

- toate tranziţiile de tip "cut" detectate în intervalul de timp ce corespunde unei tranziţii graduale sunt eliminate,

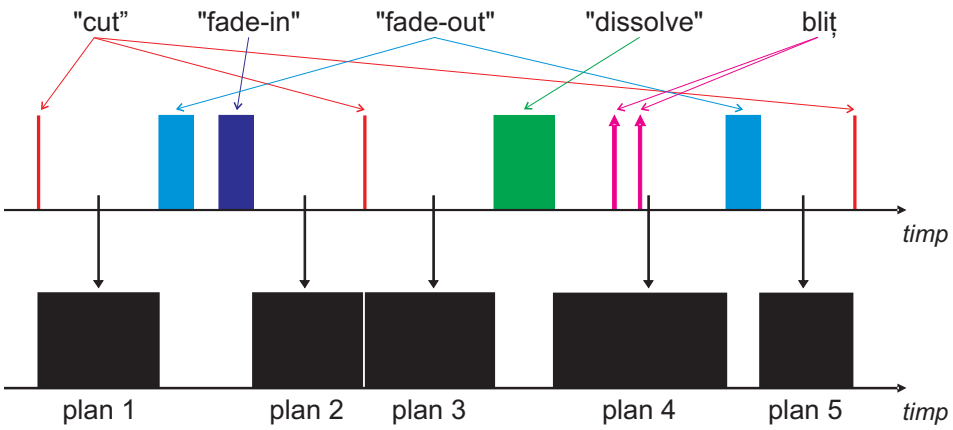


Figura 2.6: Constituirea planelor video (fiecare tip de tranziţie este reprezentat cu o culoare diferită).

- imaginile tranziţiilor video nu vor face parte din plan, fiind imagini nesemnificative pentru conţinutul acestuia,

- efectele vizuale cu proprietatea că imaginea de început este similară cu imaginea de sfârşit, (de exemplu bliţul aparatului foto din secvenţele de ştiri [Heng 99] sau efectele de tip "short color change" din filmele de animaţie [Ionescu 07c]), nu produc o schimbare de plan,

- planele video cu o durată inferioară unui anumit prag $T_{plan} \approxeq 5$ imagini (determinat empiric) pot fi eliminate deoarece nu conţin informaţie esenţială pentru secvenţă, fiind aproape neperceptibile,

- planele video cuprinse între două tranziţii de tip "fade-out" şi "fade-in", sunt eliminate dacă durata lor este inferioară unui prag T_{fade}, acestea conţinând doar imagini constante şi de regulă negre. Dacă această inserţie de imagini constante are o durată importantă, atunci poate fi considerată ca fiind un plan, având în acest caz o semnificaţie semantică şi anume introducerea unui moment de pauză în desfăşurarea acţiunii.

2.3 Detecţia scenelor video

Segmentarea temporală a unui film clasic furnizează în general între 600 şi 1500 de plane video. În diversele metode de prelucrare, pentru a reduce redundanţa temporală, strategia cel mai des folosită constă în rezumarea fiecărui plan cu una sau mai multe imagini reprezentative. Astfel, se poate obţine o reprezentare compactă a secvenţei în cel puţin 600 până la 1500 de imagini, în funcţie de numărul de plane. Volumul informaţional este însă în acest caz încă mult prea ridicat pentru a permite o analiză şi vizualizare eficientă a conţinutului secvenţei. Astfel, de multe ori este necesară o reprezentare structurală a conţinutului secvenţei pe un nivel superior planelor video. O astfel de reprezentare o constituie descompunerea secvenţei în **scene**.

Detecţia elementelor structurale de nivel semantic superior, precum scenele, nu este folosită doar pentru a evalua conţinutul secvenţei, acestea permiţând şi accesarea secvenţei la un nivel semantic. În general, durata şi frecvenţa de apariţie a scenelor constituie elemente importante pentru studiul ritmului de desfăşurare al acţiunii, sau al tehnicilor de realizare a secvenţei. De exemplu, o scenă constituită din plane de scurtă durată implică un conţinut de acţiune important. Pe de altă parte, o scenă în care durata planelor descreşte progresiv, are ca efect creşterea suspansului acţiunii.

În literatura de specialitate, scenele sunt numite frecvent şi *unităţi ale naraţiunii* sau *paragrafe video*. Folosind terminologia specifică domeniului producţiei de film, putem defini scenele ca *fiind un ansamblu redus de plane video ce sunt unificate de locul acţiunii sau de anumite evenimente de interes* [Beaver 94].

În limbaj ştiinţific, *o scenă se traduce printr-un grup de plane video ce prezintă caracteristici semantice comune*. Inspirat din teoria teatrului clasic, conţinutul unei scene trebuie să respecte regula celor trei unităţi: *unitatea de timp*, *unitatea de loc* şi *unitatea de acţiune* [Corridoni 95]. În Figura 2.7 am ilustrat un exemplu de repartiţie în scene pentru un extras dintr-un film abstract de animaţie [Ionescu 05b]. Astfel, o scenă X va conţine planele de tip X, unde $X \in \{A, B, C, D, E, F\}$. Pentru fiecare plan video am figurat doar imaginile de început şi respectiv de sfârşit. Se poate observa că o scenă va conţine acele plane video, vecine temporal, cu un conţinut similar, în care sunt prezente aceleaşi personaje şi a căror acţiune se desfăşoară în acelaşi loc.

Conceptul teoretic prin care filmul este constituit ca o combinaţie armonioasă de elemente sintactice (plane) legate între ele printr-un meta-limbaj[14]

[14]în domeniul lingvistic sau al logicii, un "meta-limbaj" este un limbaj folosit pentru a caracteriza sau pentru a face afirmaţii cu privire la alte limbaje.

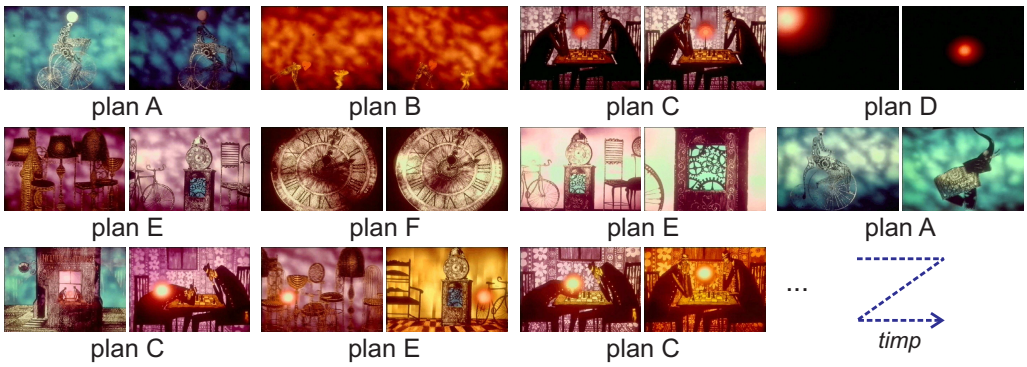


Figura 2.7: Exemplu de descompunere în scene (axa orizontală corespunde axei temporale, imagini din filmul de animaţie "Cœur de Secours" [CICA 06]).

poartă numele de paradigma de montaj [Corridoni 95]. În acest sens, putem spune că, dacă planele video sunt considerate ca fiind unităţile sintactice ale secvenţei, folosite cu predilecţie pentru înţelegerea structurală a conţinutului, atunci scenele constituie unităţile semantice ale secvenţei, folosite pentru înţelegerea la nivel perceptual a conţinutului secvenţei.

După cum am precizat în capitolul introductiv al acestei cărţi, tehnicile existente de prelucrare şi analiză a imaginilor, nu sunt încă suficient de performante pentru a permite o înţelegere semantică completă a conţinutului secvenţelor de imagini, lucru ce este strict necesar pentru o detecţie corectă a scenelor video. Cu toate acestea, tehnicile existente de detecţie încearcă să "trişeze" prin transpunerea problemei de la un nivel semantic de analiză, la un nivel de analiză reprezentat prin mărimi numerice de nivel scăzut (nivel sintactic). Astfel, conceptele semantice de unitate temporală, loc şi acţiune vor fi exprimate în termeni de similaritate între diverşi parametri de culoare, textură, formă, mişcare, etc. Revenirea de la nivelul sintactic la cel semantic se va face în acest caz pe baza expertizei "a priori" a domeniului de aplicaţie.

Metodele existente de analiză a scenelor video sunt orientate spre două direcţii de studiu distincte, şi anume:

- pe de-o parte sunt metodele de *clasare automată a scenelor* în funcţie de conţinutul semantic al acestora, cum ar fi de exemplu separarea scenelor de dialog, scenelor de vânătoare, etc. Clasele predefinte folosite sunt de regulă specifice fiecărui domeniu (ştiri, film, etc.),

- pe de altă parte, sunt metodele ce au ca scop *descompunerea în scene* sau care detectează schimbările de scenă.

Pentru un studiu bibliografic complet al problematicii de segmentare în elemente structurale de nivel semantic, cititorul se poate raporta la lucrările [Bimbo 99], [Kang 01] sau [Snoek 05b]. În cele ce urmează vom face o trecere în revistă a principalelor caracteristici ale tehnicilor folosite de metodele din cele două categorii enunţate.

2.3.1 Tehnici de clasare automată a scenelor

Obiectivul metodelor de clasare automată a conţinutului scenelor este de a grupa diversele pasaje ale secvenţei în categorii semantice predefinite. Aceasta poate fi realizată pe mai multe niveluri semantice, după cum urmează [Wang 00]:

- **nivel semantic de bază**: la acest nivel pasajele secvenţei sunt grupate în clase elementare, precum: scene filmate în exterior/interior, scene de acţiune, scene fără acţiune, etc.,
- **nivel semantic intermediar**: la acest nivel sunt detectate scenele ce conţin evenimente de interes uzual, precum scenele de dialog dintre personaje, scenele ce se derulează în anumite locaţii specifice, scenele unor evenimente artistice, etc.,
- **nivel semantic propriu-zis**: la acest nivel, pentru localizarea scenelor este necesară înţelegerea conţinutului de acţiune al secvenţei. Scenele căutate conţin în acest caz evenimente complexe, de exemplu se caută scena "uraganului apărut în Florida în 2004" sau "celebrării Noului An la baza turnului Eiffel".

Metodele propuse pentru clasarea automată a scenelor sunt foarte diverse. De exemplu, [Alatan 01] propune pentru detectarea scenelor de dialog din filme, folosirea informaţiilor obţinute în urma detecţiei şi localizării feţelor cu modele Markov ascunse. Scenele de violenţă sunt detectate în [Nam 98] pe baza analizei activităţii spaţio-temporale a secvenţei, prezenţei în imagine a urmelor de sânge şi a flăcărilor, precum şi a analizei schimbărilor de energie ale semnalului audio.

Abordarea propusă în [Saraceno 98] foloseşte atât informaţia audio cât şi vizuală pentru a clasa scenele în categoriile următoare: scene de dialog, scene de naraţiune, scene de acţiune şi scene generice. Pentru aceasta, secvenţele sunt mai întâi descompuse în plane video şi audio. Planele audio sunt detectate pe baza analizei energiei semnalului audio. Acestea sunt încadrate într-una dintre categoriile următoare: pasaje fără sunet, pasaje de vorbire, pasaje de muzică sau zgomot.

Descompunerea în plane video este realizată pe baza informaţiei de culoare. Folosind o metodă de cuantificare vectorială, în planele obţinute anterior sunt localizate modele de blocuri de pixeli ce sunt specifice scenelor vizate. Mai departe, planele sunt grupate în funcţie de caracteristicile audio-vizuale comune în grupuri omogene.

Detecţia propriu-zisă a scenelor vizate este realizată pe bază de reguli. De exemplu, pe parcursul unei scene de dialog semnalul audio conţine cu predilecţie pasaje de vorbire iar planele audio respectă un model de tip $ABABAB$, unde A şi B reprezintă două modele diferite de plane audio. Pe parcursul unei scene de acţiune, semnalul audio nu conţine pasaje de vorbire iar informaţia vizuală are o evoluţie progresivă după un model de tip $ABCDE$, unde literele desemnează modele de plane video diferite.

Un alt exemplu este metoda propusă în [Lienhart 99b] ce localizează în secvenţă scenele ce au caracteristici sonore similare, scenele filmate în locuri similare şi scenele de dialog. Metoda propusă presupune patru etape de analiză. În prima etapă, secvenţa este descompusă în plane prin detecţia tranziţiilor video de tip "cut", "fade" şi "dissolve". Mai departe, sunt extrase o serie de caracteristici la nivel de sunet, culoare, orientare a contururilor precum şi a feţelor persoanelor prezente în imagine. Etapa următoare constă în calcularea unei măsuri de distanţă între fiecare două plane video. Aceasta va fi adaptată la fiecare categorie de informaţii disponibile.

Distanţele astfel obţinute sunt organizate sub forma unui tabel pe baza căruia planele vor fi grupate în scene în funcţie de valoarea de distanţă obţinută. De exemplu, pentru a detecta scenele ce conţin pasaje audio similare, numite şi "secvenţe audio", măsura de distanţă folosită este distanţa minimală dintre vectorii de caracteristici audio ai planelor. Astfel, planele vecine temporal ce au valori ale distanţelor inferioare unui anumit prag (fiind astfel apropiate de distanţa minimă), sunt regrupate în aceeaşi secvenţă audio. Autorii pretind că folosirea independentă a diverselor modalităţi ale secvenţei conduce la obţinerea de rezultate mai performante decât în cazul fuzionării acestora.

Din punct de vedere global, metodele de clasare automată a scenelor se bazează pe categorii predefinite, neexistând în acest moment metode ce permit localizarea oricărui tip de scenă fără a utiliza un minim de expertiză a conţinutului secvenţei.

Mai mult, algoritmii de detecţie sunt specifici fiecărui tip de scenă în parte, nefiind aplicabili pentru a detecta şi alte categorii de scene. De exemplu, un algoritm de detecţie al scenelor de dialog ce foloseşte de regulă localizarea feţelor şi detecţia pasajelor de vorbire, nu poate fi folosit pentru a detecta o scenă de vânătoare.

2.3.2 Tehnici de descompunere în scene

Metodele de descompunere în scene nu caută să claseze conţinutul scenelor ci dimpotrivă să reprezinte secvenţa la un nivel structural superior descompunerii în plane video (vezi Figura 2.1). În general, metodele de descompunere în scene sunt independente de aplicaţie sau de genul secvenţei analizate.

Pentru a localiza scenele video dintr-o secvenţă, [Aigrain 95] defineşte un set de reguli globale ce trebuiesc îndeplinite de acestea. Regulile propuse sunt suficient de invariante la genul secvenţei şi se folosesc de următoarele informaţii:

- modalitate în care sunt **inserate tranziţiile video** graduale între tranziţiile de tip "cut",
- **distanţa dintre planele video similare**: o schimbare de scenă este detectată atunci când se localizeaza în secvenţă un plan similar cu planul curent analizat ce se află la o distanţă de numai 2 sau 3 plane. Similaritate între plane este exprimată ca distanţă între imagini cheie de luminanţă ce sunt extrase din fiecare plan,
- **similaritatea planelor vecine**: continuitatea conţinutului planelor video este detectată pe baza unei măsuri de similaritate, notată S. Aceasta este calculată în funcţie de valoarea medie, $m()$, şi de dispersia, $\sigma()$, a saturaţiei şi respectiv a nuanţei de culoare a pixelilor din anumite imagini cheie ale fiecărui plan. Mărimea S este dată de relaţia:

$$S_{i,j} = |m(i) - m(j)| + |\sigma(i) - \sigma(j)| \tag{2.40}$$

  unde i şi j reprezintă indicii planelor.

- **ritmul secvenţei**: o schimbare de scenă este detectată numai dacă planul curent analizat are o durată de cel puţin trei ori mai mare decât o durată de referinţă, L, sau de patru ori mai mică decât L. Durata de referinţă L este estimată pe baza unui model autoregresiv[15] astfel:

$$L_n = a \cdot L_{n-1} + b \cdot L_{n-2} \tag{2.41}$$

  unde L_n este valoarea lui L pentru iteraţia n iar parametrii a şi b sunt estimaţi în ferestre de 10 plane, variabile temporal.

[15] un model autoregresiv, sau AR, este un predictor liniar ce încearcă să estimeze ieşirea unui sistem la un moment dat, pe baza valorilor anterioare de ieşire şi respectiv intrare ale acestuia.

- **prezenţa pasajelor muzicale** după un pasaj lipsit de sunet,
- **similaritatea mişcării camerei video**: seriile temporale de cel puţin trei plane video ce conţin aceleaşi mişcări ale camerei video sunt considerate ca aparţinând aceleiaşi scene.

Un alt exemplu este metoda propusă în [Huang 98] ce realizează o segmentare ierarhică a secvenţei. Astfel, schimbările de plan şi respectiv de scenă sunt detectate la diverse niveluri de detaliu. Schimbările de scenă sunt asociate schimbărilor simultane a culorilor, mişcării şi a caracteristicilor semnalului sonor. Metoda propusă în [H. Sundaram 00] modelează relaţiile de coerenţă dintre plane folosind un model de memorie cauzală de tip FIFO[16].

O altă abordare constă în rezumarea conţinutului fiecărui plan video cu imagini "mozaic"[17] [Aner 01]. Astfel, planele sunt grupate în scene pe baza analizei similarităţii dintre imaginile de tip "mozaic" asociate acestora. Metoda propusă în [Ionescu 05b] testează utilizarea histogramelor color medii clasice şi respectiv ponderate, pentru a măsura similaritatea dintre planele video. Astfel, două plane sunt considerate ca aparţinând aceleiaşi scene dacă au un conţinut de culoare similar (valoare a distanţei Euclidiene dintre histograme redusă) şi sunt vecine temporal.

O abordare complexă inedită a problematicii descompunerii în scene este metoda hibridă propusă în [Kang 01]. Aceasta foloseşte pentru detecţie colaborarea mai multor tehnici de detecţie. Segmentarea ierarhică în scene este realizată în trei etape.

Prima etapă constă într-o *segmentare iniţială* în scene a secvenţei. Metoda folosită este o metodă de detecţie continuă. Gradul de coerenţă dintre diversele plane este calculat folosind modelul de memorie cauzală de tip FIFO propus în [H. Sundaram 00]. Valoarea de coerenţă este o valoare continuă în timp, ce este exprimată în funcţie de valoarea parametrului de "recall". Valoarea acestuia pentru două plane A şi respectiv B, notată $recall(A, B)$, este calculată la momentul S_a (punctul de pornire al analizei din memoria tampon) folosind ecuaţia:

$$recall(A, B) = (1 - dissim(A, B)) \cdot L_A \cdot L_B \cdot (1 - \frac{\Delta n}{N_m}) \cdot (1 - \frac{\Delta t}{T_m}) \quad (2.42)$$

unde $dissim(A, B)$ este o măsură de disimilaritate a planelor A şi B, L_A şi L_B reprezintă durata planelor A şi respectiv B ce este ponderată de coeficienţi

[16]FIFO este acronimul pentru "First-In-First-Out" şi reprezintă o modalitate de organizare şi manipulare a datelor ce ţine cont de timp şi de prioritate. Aceasta lucrează pe principiul unei stive de date şi soluţionează conflictele generate de manipularea datelor folosind regula "primul sosit este şi primul servit".

[17]vezi explicaţia de la pagina 150.

ce iau în calcul următoarele valori: T_m-dimensiunea memoriei tampon, N_m-numărul total de plane conţinute în memorie, Δn-numărul de plane dintre planele A şi B, Δt-distanţa temporală dintre planele A şi B.

Valoarea de coerenţă, $Co(S_a)$, la momentul S_a, este dată de relaţia:

$$Co(S_a) = \frac{recall(A, B)}{R_{max}(S_a)} \tag{2.43}$$

unde $R_{max}(S_a)$ reprezintă valoarea maximală a parametrului de "recall" ce este obţinută în cazul particular în care $dissim(A, B) = 0$.

Astfel, schimbările de scenă sunt detectate pe baza analizei valorilor funcţiei de coerenţă într-o fereastră de decizie de dimensiune predefinită. O schimbare de scenă este detectată când un minim local al acesteia se găseşte în mijlocul ferestrei considerate.

A doua etapă de analiză se foloseşte de această dată de o *abordare discretă*, şi este folosită pentru a rafina rezultatele descompunerii în scene obţinute anterior. Scopul acesteia este de a ameliora detecţia prin reducerea numărului de scene rămase nedetectate. Astfel, planele conţinute de scenele din prima etapă vor fi regrupate pe baza analizei unui anumit set de imagini cheie. Regruparea lor se va face cu un clasificator k-means pentru care numărul optimal de clase a fost ales pe baza unei metode de analiză a pertinenţei claselor[18]. În urma clasificării, fiecare plan va dispune de o anumită etichetă corespunzătoare clasei din care face parte, transformând astfel scenele în serii de etichete. Planele ce au aceeaşi etichetă vor fi asociate între ele prin crearea unei corespondenţe. Dacă în interiorul unei scene găsim două plane ce nu sunt corespondente, atunci scena va fi scindată în acel punct (vezi Figura 2.8, etichetele diverselor plane au fost figurate cu litere majuscule iar relaţiile de corespondenţă sunt marcate cu săgeţi duble).

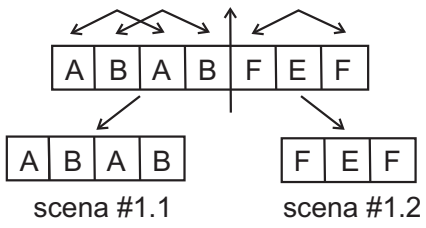


Figura 2.8: Exemplu de rafinare a detecţiei [Kang 01].

Ultima etapă de analiză constă în ajustarea segmentării în scene obţinută până în acest punct. Aceasta, are ca scop corectarea falselor detecţii.

[18]pentru o descriere detaliată a algoritmului k-means, vezi Secţiunea 7.1.2.

Mai întâi planele scenelor adiacente sunt regrupate pe baza unui clasificator k-means folosind acelaşi principiu ca în etapa precedentă. Falsele schimbări de scenă sunt apoi corectate pe baza examinării corespondenţelor dintre planele diverselor scene.

Din punct de vedere global, metodele existente de descompunere în scene, transformă problema detecţiei într-o problemă de similaritate între conţinutul planelor video, sau la un nivel mai general, într-o problemă de clasificare a conţinutului acestora. În acest context, scenele sunt văzute ca ansambluri de plane, vecine temporal, ce au proprietăţi similare ale conţinutului de culoare, mişcare, etc. Definirea unei măsuri eficiente de similaritate între plane joacă în acest caz un rol decisiv în descompunerea corectă în scene a secvenţei. Cu toate că o astfel de abordare este încă departe de sensul semantic pe care îl au scenele, aceasta constituie totuşi un pas important spre înţelegerea automată a conţinutului secvenţei.

Pe lângă clasarea automată după conţinut şi decupajul secvenţei în scene, tehnicile de analiză a scenelor video îşi găsesc aplicaţii şi în problematici conexe din analiza secvenţelor de imagini. În cele ce urmează vom prezenta câteva dintre acestea.

2.3.3 Aplicaţii ale analizei scenelor video

Detecţia de scene, şi în particular analiza similarităţii planelor video, permite detecţia unei tehnici de filmare particulare cunoscută sub numele de **"shot-reverse-shot"** [Bimbo 99]. Această tehnică este folosită frecvent în pasajele video cu un conţinut cu predilecţie static, precum scenele de conversaţie dintre personaje, anumite pasaje ale secvenţelor documentare culturale sau în scenele din sporturile de interior, precum jocul de biliard, tenisul de masă, etc.

Un "shot-reverse-shot" presupune folosirea alternativă a mai multor camere video. Un exemplu clasic este filmarea unui meci de biliard în care camerele video filmează alternativ jucătorul, masa de biliard şi publicul (vezi Figura 2.9).

Figura 2.9: Exemplu de tehnică "shot-reverse-shot" dintr-un meci de biliard (pentru fiecare plan am ilustrat doar o singură imagine reprezentativă).

Tehnica "shot-reverse-shot" introduce un fel de periodicitate a conţinutului secvenţei. Structura de plane va avea o evoluţie de tip $ABABAB$, unde A şi B denotă două plane diferite din punct de vedere al conţinutului. Analizând distanţa, în sens de similaritate, a planului curent faţă de planele vecine (tehnică folosită la detectarea scenelor), putem localiza uşor semnătura temporală specifică unui "shot-reverse-shot", şi anume o succesiune de valori de tip valoare redusă, valoare semnificativă, valoare redusă. Interesul asupra acestei tehnici de filmare nu se rezumă doar la detecţia propriu-zisă, ci are şi o aplicaţie în caracterizarea semantică a conţinutului. Aceasta frunizează informaţii preţioase despre natura secvenţei, despre conţinutul de acţiune al acesteia precum şi despre ritmului de desfăşurare al secvenţei.

Detecţia scenelor poate fi utilă şi în **cazul segmentării temporale** a secvenţei, în particular pentru detecţia tranziţiilor abrupte de tip "cut" (vezi Secţiunea 2.2.1). Metodele de detecţie de "cut" existente se confruntă cu problema falselor detecţii, ce se concretizează prin asimilarea uneia sau a mai multor schimbări vizuale drept tranziţii de tip "cut". Acestea vor produce global o supra-segmentare temporală a secvenţei prin scindarea artificială a planelor video în mai multe plane. Din punct de vedere al detecţiei de scene pe baza analizei similarităţii dintre plane, aceste situaţii se traduc prin scene compuse din plane succesive temporal. Astfel, dacă două plane succesive sunt găsite ca aparţinând aceleaşi scene, atunci este foarte probabil ca acestea să fie rezultatul unei supra segmentări, fuzionarea acestora într-unul singur corectând această problemă. Detecţia scenelor în acest caz, furnizează informaţii suplimentare metodelor clasice de detecţie a tranziţiilor video, datorită faptului că măsurile de disimilaritate vizuală sunt calculate de această dată la nivel de segment şi nu la nivel de cadru. Totuşi, detecţia scenelor nu poate ţine loc de detecţie a schimbărilor de plan, ci trebuie văzută ca o etapă de rafinare a rezultatelor detecţiei.

Descompunera în scene video îşi găseşte aplicaţie şi în metodele de **generare automată a rezumatelor de conţinut**, permiţând accesarea conţinutului secvenţei pe mai multe niveluri de detaliu. Rezumatul unei secvenţe video este definit ca fiind o reprezentare compactă a conţinutului acesteia (vezi Capitolul 5). Metodele existente de rezumare folosesc ca informaţie de plecare segmentarea temporală în plane a secvenţei. Astfel, principiul rezumării constă în rezumarea planelor video cu un anumit număr de imagini reprezentative, numite şi imagini cheie. Totuşi, acest tip de abordare nu permite utilizatorului să aleagă nivelul de detaliu dorit pentru rezumat, fiind limitată în a furniza informaţii la nivel de plan.

Pe baza analizei similarităţii planelor, folosită de descompunerea în scene, se poate construi o reprezentare compactă a secvenţei pe mai multe niveluri de detaliu. Principiul este ilustrat în Figura 2.10. Astfel un prim nivel de

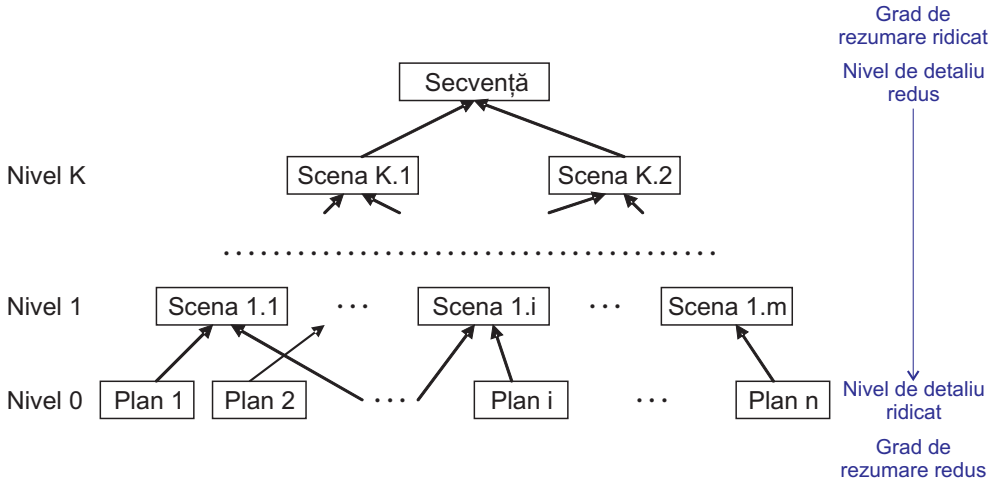


Figura 2.10: Reprezentare ierarhică a conţinutului secvenţei pe baza regrupării în scene.

detaliu il constituie nivelul planelor video. Acesta este urmat în mod natural de nivelul scenelor video. Nivelurile de detaliu superioare ierarhic sunt obţinute pe baza regrupării scenelor similare. Pentru aceasta se poate folosi acelaşi mecanism ca pentru determinarea scenelor iniţiale. Procesul poate fi repetat iterativ până se ajunge la cel mai înalt nivel ierarhic, reprezentat de secvenţa în totalitate.

Folosind principiul dendogramei[19], această reprezentare ierarhică a conţinutului permite rezumarea secvenţei în funcţie de durata rezumatului dorit, cât şi a cantităţii de informaţie furnizată de acesta. De exemplu, secvenţa poate fi rezumată cu o singură imagine cheie pentru fiecare scenă din nivelul k, obţinând pentru $k = 0$, n imagini, pentru $k = 1$, m imagini, cu $m << n$, până la $k = K$, unde obţinem doar două imagini (vezi Figura 2.10). Astfel, în funcţie de cerinţele aplicaţiei, putem opta între un rezumat cu un grad de detaliu ridicat, dar de o durată mai semnificativă, sau pentru un nivel de detaliu mai scăzut, dar cu o durată mult mai redusă (în acest caz rezumatul poate conţine doar câteva imagini).

Tot în cazul tehnicilor de rezumare, detecţia scenelor video poate fi folosită la rafinarea conţinutului rezumatelor dinamice. Spre deosebire de rezumatele ce folosesc un anumit număr de imagini cheie, rezumatele dinamice furnizează o serie de pasaje ale secvenţei ce sunt considerate ca fiind reprezentative pentru conţinut. Problema care apare este redundanţa vizuală a planelor

[19] o dendogramă (în greceşte dendron - arbore şi gramma - a desena) este o reprezentare sub formă arborescentă a claselor obţinute în urma unui algoritm de clasificare.

din aceeaşi scenă video. Astfel, localizarea scenelor prin analiza similarităţii planelor, poate fi folosită pentru a localiza şi elimina planele redundante, cum este cazul metodei propuse în [Laganière 08].

2.4 Concluzii

În acest capitol am discutat problematica segmentării temporale a secvenţelor de imagini. Aceasta stă la baza majorităţii metodelor de analiză şi prelucrare video.

Segmentarea temporală constă în descompunerea secvenţei în plane video pe baza detecţiei tranziţiilor video. Dintre tranziţiile existente se remarcă:

- tranziţiile abrupte de tip "cut", ce reprezintă concatenarea directă a două plane succesive, fiind şi tranziţiile cel mai frecvent utilizate,
- tranziţiile graduale, precum tranziţiile de tip "fade" şi "dissolve", ce reprezintă o trecere progresivă de la un plan la altul, având un procent de apariţie cu cel puţin un ordin de măsură inferior tranziţiilor de tip "cut".

În ceea ce priveşte metodele folosite, direcţia de studiu cea mai profitabilă pentru detecţia tranziţiilor video (atât abrupte cât şi graduale), se dovedeşte a fi analiza evoluţiei intensităţii pixelilor din imagine, aceasta surclasând la nivel de performanţe celelalte metode, precum metodele ce folosesc analiza de contur sau analiza mişcării. Din punct de vedere al performanţelor metodelor existente, de remarcat este faptul că cele mai precise sunt metodele de detecţie a tranziţiilor abrupte de tip "cut", algoritmii existenţi furnizând în medie o precizie de detecţie de peste 95%. Pe de altă parte, în cazul detecţiei tranziţiilor graduale de tip "fade" şi "dissolve", precizia de detecţie este inferioară, în medie situându-se undeva în jurul valorii de 75%. Acest lucru se datorează în principal modificărilor complexe ale scenei realizate de tranziţiile graduale.

Detecţia tranziţiilor video are pe lângă rolul de a furniza descompunerea temporală în plane a secvenţei şi un aport în caracterizarea semantică a conţinutului. Fiecare tip de tranziţie video în parte, este folosit cu un scop precis. De exemplu, tranziţiile de tip "cut" sunt schimbări bruşte de scenă şi astfel sunt folosite pentru a face tranziţia rapidă de la un moment al scenei la altul. Tranziţiile de tip "dissolve" sunt tranziţii lente şi sunt folosite de regulă pentru a schimba timpul acţiunii sau tranziţiile de tip "fade" ce introduc o pauză în desfăşurarea acţiunii secvenţei.

Pe de altă parte, cunoaşterea structurii temporale a secvenţei oferă informaţii şi referitor la ritmul de desfăşurare al acţiunii. Astfel, pasajele secvenţei

bogate în schimbări de plan, reflectă un conţinut bogat în acţiune şi totodată un ritm alert, pe când planele video de lungă durată implică o rată de schimbare de plan redusă şi astfel un ritm lent. Analiza cadenţei schimbărilor de plan este utilizată cu succes de metodele existente pentru caracterizarea conţinutului de acţiune al secvenţei.

Dacă planele video sunt considerate ca fiind unităţile sintactice ale secvenţei, fiind folosite cu predilecţie pentru analiza de nivel scăzut a conţinutului, scenele video sunt unităţile semantice ale secvenţei ce permit o înţelegere de nivel semantic superior a conţinutului. O scenă video este constituită dintr-un grup de plane video ce respectă unitatea de timp, de loc şi de acţiune. Din păcate nivelul ştiinţific actual nu permite implementarea de metode automate capabile să înţeleagă în totalitate conţinutul secvenţelor, lucru necesar la detectarea scenelor. Din acest motiv, metodele existente se limitează cu predilecţie doar la analiza similarităţii dintre planele video.

În concluzie, descompunerea secvenţei în unităţi temporale, fie că este vorba de plane sau de scene, reprezintă o etapă de analiză necesară şi totodată premergătoare analizei conţinutului video. Aceasta oferă informaţii despre modalitatea în care a fost constituită secvenţa precum şi despre modul de desfăşurare al acţiunii, fiind într-un fel etapa inversă procesului de realizare a acesteia ce are loc în studio.

CAPITOLUL 3

Analiza mișcării

Rezumat: *Evoluţia temporală a informaţiei vizuale este una dintre particularităţile fundamentale a secvenţelor de imagini. Din acest punct de vedere, secvenţele de imagini mai sunt denumite şi imagini în mişcare, fiind constituite ca o succesiune de evoluţii temporale a conţinutului unor imagini fixe. Pornind de la problematica estimării mişcării la nivel de pixel, în acest capitol vom face o trecere în revistă a diverselor direcţii de studiu abordate de metodele de analiză şi caracterizare a conţinutului de mişcare din secvenţele de imagini.*

Una dintre particularităţile de bază a secvenţelor de imagini o constituie informaţia de mişcare. Raportat la imaginile statice, secvenţele de imagini oferă o evoluţie temporală a conţinutului uneia sau a mai multor imagini fixe. În acest sens, secvenţele de imagini sunt denumite şi *imagini în mişcare*.

Dacă în acest moment motoarele de căutare folosite de sistemele de indexare actuale, permit accesul rapid şi eficient la informaţiile textuale, nu putem spune acelaşi lucru şi despre accesarea conţinutului multimedia. Pentru a compensa această lipsă, şi anume existenţa unui mecanism eficient de accesare a datelor multimedia, dintre care în special a informaţiilor video, grupul de dezvoltare al standardului de codare video MPEG ("Moving Picture Experts Group") lucrează la dezvoltarea şi ameliorarea unui nou standard cunoscut sub numele de MPEG-7. După cum precizează creatorii aces-

tuia, "principala ambiţie a lui MPEG-7 este de a face conţinutul informaţiilor multimedia să fie la fel de uşor de accesat pe Internet precum informaţiile textuale". În ceea ce priveşte informaţia de mişcare, standardul MPEG-7 selectează şi integrează *unele dintre cele mai performante metode existente* de analiză a mişcării. Astfel, tehnicile existente se grupează în două categorii principale [Jeannin 01]:

- pe de-o parte sunt metodele de analiză globală, bazate pe **analiza mişcării globale** a camerei video. În acest caz, analiza mişcării este realizată la nivel de segment video (pasaj al secvenţei). Dintre aplicaţiile analizei de mişcare globală putem enumera: recunoaşterea mişcării camerei video, detecţia activităţii de mişcare sau generarea imaginilor de tip "mozaic".

- pe de altă parte sunt metodele de analiză locală, ce sunt bazate pe **analiza mişcării obiectelor** din scenă. Acestea analizează mişcarea la nivel de regiuni spaţiale de pixeli din imagine. De regulă, analiza locală este folosită pentru segmentarea şi urmărirea temporală a obiectelor în mişcare.

Aceste două direcţii de studiu sunt sintetizate în Figura 3.1. În cele ce urmează vom face o trecere în revistă a tehnicilor folosite de fiecare dintre acestea.

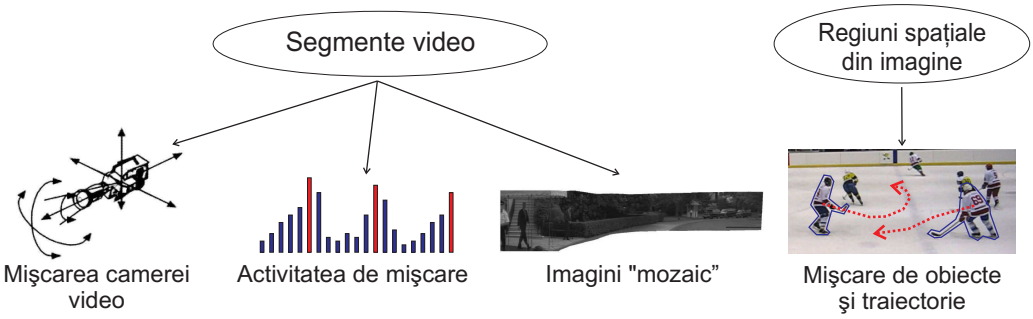


Figura 3.1: Principalele direcţii de analiză a mişcării în secvenţele de imagini: nivel global (segment) şi nivel local (regiune) (sursă standard MPEG-7 [Jeannin 01]).

Mişcarea globală. Analiza mişcării globale a scenei este realizată la nivel de segment video sau de grup de imagini. O primă informaţie extrasă din secvenţă este *tipul mişcării camerei video*, ca de exemplu: mişcare translaţională, mişcare de rotaţie, mişcare de apropiere etc. (vezi Secţiunea 3.2).

Informaţiile reţinute în acest caz pentru o anumită categorie de mişcare sunt de regulă amplitudinea mişcării, durata mişcării precum şi localizarea acesteia în secvenţă. Analiza mişcării camerei video este importantă deoarece permite în anumite situaţii înţelegerea conţinutului secvenţei prin identificarea anumitor pasaje de interes din aceasta. De exemplu, focalizarea asupra unui anumit personaj se traduce printr-o mişcare a camerei de tip "zoom-in" (mărire), sau, creşterea suspansului acţiunii poate fi marcată de o mişcare de translaţie foarte rapidă.

O altă informaţie exploatată este *activitatea de mişcare*. Aceasta este o măsură a percepţiei vizuale pe care o avem asupra mişcării conţinute în secvenţă. Activitatea de mişcare este determinată pe baza clasificării mişcării globale în funcţie de o serie de parametri de nivel scăzut (de exemplu, dispersia amplitudinii vectorilor de mişcare). Clasificarea este realizată pe mai multe niveluri de activitate, în funcţie de intensitatea acţiunii. La clasificare este luată în calcul şi situaţia în care acţiunea este absentă, aceasta reprezentând nivelul minim de activitate. Un nivel de activitate intens corespunde evenimentelor dinamice, ca de exemplu scenele de gol din secvenţele de fotbal sau scenele de urmăriri de maşini din secvenţele de ştiri. Pe de altă parte, un nivel de activitate redus corespunde scenelor cu un conţinut "sărac" în mişcare, ca de exemplu scenele de dialog dintre personaje sau scenele de interviu din secvenţele de ştiri sau documentare.

Tot pe baza analizei mişcării globale este şi *construcţia imaginilor de tip "mozaic"* [Aner 01]. O imagine de tip "mozaic" este o imagine statică ce rezumă conţinutul de mişcare global al unui pasaj al secvenţei (de regulă un plan video). Aceasta este realizată prin regruparea şi suprapunerea diverselor imagini ale segmentului, după recalarea geometrică în funcţie de deplasarea globală a scenei (vezi Figura 5.2 de la pagina 150). Imaginile de tip "mozaic" sunt folosite drept rezumate compacte ale diverselor pasaje ale secvenţei şi de regulă au o complexitate de calcul ridicată. Totuşi, aceasta poate fi redusă prin folosirea parametrilor de deformare furnizaţi de standardul MPEG-7.

Mişcarea locală. Analiza mişcării locale sau a deplasării obiectelor, este efectuată la nivel de regiuni de pixeli. Dacă pentru caracterizarea globală a mişcării, vectorii de mişcare puteau fi estimaţi la un nivel de detaliu mai redus (de exemplu, la nivel de blocuri de pixeli), furnizând astfel o aproximaţie grosieră a fluxului optic, în cazul analizei mişcării locale a obiectelor, vectorii de mişcare sunt estimaţi de regulă la nivel de pixel pentru obţinerea unui nivel de detaliu ridicat. Pentru analiză, metodele existente folosesc de regulă o modelare parametrică a mişcării. Aceasta permite localizarea în secvenţă a obiectelor cu deplasări similare, în ciuda diverselor deformări geometrice suportate de acestea. În general, rezultatul analizei mişcării obiectelor este

cuantificat prin furnizarea traiectoriei acestora sub formă de evoluţie temporală a anumitor puncte de interes, precum centrul de greutate sau anumite puncte de contur.

În contextul indexării după conţinut a secvenţelor de imagini, metodele de analiză a traiectoriei obiectelor ("object tracking") sunt cu mult mai studiate decât metodele de analiză globală a mişcării camerei video. Acest lucru se datorează în principal faptului că într-o secvenţă marea majoritate a evenimentelor de interes implică, şi sunt legate, de mişcarea obiectelor. De exemplu, într-o secvenţă sportivă, va fi mult mai interesant şi totodată reprezentativ pentru analiză să dispunem de traiectoria unui anumit jucător care este într-o acţiune de atac, decât să caracterizăm mişcarea globală a camerei video ce urmăreşte jucătorul. Pentru un studiu biliografic complet al tehnicilor de analiză a mişcării obiectelor, cititorul se poate raporta la studiile [Dagtas 00], [Fablet 02] sau [Smith 04].

În concluzie, toate metodele existente de analiză a mişcării, fie că este vorba de mişcare globală sau locală, folosesc ca punct de plecare *estimarea mişcării*. Aceasta, pe baza măsurării deplasării pixelilor, sau a regiunilor de pixeli, de la un cadru la altul, furnizează un câmp vectorial de mişcare. În cele ce urmează vom face o trecere în revistă a tehnicilor de estimare a mişcării existente.

3.1 Estimarea mişcării

Principiul estimării de mişcare constă în determinarea deplasării unui pixel, sau a unui bloc de pixeli, între două imagini succesive ale secvenţei, pe baza minimizării variaţiei intensităţii acestuia, numită şi DFD sau "Displaced Frame Difference". Această variaţie poate fi reprezentată sub forma următoare:

$$DFD(\vec{r}, \vec{d}, \Delta t) = I(\vec{r} + \vec{d}, t + \Delta t) - I(\vec{r}, t) \tag{3.1}$$

unde $\vec{r}$ reprezintă poziţia pixelului sau a blocului de pixeli în imaginea analizată, $\vec{d}$ reprezintă vectorul de deplasare între momentele t şi $t + \Delta t$ exprimat în funcţie de deplasarea pe cele două axe, oX şi respectiv oY, $\vec{d} = (dx, dy)$, iar $I(t)$ reprezintă imaginea la momentul t.

Acest principiu de estimare se bazează pe ipoteza conform căreia intensitatea pixelilor nu variază semnificativ de la o imagine la alta. Un exemplu de vectori de mişcare obţinuţi la nivel de blocuri de pixeli este prezentat în Figura 3.2. Secvenţa folosită, pentru care am ilustrat câteva imagini reprezentative, conţine o deplasare a camerei video către dreapta, de aici

şi orientarea predominantă a vectorilor de mişcare spre stânga (imaginea se deplasează în sens invers camerei video).

Figura 3.2: Exemplu de vectori de deplasare (axa oX este axa temporală, vectorii indică direcţia şi amplitudinea deplasării blocurilor de pixeli, sursă imagini: filmul "The Wicker Man", Copyright 2006 Warner Bros Pictures).

Dacă considerăm imaginea ca fiind o funcţie continuă, atunci putem folosi descompunerea în serie Taylor de ordinul întâi, astfel:

$$I(\vec{r}+\vec{d},t+\Delta t) = I(\vec{r},t) + \frac{\partial I(\vec{r},t)}{\partial x}\cdot dx + \frac{\partial I(\vec{r},t)}{\partial y}\cdot dy + \frac{\partial I(\vec{r},t)}{\partial t}\cdot dt \quad (3.2)$$

şi mai departe folosind ecuaţia 3.1, obţinem:

$$DFD(\vec{r},\vec{d},\Delta t) = \frac{\partial I(\vec{r},t)}{\partial x}\cdot dx + \frac{\partial I(\vec{r},t)}{\partial y}\cdot dy + \frac{\partial I(\vec{r},t)}{\partial t}\cdot dt \quad (3.3)$$

Minimizând funcţia DFD obţinem ecuaţia fluxului optic:

$$\frac{\partial I(\vec{r},t)}{\partial x}\cdot u + \frac{\partial I(\vec{r},t)}{\partial y}\cdot v + \frac{\partial I(\vec{r},t)}{\partial t} = 0 \quad (3.4)$$

unde $(u,v) = (\frac{dx}{dt},\frac{dy}{dt})$ defineşte vectorul de deplasare în imagine. Fluxul optic astfel definit va permite estimarea mişcării doar în direcţia gradientului spaţial.

În general, metodele de estimare existente folosesc algoritmi de minimizare a unei funcţii de cost ce este determinată pe baza funcţiei DFD de deplasare a pixelilor sau a blocurilor de pixeli. Astfel întâlnim mai multe abordări [Marichal 98]:

- **metodele diferenţiale**: sunt bazate pe ecuaţia fluxului optic (vezi ecuaţia 3.4) şi au ca rezultat un câmp vectorial de mişcare dens. Metodele diferenţiale sunt foarte sensibile la prezenţa zgomotului în imagine, acesta diminuând considerabil precizia estimării. De asemenea, având o complexitate de calcul importantă, timpul necesar estimării este de regulă semnificativ.

- **metodele parametrice**: modelează deplasarea pixelilor în imagine folosind o serie de parametri. Problema estimării mişcării este astfel transformată într-o problemă de estimare a parametrilor unui anumit model, ca de exemplu modelul afin, cuadratic, etc.

- **metodele stohastice**: folosesc modele probabilistice. Explorarea spaţiului parametrilor este ghidată în acest caz de procese aleatoare, precum modele Bayesiene, modele Markov sau algoritmi genetici. Metodele stohastice au o complexitate de calcul ridicată, dar rezultatele obţinute corespund mult mai bine realităţii.

- **metodele bazate pe blocuri de pixeli**: estimează mişcarea la nivel de blocuri de pixeli. Această abordare a fost propusă pentru prima dată în [Jain 91] şi s-a dovedit a fi cel mai bun compromis între complexitatea de calcul şi precizia estimării obţinute. Reglarea dimensiunii blocurilor de pixeli folosite la estimare permite în acest caz reglarea sensibilităţii şi a robusteţei metodei. Astfel, folosirea de blocuri de dimensiuni reduse are ca rezultat o sensibilitate ridicată a algoritmului de estimare, situaţie favorabilă în cazul micilor deplasări ale obiectelor din scenă. Pe de altă parte, folosirea de blocuri de dimensiuni mai mari conferă robusteţe metodei la prezenţa zgomotului în imagine. Datorită acestor proprietăţi, metodele de estimare pe blocuri de pixeli sunt folosite cu succes în marea majoritate a standardelor de codare video, precum H.263, MPEG 1, 2 şi 4.

Din punct de vedere global, pentru a facilita calculele, metodele existente de estimare a mişcării adoptă o serie de *ipoteze de plecare*. În realitate, acestea nu sunt întotdeauna valabile, fapt ce duce uneori la obţinerea de rezultate eronate. Dintre ipotezele cele mai importante, putem enumera următoarele:

- pixelii din blocurile de pixeli considerate, au acelaşi tip de mişcare de translaţie de la o imagine la alta,
- funcţia DFD de deplasare a pixelilor, are o evoluţie monoton crescătoare,
- intensitatea pixelilor din imagine este constantă cu mişcarea,
- mişcarea este considerată a fi constantă pentru mici volume spaţio-temporale.

În funcţie de nivelul de detaliu şi de precizia câmpului vectorial de mişcare, întâlnim două tipuri de implementări (vezi Figura 3.3). Pe de-o parte sunt implementările *multi-rezoluţie*, în care estimarea mişcării se face pentru mai multe reprezentări, de rezoluţii diferite, ale aceleaşi imagini analizate. Acestea sunt organizate sub formă piramidală, imaginea din vârful piramidei fiind imaginea cu rezoluţia cea mai mică. Diversele rezoluţii sunt obţinute în urma filtrării de tip trece-jos şi a subeşantionării spaţiale.

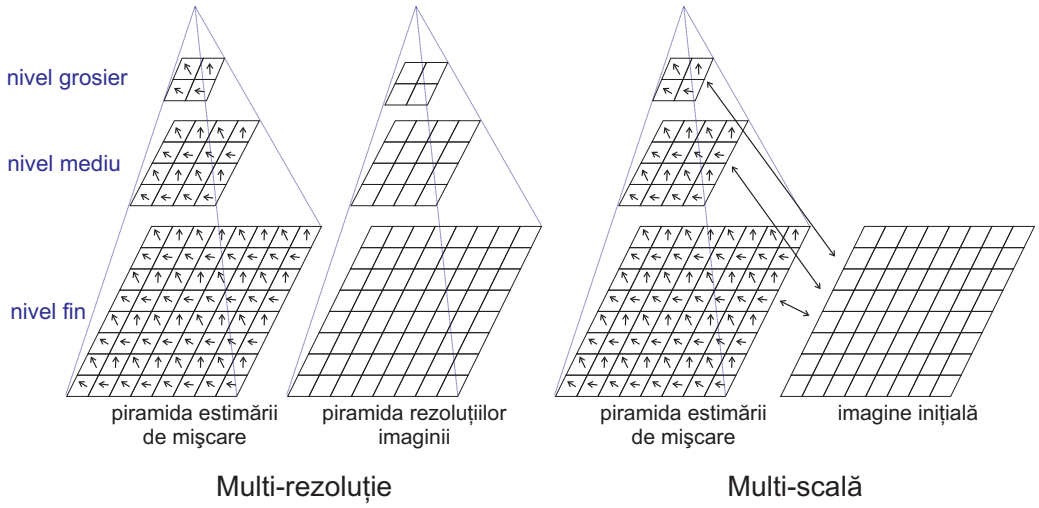


Figura 3.3: Modalităţi de implementare a estimării câmpului vectorial de mişcare (sursă [Marichal 98]).

Pe de altă parte, întâlnim implementarea *multi-scală*, ce foloseşte imaginile în rezoluţia iniţială, dar furnizează mai multe niveluri de detaliu pentru vectorii de mişcare. Ca şi în cazul implementării multi-rezoluţie, nivelurile de detaliu sunt organizate sub formă piramidală şi ordonate în funcţie de densitatea câmpului vectorial furnizat.

Din punct de vedere temporal, estimarea mişcării între două imagini succesive, poate fi realizată în două moduri:

- fie de tip *"forward"*, caz în care căutarea noii poziţii a blocului de pixeli analizat din imaginea la momentul t se face în imaginea următoare la momentul $t + 1$,
- sau de tip *"backward"*, când poziţia blocului de pixeli la momentul t este căutată în imaginea anterioară, la momentul $t - 1$.

Aceste modalităţi de estimare îşi găsesc utilitate cu predilecţie în metodele de codare video, precum standardele H.263 şi MPEG, în care este folosită codarea bidirecţională. Astfel, un bloc de pixeli poate fi obţinut atât cu predicţie anterioară ("backward") cât şi cu predicţie ulterioară ("forward").

3.1.1 Metodele diferenţiale

Metodele diferenţiale de estimare a mişcării sunt metode bazate pe *calculul gradientului*. Acestea pornesc de la ipoteza conform căreia intensitatea pixelilor este constantă cu mişcarea. În această categorie putem enumera *metodele directe*, ce folosesc ca principiu anularea valorilor gradientului funcţiei DFD ce trebuie minimizată, şi respectiv *metodele indirecte*, ce urmăresc convergenţa funcţiei DFD către o soluţie în direcţia gradientului. Din categoria metodelor indirecte putem menţiona metodele iterative şi metodele pel-recursive [Marichal 98].

Metodele iterative

Primele abordări ale problematicii estimării mişcării pe baza estimării fluxului optic sunt cele propuse în [Horn 81]. Pentru a rezolva ecuaţia 3.4 este adoptată o ipoteză suplimentară, şi anume ca modulul gradientului să aibă valori mici, condiţie valabilă în realitate doar pentru micile deplasări ale pixelilor din imagine. Astfel, problema estimării este transformată într-o problemă de minimizare a unei funcţii de cost exprimată cu ajutorul coeficienţilor Lagrange astfel:

$$\int\int \left[(I_x \cdot u + I_y \cdot v + I_t)^2 + \lambda \cdot (u_x^2 + u_y^2 + v_x^2 + v_y^2)^2\right] dx \cdot dy \qquad (3.5)$$

unde I_i reprezintă derivata parţială de ordinul întâi a imaginii I în funcţie de componenta i; u_x, u_y, v_x şi respectiv v_y reprezintă derivatele parţiale de ordinul întâi ale celor două componente, (u, v), ale fluxului optic, calculate în

direcţia oX şi respectiv oY, iar λ este multiplicatorul lui Lagrange ce reglează aportul erorii în ecuaţia de mişcare.

O soluţie posibilă este dată de ecuaţiile următoare:

$$\widehat{u} = u_m - I_x \cdot \frac{P}{D} \tag{3.6}$$

$$\widehat{v} = v_m - I_y \cdot \frac{P}{D} \tag{3.7}$$

unde u_m şi v_m reprezintă valorile medii ale lui u şi respectiv v, iar P şi D sunt doi parametri daţi de relaţiile următoare:

$$P = I_x \cdot u_m + I_y \cdot v_m + I_t \tag{3.8}$$

$$D = \lambda + I_x^2 + I_y^2 \tag{3.9}$$

Calculul fluxului optic final este realizat în mod iterativ folosind pentru rafinare metoda Gauss-Seidel[1]. Soluţia la iteraţia i, $(\widehat{u}_i, \widehat{v}_i)$, este exprimată în funcţie de soluţia de la iteraţia precedentă, $i-1$, $(\widehat{u}_i, \widehat{v}_i) = f\{(\widehat{u}_{i-1}, \widehat{v}_{i-1})\}$, soluţia finală adoptată fiind soluţia furnizată de iteraţia pentru care este îndeplinit un anumit criteriu de convergenţă.

Câmpul vectorial obţinut în acest caz *este unul dens*. Fiecare pixel din imagine va avea asociat un vector de deplasare. Complexitatea de calcul este de asemenea semnificativă. Din această cauză, acest tip de abordare nu a fost folosită în tehnicile de codare, fiind utilizată cu predilecţie în metodele de analiză a mişcării. Mai mult, datorită ipotezelor iniţiale adoptate, metodele bazate pe analiza gradientului nu furnizează rezultate precise pentru deplasări importante ale pixelilor din imagine.

Ca exemple de metode de estimare a mişcării ce folosesc estimarea fluxului optic putem menţiona:

- [Lim 01] ce propune îmbunătăţirea calităţii fluxului optic obţinut cu metoda Lucas-Kanade [Barron 94] pentru deplasări importante ale pixelilor din imagine,
- [Timoner 01] propune folosirea de filtre multi-dimensionale discrete pentru a ameliora precizia estimării cât şi invarianţa la efectul de "motion blur"[2],

[1] metoda Gauss-Seidel este o tehnică iterativă folosită pentru a rezolva sisteme de ecuaţii liniare de tip $Ax = b$, unde A este o matrice de coeficienţi, x reprezintă un vector de necunoscute iar b este un vector de valori. Soluţiile obţinute la o anumită iteraţie sunt exprimate în funcţie de valorile calculate anterior, $x_i^{(k)} = \frac{b_i - \sum_{j<i} a_{i,j} x_j^{(k)} - \sum_{j>i} a_{i,j} x_j^{(k-1)}}{a_{i,j}}$, unde k reprezintă iteraţia curentă. Aceasta este o versiune îmbunătăţită a metodei Jacobi.

[2] efectul de "motion blur" corespunde urmelor de culoare ce apar în imagine în urma mişcării rapide a obiectelor sau a camerei video.

- [Elad 99] propune o soluţie optimală pentru realizarea unor filtre specifice calculului gradientului, în scopul estimării mişcării.

Metodele "pel-recursive"

Metodele "pel-recursive" [Netravali 79] realizează estimarea mişcării 2D a scenei la nivel de pixel şi în *mod recursiv*. Dacă dispunem de o estimare iniţială a mişcării pentru fiecare pixel al imaginii, $d_i = (u_i, v_i)$, unde d_i reprezintă vectorul de deplasare, atunci este realizată o corecţie pe baza funcţiei DFD rezultate:

$$d_{i+1} = d_i + \Delta d_i \tag{3.10}$$

unde

$$\Delta d_i = (\Delta u_i, \Delta v_i) \tag{3.11}$$

reprezintă termenul de actualizare al iteraţiei i. Iterarea poate fi executată, fie pentru o linie de pixeli, fie pentru anumite linii de pixeli sau între imagini, operaţii numite şi recurenţă orizontală, verticală şi respectiv temporală.

Acest tip de abordare porneşte de la ipoteza conform căreia funcţia DFD converge local spre zero când mişcarea estimată tinde spre mişcarea reală prezentă în scenă. Principiul estimării constă în minimizarea recursivă a valorii cuadratice a funcţiei DFD pe baza metodei de gradient, astfel:

$$d_{i+1} = d_i + \varepsilon \cdot DFD(\vec{r}, \vec{d}, \Delta t) \cdot \bigtriangledown_{d_i}(DFD(\vec{r}, \vec{d}, \Delta t)) \tag{3.12}$$

unde $\vec{r} = (x, y)$ reprezintă poziţia iniţială, $\vec{d} = (u, v)$ reprezintă deplasarea între momentele t şi $t+\Delta t$, $\bigtriangledown_{d_i}$ este operatorul de gradient calculat în funcţie de d_i iar ε este în general o constantă pozitivă.

Exprimând gradientul în funcţie de gradientul spaţial, obţinem ecuaţia următoare:

$$d_{i+1} = d_i + \varepsilon \cdot DFD(\vec{r}, \vec{d}, \Delta t) \cdot \bigtriangledown I(x + u, y + v, t + \tau) \tag{3.13}$$

unde $\bigtriangledown$ este operatorul de gradient spaţial iar I reprezintă imaginea. Alegerea adecvată a valorii lui ε permite obţinerea convergenţei estimării. Daca valoarea utilizată este ridicată, atunci convergenţa este rapidă dar mai puţin precisă şi vice-versa, dacă valoarea lui ε este mică, atunci convergenţa este sigură dar foarte lentă.

Această abordare, ca şi estimarea iterativă, are ca rezultat un câmp vectorial dens. Principala problemă în acest caz este convergenţa metodei care este dependentă de ipoteza de plecare. Mai mult, metodele "pel-recursive" sunt foarte sensibile la prezenţa zgomotului în imagine, acesta reducând drastic precizia estimării.

Ca exemple de metode de estimare a mişcării de tip "pel-recursiv" putem menţiona:

- [Estrela 04] ce propune o metodă de estimare robustă la prezenţa zgomotului în imagine. Vectorii de mişcare sunt estimaţi iterativ folosind metoda EM ("Expectation-Maximization"[3]) şi modelarea Gaussiană a datelor.

- [Hampson 00] ia în calcul influenţa variaţiilor de luminozitate a imaginii asupra calităţii estimării mişcării. Pentru aceasta, estimatorului de tip "pel-recursiv" i se adaugă un coeficient multiplicativ ce modelează variaţiile intensităţii luminoase ale scenei.

3.1.2 Metodele parametrice

Metodele parametrice modelează deplasarea pixelilor din imagine pe baza unui anumit set de parametri. Problema estimării mişcării este transformată astfel într-o problemă de estimare a acestor parametrii. Numărul de parametri folosiţi la estimare constituie caracteristica definitorie a metodei folosite. Spre deosebire de metodele din categoriile enunţate anterior, metodele parametrice folosesc o restricţie de natură geometrică, şi anume pornesc de la ipoteza conform căreia obiectele din imagine sunt suprafeţe plane şi rigide.

Modelul cel mai simplu de parametrizare a mişcării îl constituie *modelul translaţional* ce defineşte noua poziţie, (x', y'), a pixelului curent analizat, (x, y), ca fiind:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{3.14}$$

unde t_x şi t_y reprezintă deplasările pe cele două axe, oX şi respectiv oY.

Introducerea unui factor unic de scalare, atât orizontală cât şi verticală, conduce la un model cu trei parametrii, astfel:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = C \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{3.15}$$

unde C este parametrul de scală.

[3]un algoritm de tip EM - Expectation-Maximization este folosit în statistică pentru a găsi estimatorul matematic cel mai adecvat al parametrilor unui anumit model probabilistic, unde acesta din urmă depinde doar de "variabile latente" (variabile ce nu sunt direct observate, ci inferate pe baza altor variabile ce sunt observate şi măsurate în mod direct).

O imbunătăţire a modelului astfel obţinut constă în adăugarea unui parametru suplimentar pentru a specifica factorul de scală pe fiecare axă în parte, astfel:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} C_x & 0 \\ 0 & C_y \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{3.16}$$

unde C_x şi C_y sunt de această dată factorii de scală pe cele două axe, oX şi respectiv oY.

O variantă similară poate fi considerat modelul în care factorul de scală este înlocuit cu un factor de rotaţie cu un anumit unghi θ, astfel:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{3.17}$$

Combinând ultimele două modele, obţinem un model cu cinci parametri, astfel:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \cdot \begin{bmatrix} C_x & 0 \\ 0 & C_y \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{3.18}$$

Separând factorii de rotaţie pe axa oX şi respectiv oY, ajungem la modelul de mişcare cel mai cunoscut, şi anume modelul dat de *transformarea afină* a conţinutului imagini, astfel:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} C_x \cdot cos\theta_x & -C_y \cdot sin\theta_y \\ C_x \cdot sin\theta_x & C_y \cdot cos\theta_y \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{3.19}$$

unde θ_x şi θ_y reprezintă unghiurile de rotaţie pe cele două axe. Modelul afin este rezultatul *proiecţiei ortogonale* a mişcării pe o suprafaţă plană.

Folosind o *proiecţie de perspectivă*, obţinem un model al mişcării ce foloseşte opt parametri (*transformare de perspectivă*), a_i cu $i = 1, ..., 8$, în care noile coordonate (x', y') ale pixelului curent analizat sunt date de relaţiile:

$$x' = \frac{a_1 + a_2 \cdot x + a_3 \cdot y}{1 + a_7 \cdot x + a_8 \cdot y} \tag{3.20}$$

$$y' = \frac{a_4 + a_5 \cdot x + a_6 \cdot y}{1 + a_7 \cdot x + a_8 \cdot y} \tag{3.21}$$

O altă transformare folosită în mod curent este *transformarea biliniară* ce este dată de relaţiile următoare:

$$x' = a_1 \cdot x + a_2 \cdot y + a_3 \cdot x \cdot y + a_4 \tag{3.22}$$

$$y' = a_5 \cdot x + a_6 \cdot y + a_7 \cdot x \cdot y + a_8 \tag{3.23}$$

Alte reprezentări iau în calcul şi efectul acceleraţiei mişcării, ca de exemplu modelul propus în [Sanson 81]:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_x^x & a_x^y \\ a_y^x & a_y^y \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b_x^{x^2} & b_x^{xy} & b_x^{y^2} \\ b_y^{x^2} & b_y^{xy} & b_y^{y^2} \end{bmatrix} \cdot \begin{bmatrix} x^2 \\ xy \\ y^2 \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \quad (3.24)$$

unde a şi b reprezintă seturile de parametrii folosiţi pentru modelizare.

Ca exemple de metode ce folosesc estimarea parametrică a mişcării putem menţiona:

- [IRISA 05] ce propune o librărie C++ completă de funcţii de calcul a estimării parametrice a mişcării ce folosesc marea parte a modelelor existente, precum modelul translaţional, afin şi cuadratic,
- [Farnebäck 00] combină calculul tensorilor[4] 3D de mişcare folosind constrângerile unui model parametric. Aceştia sunt folosiţi pentru a estima viteza de mişcare în imagine,
- [Wallin 01] propune o extensie a algoritmului de căutare ierarhică folosit de standardul MPEG-7 pe baza modelelor parametrice de mişcare.

3.1.3 Metodele stohastice

Metodele stohastice de estimare a mişcării folosesc teoria statistică. Explorarea spaţiului parametrilor este ghidată în acest caz de procese aleatoare. În această categorie de metode putem enumera: *estimarea Bayesiană*, *modelele Markoviene* şi *algoritmii genetici* [Marichal 98].

O soluţie pentru a modela discontinuităţile câmpului vectorial de mişcare constă în folosirea câmpurilor aleatoare Markoviene sau MRF ("Markov Random Fields") [Graffigne 95] în ipotezele Bayesiene. În acest model, imaginea este reprezentată ca fiind un ansamblu de "locaţii", o "locaţie" reprezentând un pixel din imagine. Fiecare pixel este considerat astfel ca fiind o variabilă aleatoare, iar probabilitatea de apariţie a acestuia cât şi relaţiile de vecinătate cu ceilalţi pixeli sunt modelate cu ajutorul probabilităţilor condiţionate. Câmpul aleator definit în acest fel este un câmp MRF dacă probabilitatea condiţionată a fiecărui pixel depinde de un număr redus de pixeli vecini.

În [Papoulis 91] se propune modelarea discontinuităţilor câmpului de mişcare vectorial, pe baza introducerii noţiunii de "proces linie" sau proces de

[4] un tensor de rangul n într-un spaţiu $m-$dimensional este un obiect matematic ce are n indici şi m^n componente. Acesta se supune unui anumt set de reguli de transformare. Fiecare indice al tensorului ia valori în numărul dimensiunior spaţiului de definiţie.

discontinuitate. Acesta va permite interpretarea diferită a pixelilor vecini pixelului curent, analizat. Costul de introducere al acestui proces este modelat folosind câmpuri MRF. Astfel, elementele câmpului MRF vor avea două stări: fie sunt active, ipoteză în care a avut loc o discontinuitate în câmpul vectorial pe linia modelată, fie sunt inactive (pasive).

Estimarea mişcării va fi rezultatul minimizării unei funcţii de energie ce va depinde de deplasările pe cele două axe, oX şi respectiv oY, date de vectorul $\vec{d}$, precum şi de procesul de discontinuitate ("procesul linie") l. Acesta este modelat de patru parametri şi anume: b, b', c şi c'. Funcţia de energie astfel definită este dată de relaţia:

$$E(u,v|l) = (1-\lambda)\cdot\sum_{x,y} DFD^2(\vec{r},\vec{d}) + \lambda^2\cdot(\sum_{i,j} u_x^2(i,j)(1-b_{i,j}) + \sum_{i,j} v_x^2(i,j)(1-b'_{i,j}) + \sum_{i,j} u_y^2(i,j)(1-c_{i,j}) + \sum_{i,j} v_y^2(i,j)(1-c'_{i,j})) + \alpha\cdot\sum_{i,j}(b_{i,j}+b'_{i,j}+c_{i,j}+c'_{i,j}) \quad (3.25)$$

unde $\vec{r} = (x,y)$ reprezintă poziţia iniţială a pixelului analizat, $\vec{d} = (u,v)$ reprezintă vectorul de deplasare, λ este un parametru Lagrange, u_x, u_y, v_x şi respectiv v_y reprezintă diversele deplasări exprimate în funcţie de procesul de discontinuitate (vezi Figura 3.4) iar α reprezintă costul de introducere a unei discontinuităţi, situaţie în care parametrii b, b', c şi c' iau valoarea 1.

Modelele Markoviene introduc o complexitate de calcul ridicată dar rezultatele obţinute sunt apropiate de realitate [Marichal 98].

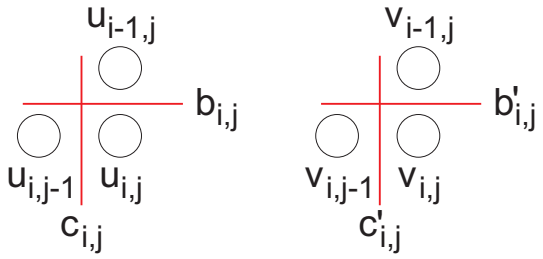


Figura 3.4: Configuraţiile particulare ale procesului de discontinuitate în modelarea Markoviană (cercurile reprezintă pixelii iar liniile procesele de discontinuitate).

Un alt tip de abordare pentru estimarea mişcării constă în folosirea algoritmilor genetici. Algoritmii genetici sunt bazaţi pe teoria evoluţiei formulată de Darwin. Principiul acestora constă în evoluţia artificială a unei populaţii

de dispozitive, proces ce este realizat pe baza unor operatori specifici, precum: selecţia, creşterea sau mutaţia (vezi Secţiunea 7.1.2). De regulă algoritmii genetici sunt folosiţi pentru probleme de optimizare a sistemelor ce implică un număr foarte mare de parametri sau obiective.

Ca exemple de metode de estimare a mişcării putem menţiona metoda propusă în [Zaim 01] unde obiectele din imagine sunt parametrizate folosind vectori de caracteristici relative la forma şi traiectoria acestora în imagine. În acest caz, algoritmii genetici sunt folosiţi pentru estimarea acestor vectori de caracteristici. Un alt exemplu este metoda propusă în [Pan 00] ce vizează estimarea mişcării 3D a oamenilor.

3.1.4 Metodele de estimare pe blocuri de pixeli

Tehnica de estimare a mişcării bazată pe analiza blocurilor de pixeli a fost propusă pentru prima oară în [Jain 91]. Spre deosebire de metodele prezentate în paragrafele anterioare, metode ce estimau mişcarea la nivel de pixel, estimarea pe blocuri calculează câmpul vectorial de mişcare la nivel de regiuni de pixeli, astfel furnizând un vector de deplasare pentru fiecare dintre acestea. Ameliorată de-a lungul timpului, această abordare s-a dovedit o metodă de estimare foarte eficientă ce furnizează cel mai bun compromis între complexitatea de calcul şi precizia rezultatelor obţinute. În funcţie de dimensiunea blocurilor de pixeli folosite la estimare, întâlnim două situaţii posibile:

- o *sensibilitate ridicată a estimării* se obţine pentru blocuri de pixeli de dimensiuni reduse, situaţie utilă în cazul în care estimarea mişcării trebuie să furnizeze rezultate precise, ca de exemplu pentru analiza deplasărilor fine ale obiectelor din scenă. În ciuda preciziei ridicate, în acest caz estimarea va fi mai sensibilă la prezenţa zgomotului în imagine.

- o *robusteţe ridicată a estimării* se obţine pentru blocuri de dimensiuni mai mari. Folosirea mai multor pixeli pentru evaluarea funcţiei de cost face ca estimarea să fie mai puţin sensibilă la modificările provocate de prezenţa zgomotului în imagine. Totuşi, având în vedere că vectorii de mişcare sunt calculaţi pentru fiecare bloc de pixeli, rezultatele reprezintă o aproximare mai grosieră a câmpului vectorial obţinut la nivel de pixel. Această situaţie este utilă în cazul în care se analizează mişcarea la nivel global, ca de exemplu pentru analiza mişcării camerei video.

Datorită acestor proprietăţi, după cum am mai menţionat şi în partea introductivă a acestui capitol, metodele bazate pe analiza blocurilor de pixeli

sunt utilizate cu predilecţie de marea parte a standardelor de codare existente (MPEG-1, 2, 4, etc.).

Principiul estimării pe blocuri de pixeli este următorul. În primă fază, imaginea curentă analizată la momentul t, $I(t)$, este împărţită în blocuri disjuncte de $B \times B$ pixeli, unde B este de regulă ales ca fiind o putere a lui 2, din motive de optimizare hardware a metodei, $B \in \{2, 4, 8, 16, ...\}$. Pentru fiecare bloc al imaginii $I(t)$, notat $I(\vec{r}, t)$, unde $\vec{r} = (x', y')$ reprezintă poziţia acestuia în imagine, se caută noua sa poziţie în imaginea următoare, $I(t+l)$, la momentul $t + l$, unde l reprezintă pasul de analiză (uzual $l = 1$).

Pentru a reduce complexitatea de calcul, căutarea acestuia nu se efectuează în toată imaginea, ci într-o *fereastră de căutare limitată*, S, de regulă de dimensiune $(2B + 1) \times (2B + 1)$ pixeli. Acest lucru este posibil datorită faptului că în condiţiile de continuitate a mişcării, deplasările de la un cadru la altul cadru sunt mici, fiind foarte puţin probabil ca blocul de pixeli analizat să se deplaseze în afara ferestrei S. Fereastra S este aleasă în imaginea $I(t+l)$ ca fiind centrată pe blocul $I(\vec{r}, t + l)$.

Noua poziţie a blocului curent analizat este dată de minimizarea unei funcţii de cost, $F_c()$, ce estimează eroarea de aproximare a blocului curent, $I(\vec{r}, t)$, cu blocurile analizate din fereastra de căutare S din imaginea la momentul $t + l$. Astfel, vectorul de deplasare al blocului curent este dat de relaţia următoare:

$$\vec{d_m} = argmin_{\vec{d} \in S} F_c(I(\vec{r} - \vec{d}, t + l), I(\vec{r}, t)) \tag{3.26}$$

unde $\vec{d_m}$ reprezintă deplasarea blocului curent $I(\vec{r}, t)$ pentru care funcţia F_c este minimală. Valorile lui $\vec{d}$ sunt toate deplasările posibile ale blocului de comparare în interiorul ferestrei de căutare. Principiul este ilustrat în Figura 3.5.

Dacă căutarea se face pentru toate valorile posibile ale vectorului de deplasare $\vec{d}$ din fereastra de căutare S, atunci căutarea este o *căutare completă*. Căutarea completă este optimală în detrimentul vitezei de calcul. Numărul de operaţii estimate pentru o căutare completă este definit în [Accame 98] ca fiind:

$$\Delta_f = M \cdot N \cdot B^2 \cdot (2W + 1)^2 \tag{3.27}$$

unde $M \cdot N$ reprezintă numărul de blocuri de pixeli din imagine, B^2 este dimensiunea unui bloc de pixeli iar $(2W+1) \times (2W+1)$ reprezintă dimensiunea ferestrei de căutare S. De exemplu, pentru o imagine de 352×288 pixeli, $B = 16$ şi $W = 16$, numărul de operaţii necesare se ridică la $\Delta_f = 1.1 \cdot 10^8$, ce reprezintă un număr semnificativ, luând în calcul că o operaţie de filtrare cu o mască convolutivă de dimensiune 3×3 necesită în acest caz în jur de $\Delta_f = 1.5 \cdot 10^6$ operaţii.

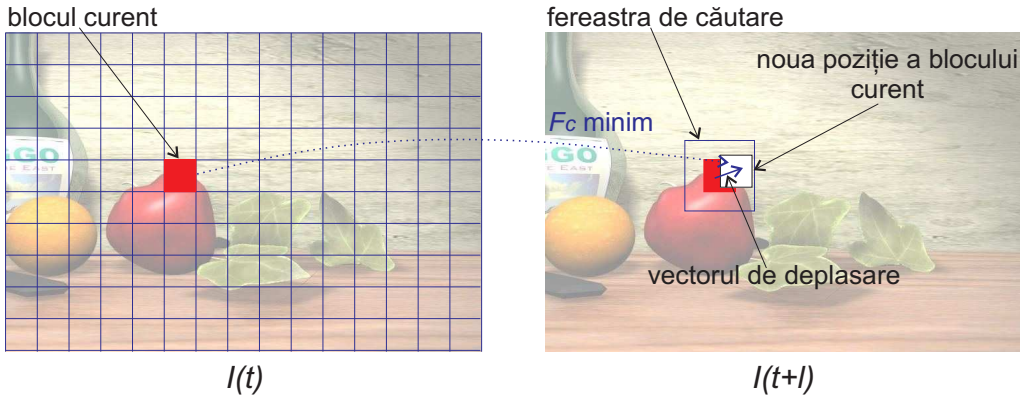


Figura 3.5: Principiul de estimare pe blocuri de pixeli a mişcării ($I(t)$ reprezintă imaginea la momentul t).

Dintre funcţiile de cost cel mai frecvent folosite, putem menţiona următoarele ($b_1()$ şi $b_2()$ reprezintă două blocuri de pixeli de dimensiuni $M \times N$):

- diferenţa medie absolută sau MAD ("Mean Absolute Difference"):

$$MAD(b_1, b_2) = \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} |b_1(i,j) - b_2(i,j)| \tag{3.28}$$

- diferenţa medie pătratică sau MSD ("Mean Square Difference"):

$$MSD(b_1, b_2) = \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} [b_1(i,j) - b_2(i,j)]^2 \tag{3.29}$$

- clasificarea distanţelor dintre pixeli sau PDC ("Pel Difference Classification"):

$$PDC(b_1, b_2) = \sum_{i=1}^{M} \sum_{j=1}^{N} ord\left(|b_1(i,j) - b_2(i,j)| \leq \tau\right) \tag{3.30}$$

  unde funcţia $ord(P)$ returnează valoarea 1 dacă propoziţia P este adevărată şi 0 altfel, iar τ este un prag ales arbitrar.

- proiecţia integrală sau IP ("Integral Projection"):

$$IP(b_1,b_2)=\sum_{i=1}^{M}\left|\sum_{j=1}^{N}b_1(i,j)-\sum_{j=1}^{N}b_2(i,j)\right|+\sum_{j=1}^{N}\left|\sum_{i=1}^{M}b_1(i,j)-\sum_{i=1}^{M}b_2(i,j)\right| \tag{3.31}$$

În general în urma minimizării funcţiei de cost, $F_c()$, pentru un anumit bloc de pixeli analizat, intervin trei situaţii particulare ce trebuiesc luate în considerare de algoritmul de estimare, astfel:

- valoarea minimală este obţinută pentru blocul de pixeli din imaginea la mometul $t+l$ ce se află în acceaşi poziţie cu blocul curent analizat. În această situaţie vectorul de deplasare este nul, fiind un caz de *"absenţă a mişcării"*,
- valoarea minimală a funcţiei de cost este foarte ridicată, fiind superioară unui anumit prag, $\tau_{discont}$, determinat empiric. În acest caz, mişcarea blocului analizat este catalogată ca fiind un caz de *"mişcare discontinuă"*, aceasta neavând continuitate temporală în imaginea următoare,
- valoarea minimală este nenulă dar şi inferioară pragului $\tau_{discont}$. În acest caz, deplasarea obţinută corespunde *mişcării blocului analizat.*

În ceea ce priveşte pragul $\tau_{discont}$, numit şi prag de discontinuitate, valoarea acestuia poate fi estimată "a priori" pe baza expertizei manuale a diverselor pasaje de discontinuitate a mişcării ce pot apărea în secvenţele de imagini, ca de exemplu pasajele tranziţiilor video [Ionescu 07b].

Complexitatea de calcul a unei metode de estimare pe blocuri de pixeli este dată în primul rând de *modalitatea de căutare* a noii poziţii a blocului curent analizat în fereastra S, şi apoi de *dimensiunea ferestrei* de căutare precum şi de funcţia de cost folosită. În acest sens, căutarea completă se dovedeşte a fi căutarea cu gradul de complexitate de calcul cel mai ridicat. Metodele existente încearcă să găsească diverse soluţii pentru a optimiza modul în care este realizată căutarea, fără a pierde însă semnificativ din precizia estimării. În cele ce urmează, vom face o trecere în revistă a diverşilor algoritmi de căutare existenţi [Turaga 98].

Căutarea completă

După cum am menţionat în paragrafele anterioare, căutarea completă constă în minimizarea funcţiei de cost pentru toate blocurile de pixeli din interiorul

ferestrei de căutare S. Astfel, căutarea completă este optimală dar are complexitatea de calcul cea mai importantă dintre metodele de căutare existente. Din această cauză, căutarea completă este folosită de regulă ca referinţă pentru evaluarea celorlalţi algoritmi de căutare.

Căutarea în trei etape

Principiul general constă în calcularea în avans a trei valori ale funcţiei de cost, $F_c()$, pentru trei valori diferite ale deplasării blocului curent analizat, valori ce sunt alese după un anumit criteriu. Pe parcursul căutării, aceste trei valori sunt recalculate în mod iterativ, căutarea finalizându-se în momentul în care este îndeplinită o anumită condiţie de convergenţă [Reoxiang 94].

În această categorie se găseşte algoritmul de căutare folosit de standardul de compresie video H.263+ [4i2i 06]. Pentru acesta, noua poziţie a blocului curent este căutată în direcţia valorii minimale a funcţiei de cost folosind doar patru direcţii de analiză, şi anume: Sud, Est, Nord şi Vest. Algoritmul este următorul (vezi Figura 3.6):

1. mai întâi este calculată funcţia de cost între blocul curent analizat din imaginea la momentul t şi cele patru blocuri vecine la distanţă de un pixel pe direcţiile orizontală şi respectiv verticală din imaginea următoare la momentul $t + 1$. Valoarea minimală a funcţiei de cost astfel obţinută, este salvată în variabila D_1 iar blocul de pixeli căruia îi corespunde această valoare devine bloc curent în imaginea la momentul $t + 1$.

2. folosind acelaşi principiu, se repetă calculul funcţiei de cost pentru vecinii noului bloc curent analizat în imaginea la momentul $t + 1$. Valoarea anterioară a variabilei D_1, va fi salvată într-o variabilă D_2 iar noua valoare minimală a funcţiei de cost este salvată în variabila D_1. Blocul pentru care se obţine această nouă valoare devine astfel bloc curent.

3. principiul se repetă mai departe pentru noul bloc curent. Astfel, valoarea variabilei anterioare, D_2, este salvată în noua variabilă D_3, valoarea variabilei D_1 este salvată în D_2 iar noua valoare minimală a funcţiei de cost este reţinută în variabila D_1. Dacă condiţia:

$$D_3 \leq D_1 \quad \cap \quad D_2 \leq D_1 \tag{3.32}$$

este indeplinită, atunci căutarea este încheiată. În caz contrar, blocul ce a furnizat eroarea minimală devine bloc curent pentru analiză.

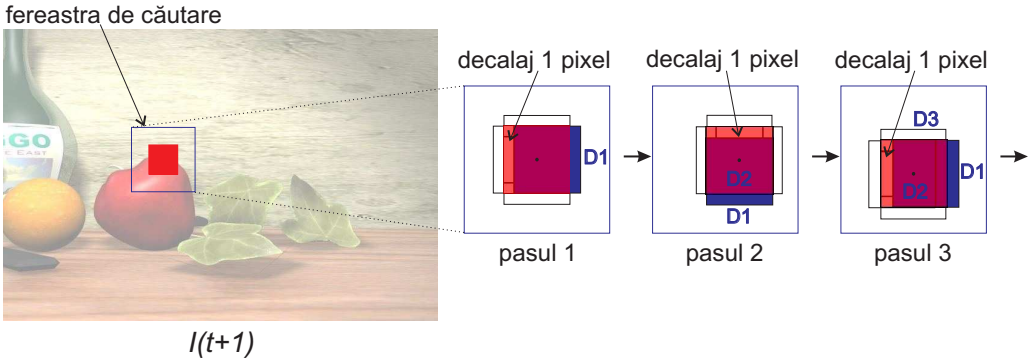


Figura 3.6: Principiul de estimare a mişcării în trei etape folosit de standardul H.273+ (blocul de pixeli curent este marcat cu roşu iar blocul ce corespunde erorii minimale cu albastru).

4. se repetă în acelaşi mod calculul funcţiei de cost pentru noul bloc curent. Valoarea lui D_2 va fi salvată în D_3, valoarea lui D_1 în D_2 iar noua valoare minimală a funcţiei de cost este reţinută în D_1. Dacă valorile D_1, D_2 şi D_3 nu satisfac condiţia 3.32, se repetă pasul 4. Dacă însă condiţia 3.32 este satisfăcută, atunci căutarea se încheie iar ultimul bloc de pixeli ce a furnizat valoarea minimală a funcţiei de cost va determina vectorul de deplasare.

Pentru a evita poziţionarea pe un minim local al funcţiei de cost, se poate folosi o condiţie suplimentară, şi anume ca valoarea minimală a funcţiei $F_c()$, obţinută pentru ultimul bloc analizat din fereastra de căutare S, să fie de asemenea minimală în raport cu valoarea $F_c()$, obţinută pentru blocul omolog blocului curent analizat din imaginea la momentul t, din imaginea la momentul $t+1$, bloc pentru care se efectuează estimarea. În caz contrar, vectorul de mişcare poate fi considerat nul.

Căutarea logaritmică

Similar cu principiul căutării în trei etape, căutarea logaritmică reduce numărul de comparaţii între blocuri, şi astfel complexitatea de calcul, prin parcurgerea ferestrei de căutare S în direcţia minimului funcţiei de cost $F_c()$ [Lundmark 01].

Pentru fiecare bloc curent analizat din imaginea la momentul t, căutarea noii sale poziţi în imaginea la momentul $t+1$ începe cu blocul omolog, pe care va fi centrată fereastra de căutare S. Se vor lua în considerare doar deplasările pe cele patru direcţii fundamentale, şi anume: Sud, Est, Vest şi Nord. Astfel,

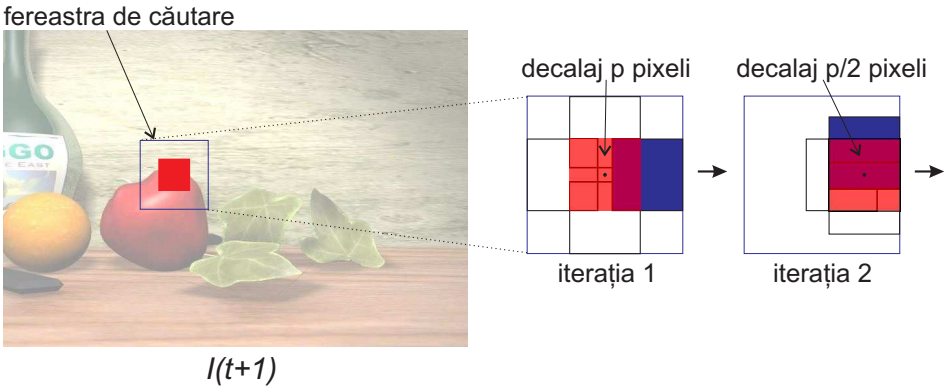


Figura 3.7: Principiul căutării logaritmice (blocul de pixeli curent este marcat cu roşu iar blocul ce corespunde erorii minimale cu albastru).

în primă etapă, funcţia de cost este calculată pentru cele patru blocuri la o distanţă de p pixeli de blocul curent pe cele patru direcţii considerate (vezi Figura 3.7). Blocul ce corespunde valorii minimale a funcţiei de cost, devine bloc curent pentru etapele următoare. La fiecare etapă, deplasarea p este înjumătăţită, astfel $p \leftarrow p/2$.

Algoritmul se repetă până în momentul în care $p = 1$. Ultimul bloc din imaginea la momentul $t + 1$ pentru care s-a obţinut valoarea minimală a funcţiei de cost, va determina vectorul de deplasare. Pentru a evita poziţionarea pe un minim local al funcţiei de cost, se poate folosi aceeaşi condiţie ca cea enunţată anterior pentru căutarea în trei etape, şi anume ca valoarea minimală obţinută pentru ultimul bloc analizat să fie minimală şi în raport cu eroarea obţinută pentru blocul iniţial, pentru care se efectuează estimarea.

Cu această metodă de căutare, numărul de blocuri comparate este de $2 + 7 \cdot log_2 W$, unde W reprezintă dimensiunea ferestrei de căutare. Dependenţa logaritmică de valoarea lui W a dat numele metodei de "căutare logaritmică".

Căutarea binară

Căutarea binară este unul dintre algoritmii de căutare cei mai populari, fiind folosit şi pentru estimarea mişcării în standardul MPEG [Zahariadis 96].

Principiul căutării binare constă în divizarea ferestrei de căutare într-o serie de regiuni şi efectuarea unei căutări complete doar într-una dintre aceste regiuni. Algoritmul este următorul:

1. mai întâi funcţia de cost este estimată în imaginea la momentul $t + 1$, poziţionându-ne pe o grilă de 9 pixeli ce sunt repartizaţi în fereastra de

căutare după cele 8 direcţii cardinale. Folosind aceste puncte, fereastra de căutare este divizată în mai multe regiuni disjuncte. Accentul se pune pe regiunea centrală, fiind şi cea mai probabilă să furnizeze noua poziţie a blocului de pixeli curent analizat din imaginea la momentul t (vezi Figura 3.8).

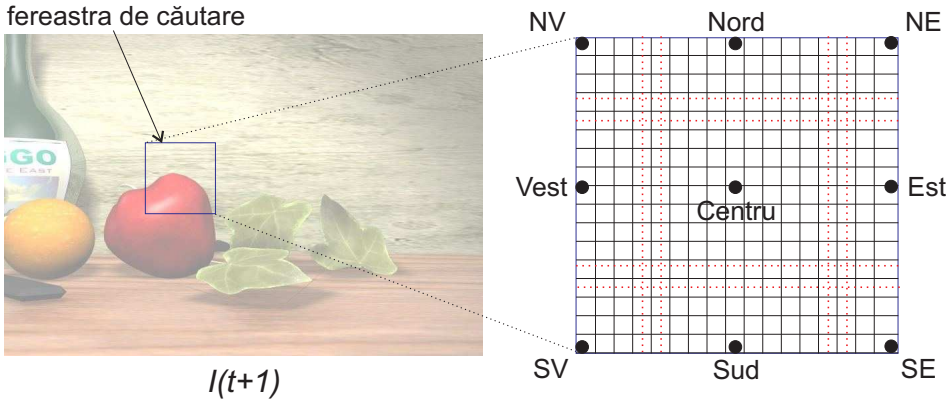


Figura 3.8: Principiul căutării binare: divizarea în regiuni a ferestrei de căutare (punctele negre reprezintă cei 9 pixeli ce formează grila de selecţie iar frontiera dintre regiuni este marcată cu linia discontinuă).

2. regiunea din care face parte blocul de pixeli ce furnizează valoarea minimală a funcţiei de cost devine regiune curentă de analiză. În aceasta, pentru determinarea vectorului de deplasare, se va efectua o căutare exhaustivă (completă), cu menţiunea că blocurile de pixeli aflate pe frontierele dintre regiuni nu sunt luate în calcul.

În ciuda numărului redus de comparaţii între blocuri, căutarea binară furnizează performanţe medii datorită faptului că anumite regiuni de pixeli nu sunt luate deloc în calcul.

Căutarea ortogonală

Căutarea ortogonală este o combinaţie între căutarea în trei etape şi căutarea logaritmică. Aceasta implică o etapă de căutare pe verticală urmată de o etapă de căutare pe orizontală a blocului optimal. Algoritmul este următorul:

1. se alege un pas de analiză, p, de regulă ca fiind jumătate din valoarea deplasării maxime în fereastra de căutare. Se evaluează funcţia de cost, $F_c()$, pentru blocurile de pixeli din imaginea la momentul $t+1$ ce se află la distanţa p, pe orizontală de centrul ferestrei de căutare. Blocul

pentru care se obţine valoarea minimală a funcţiei de cost devine blocul curent de analiză,

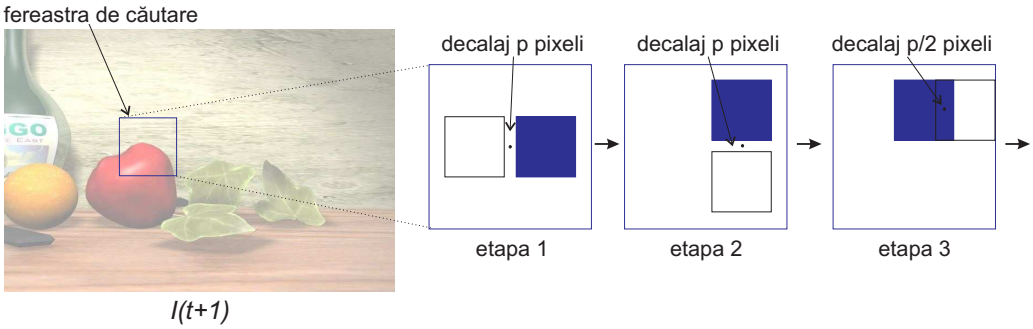


Figura 3.9: Principiul căutării ortogonale (blocul de pixeli ce corespunde erorii minimale este marcat cu albastru).

2. se evaluează funcţia de cost pentru blocurile de pixeli ce se află de această dată la distanţa p, pe verticală faţă de blocul curent din imaginea la momentul $t + 1$. Blocul ce furnizează valoarea minimală a funcţiei de cost devine astfel noul bloc curent de analiză (vezi Figura 3.9).

3. pasul de analiză este înjumătăţit, $p \leftarrow p/2$. Dacă $p > 1$, atunci procesul se repetă. În caz contrar, căutarea se încheie, ultimul bloc curent reprezentând noua poziţie căutată a blocului curent analizat din imaginea la momentul t.

Căutarea intercalată

Căutarea intercalată se bazează de asemenea pe principiul căutării logaritmice, doar că în acest caz, blocurile de pixeli sunt alese ca formând un "$\times$" şi nu un "+". Algoritmul este următorul:

1. funcţia de cost este calculată mai întâi pentru blocul de pixeli omolog, din imaginea la momentul $t+1$, blocului curent analizat din imaginea la momentul t. Dacă valoarea obţinută este inferioară unui anumit prag, atunci căutarea se încheie aici,

2. funcţia de cost este calculată pentru 4 blocuri de pixeli din imaginea la momentul $t + 1$, ce formează un "$\times$" în jurul centrului ferestrei de căutare S, la o distanţă p de acesta. Blocul ce furnizează valoarea minimală a erorii devine astfel blocul curent de analiză.

3. dacă pasul de analiză, p, este mai mare ca 1, atunci acesta este înjumătăţit şi se repetă etapa 2. În caz contrar, se trece la etapa 4,

4. dacă blocul curent se află, fie în colţul din stânga jos al ferestrei de căutare, sau în colţul din dreapta sus, atunci se mai evaluează o dată funcţia de cost pentru încă 4 blocuri, distribuite de această dată în "+" faţă de blocul curent şi la distanţa p. Dacă totuşi blocul curent se află în colţul din stânga sus sau dreapta jos al ferestrei de căutare, similar cazului precedent, funcţia de cost este evaluată pentru încă 4 blocuri dispuse de aceasta dată în "$\times$" la distanţa p.

Căutarea intercalată necesită un număr de aproximativ $5 + 4 \cdot log_2 p$ comparaţii, unde p reprezintă valoarea maximă a pasului de căutare. Aceasta are o complexitate de calcul redusă, dar nu este cel mai bun algoritm din punct de vedere al preciziei estimării.

Căutarea ierarhică

Căutarea ierarhică are ca scop reducerea complexităţii de calcul prin furnizarea mai multor niveluri de detaliu al căutării.

Un exemplu sunt metodele piramidale în care estimarea se face pe imagini piramidale construite pe baza sub-eşantionării progresive a imaginii iniţiale (vezi Figura 3.3). Estimarea mişcării este realizată astfel începând cu vârful piramidei, ce corespunde nivelului de detaliu cel mai scăzut, avansând spre baza piramidei, spre nivelul de detaliu cel mai ridicat ce finalizează cu imaginea iniţială.

Pentru a reduce influenţa zgomotului în imaginile la nivelurile inferioare de detaliu, piramidele sunt construite folosind filtrări de tip "trece-jos" (FTJ). Estimarea mişcării de pe un nivel superior poate fi realizată cu una dintre metodele existente, precum căutarea în trei etape. Aceasta va fi folosită ca punct de plecare pentru nivelul imediat inferior, noii vectorii de mişcare fiind o "rafinare" a celor anteriori [Lin 98].

Avantajul căutării ierarhice constă în faptul că se poate adapta la constrângeri de timp variabile, furnizând în funcţie de aplicaţie, nivelul de detaliu dorit. Dacă timpul de calcul este critic, estimarea se poate realiza rapid sacrificând din precizia rezultatelor.

Căutarea hibridă

Căutarea hibridă nu este o metodă propriu-zisă, aceasta folosindu-se pentru estimare, de metodele deja existente, profitând astfel de avantajele furnizate de fiecare dintre acestea. Principiul constă în determinarea mai întâi a tipului

de mişcare prezent în imagine, de exemplu: mişcare lentă, mişcare rapidă, staţionară, etc. Mai departe, în funcţie de aceasta, se alege pentru calculul vectorilor de mişcare metoda cea mai eficientă în acest caz [Ge 02]. Astfel, precizia metodele hibride depinde de eficienţa cu care se evaluează tipul de mişcare prezent în scenă.

3.1.5 Fluxul video MPEG

O alternativă pentru a recupera informaţia de mişcare constă în analiza informaţiei din domeniul comprimat, sau în particular analiza fluxului MPEG - "Moving Picture Experts Group" [Pilu 97]. Metodele de codare video existente folosesc pentru compresie estimarea şi compensarea mişcării, astfel vectorii de mişcare vor fi conţinuţi în fluxul video comprimat.

În mare, principiul compresiei temporale este următorul: în loc să fie stocate toate imaginile seventei (de exemplu, 25 de imagini pe secundă), vor fi stocate integral doar anumite imagini, numite şi imagini cheie, precum şi vectorii de deplasare a blocurilor de pixeli din acestea în imaginile următoare. În momentul decompresiei, imaginile secventei sunt reconstituite pornind de la o imagine de referinţă, ce poate fi o imagine cheie sau o imagine reconstruită anterior, pe baza vectorilor de deplasare, folosind compensarea mişcării.

Avantajul folosirii vectorilor de mişcare direct din fluxul video constă în faptul că aceştia au fost calculaţi în momentul codării. Astfel, imaginile folosite pentru estimare sunt calitativ net superioare imaginilor obţinute în urma decompresiei, ce stau de regulă la baza estimării mişcării cu metodele prezentate anterior. Acest lucru se datorează în principal compresiei cu pierdere de informaţie folosită de codarea video. În funcţie de constrângerile de debit impuse, aceasta poate deteriora semnificativ conţinutul imaginii (necomprimate). Mai mult, codarea video este o codare intensivă ce presupune o compresie atât spaţială a imaginii (de regulă JPEG), cât şi temporală pe baza vectorilor de mişcare.

Principiul de funcţionare general al unui decodor MPEG este ilustrat în Figura 3.10. Mai întâi, datele video sunt trecute printr-un decodor Huffman[5] ce permite recuperarea coeficienţilor DCT[6] cuantificaţi cât şi a vectorilor de mişcare. Mai departe, vectorii de mişcare sunt separaţi de fluxul de date şi sunt furnizaţi blocului de compensare a mişcării. În acelaşi timp, coeficienţii

[5]în teoria informaţiei, codarea Huffman este un algoritm de codare bazat pe calculul entropiei folosit la compresia fără pierderi a datelor. Pentru compresie se foloseşte un dicţionar de coduri, de dimensiune variabilă, pe baza căruia este codat fiecare simbol al sursei. Dicţionarul este construit pe baza estimării probabilităţii de apariţie a fiecărei valori posibile a simbolurilor emise de sursă.

[6]vezi explicaţia de la pagina 41.

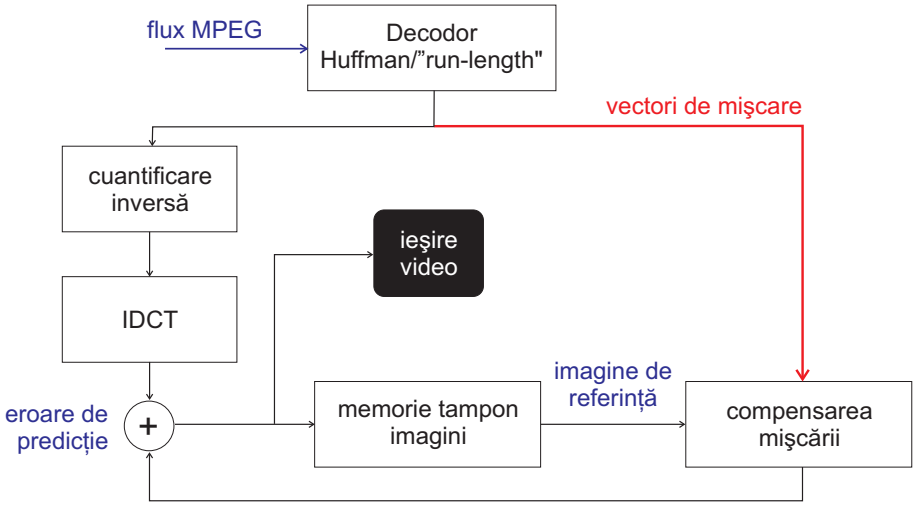


Figura 3.10: Diagrama de principiu a unui decodor MPEG-1 (sursă [Bretl 99]).

DCT sunt extrapolaţi şi trecuţi prin blocul de calcul al transformatei cosinus discretă inversă (IDCT) pentru a obţine înapoi informaţia spaţială din imagine.

În cazul imaginilor de tip P, ce sunt imagini reconstruite în funcţie de imagini anterioare ("forward prediction"), sau de tip B, ce sunt imagini cu predicţie bidirecţională, reconstruite atât din imagini anterioare cât şi următoare ("forward and backward prediction"), vectorii de mişcare sunt mutaţi la o anumită adresă de memorie de către blocul de predicţie, pentru a accesa macro-blocul de predicţie dintr-o imagine de referinţă stocată anterior. Sumatorul adaugă această predicţie la valoarea reziduală pentru a reconstrui datele din imagine. În cazul imaginilor de tip I, ce sunt imagini fără predicţie, nu există vectori de mişcare disponibili şi nici imagine de referinţă, astfel că predicţia este zero. De asemenea, pentru imaginile de tip I şi P, ieşirea sumatorului este stocată ca imagine de referinţă pentru predicţiile ulterioare.

Astfel, vectorii de mişcare pot fi extraşi direct din fluxul de date după decodarea Huffman. Pentru mai multe detalii referitor la implementarea principială cât şi practică a metodelor de recuperare a vectorilor de mişcare din fluxul video MPEG, cititorul se poate raporta la lucrările [Gilvarry 99] şi [Bretl 99].

În realitate, vectorii de mişcare obţinuţi direct din fluxul MPEG nu sunt întotdeauna coerenţi. Aceştia necesită de regulă o serie de etape de corecţie

pentru ameliorare. O situaţie frecventă ce cauzează ambiguitatea vectorilor de mişcare este estimarea mişcării pentru regiuni netexturate din imagine, unde blocurile de pixeli vecine nu prezintă suficientă variabilitate pentru a evidenţia deplasarea acestora [Pilu 97]. Un exemplu de incoerenţă a vectorilor de mişcare extraşi direct din fluxul video este prezentat în Figura 3.11, unde am ilustrat câmpul vectorial de mişcare obţinut pentru o imagine dintr-o secvenţă ce conţine deplasarea a două personaje simultan cu mişcarea globală a camerei video. Astfel, vectorii incoerenţi sunt obţinuţi în special pentru pixelii de pe bordura imaginii, ce nu pot fi localizaţi în imaginea următoare datorită deplasării globale a scenei. Pentru a corecta această problemă, şi astfel pentru a obţine o calitate suficientă a vectorilor de mişcare, soluţia cea mai frecvent adoptată este decompresia datelor până la un anumit nivel de detaliu [Pineau 05].

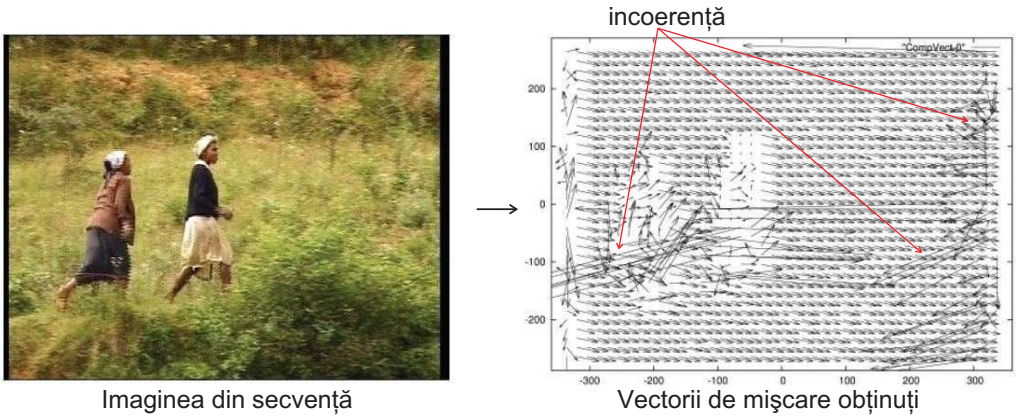


Figura 3.11: Exemplu de vectori de mişcare incoerenţi extraşi direct din fluxul MPEG-2 (sursă "Projet Analyse et Indexation Vidéo" [Pineau 05]).

3.2 Analiza mişcării camerei video

Una dintre aplicaţiile de mare interes ale estimării mişcării o constituie caracterizarea mişcării globale a scenei dată de deplasarea camerei video.

Acest tip de analiză permite obţinerea unei caracterizări de ansamblu a acţiuni din secvenţă, anumite tipuri de mişcări ale camerei video fiind folosite voluntar pentru a marca evenimente importante din derularea secvenţei. De exemplu, în anumite situaţii, schimbarea unui plan video este realizată folosind o mişcare globală de translaţie ce permite focalizarea pe un anumit punct de interes din afara scenei curente; un conţinut bogat în acţiune este

deseori marcat de o mişcare rapidă a camerei video; personajele din secvenţă sunt aduse în prim plan de regulă folosind o mişcare a camerei video de tip "zoom-in" (mărire optică a imaginii curente), etc.

Din punct de vedere global, mişcarea camerei video este o mişcare liberă în spaţiul real 3D. Totuşi, în realitate, din cauza constrângerilor tehnice şi fizice ale construcţiei camerei video, aceasta se rezumă la un număr limitat de mişcări de bază. Mişcările mai complexe, ce tind să se apropie de o mişcare liberă în spaţiul infinit 3D, sunt constituite ca o compunere de mişcări elemetare. Cele mai importante dintre acestea sunt ilustrate în Figura 3.12, astfel întâlnim:

- *mişcările translaţionale* după cele trei axe de coordonate XYZ, precum mişcarea de tip "dollying" (depărtare de obiectiv), "zoom-in/ zoom-out" (mărire/micşorare), "tracking" (translaţie spre dreapta sau stânga) sau "booming" (translaţie în plan vertical, sus-jos). Mişcarea de tip "zoom-in/ zoom-out" poate fi considerată ca o mişcare de translaţie doar din punct de vedere al efectului produs, acesta fiind similar cu cel al mişcării translaţionale de tip "dollying" (în cazul "zoom-out"). Din punct de vedere tehnic, această mişcare se realizează fără deplasarea camerei video, prin mărirea/micşorarea optică a imaginii.

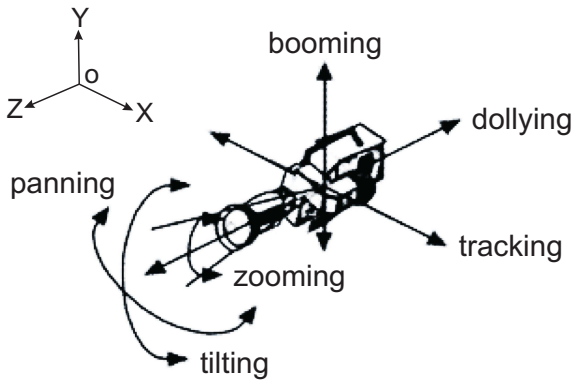


Figura 3.12: Mişcările camerei video.

- *mişcările de rotaţie* ale camerei video, precum rotaţia în sens orar şi antiorar în planul XoY, mişcarea de tip "panning" (rotaţie în planul XoZ) sau mişcarea de tip "tilting" (rotaţie în planul YoZ).

Metodele existente de analiză a mişcării camerei video pot fi regrupate în două categorii principale, astfel:

- metode ce analizează informaţia de mişcare direct în *domeniul comprimat* (fluxul MPEG),
- metode ce realizează analiza în *domeniul spaţio-temporal* al cadrelor video ale secvenţei.

Pentru un studiu complet al literaturii de specialitate al acestui domeniu, cititorul se poate raporta lucrările [Ngo 00], [Kramer 05], [Tardini 05] sau [Duan 06]. În cele ce urmează vom face o trecere în revistă a particularităţilor fiecărei dintre cele două categorii de metode existente.

3.2.1 Analiza mişcării camerei în domeniul comprimat

Din această categorie putem da ca exemplu abordarea probabilistică de detecţie a mişcării de tip "zoom-in/zoom-out", propusă în [Jin 02]. Aceasta foloseşte algoritmul EM ("Expectation-Maximization"[7]) pentru estimarea probabilităţii de apariţie a unei mişcări de tip "zoom" raportat la celelalte mişcări existente. În acest caz, informaţia de mişcare este extrasă direct din fluxul video MPEG-1 sau MPEG-2. Avantajul acestei abordări probabilistice constă în principal în îmbunătăţirea invarianţei la zgomotul ce afectează vectorii de mişcare, zgomot cauzat de regulă de erorile de cuantificare sau de prezenţa "artefactelor" datorate codării[8].

O abordare mai generală, ce vizează detecţia a şase tipuri de mişcări primare ale camerei video este propusă în [Kim 04]. Metoda propusă se bazează pe interpretarea calitativă a unui anumit număr de parametri extraşi din modelele parametrice de mişcare (vezi Secţiunea 3.1.2), parametrii ce sunt estimaţi direct, pe baza fluxului MPEG-2. Astfel, vectorii de mişcare sunt extraşi din fluxul MPEG-2 şi sunt comparaţi cu modelul afin de mişcare:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} a_2 & a_3 \\ a_5 & a_6 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a_1 \\ a_4 \end{bmatrix} \tag{3.33}$$

unde (u, v) reprezintă vectorul de mişcare al blocului de pixeli centrat în punctul de coordonate (x, y) iar a_i, cu $i = 1, ..., 6$, reprezintă parametrii modelului afin (vezi ecuaţia 3.19).

Mişcarea globală pentru o anumită imagine a secvenţei, poate fi exprimată ca un vector de parametri, $\phi = (a_1, a_2, ..., a_6)$, ce poate fi la rândul său

[7] vezi explicaţia de la pagina 83.

[8] "artefactele" de compresie sunt rezultatul eliminării informaţiei utile în cazul compresiei cu pierderi. Acestea pot fi vizibile, fie ca o alterare a tranziţiilor graduale din imagine, fie ca zgomot pe contururile obiectelor, sau în cazul compresiei pe blocuri de pixeli, ca o conturare a acestora în imagine (efect de tablă de şah).

estimat pe baza câmpului vectorial de mişcare folosind metoda celor mai mici pătrate. [Kim 04] propune reprezentarea informaţiei de mişcare cu vectori de parametri ce sunt exprimaţi în funcţie de diversele tipuri de mişcări ale camerei video. Fiecare imagine a secvenţei va fi astfel caracterizată de un vector, $\phi_c = (pan,\ tilt,\ div,\ rot,\ hyp)$, unde cei cinci parametri folosiţi au următoarea semnificaţie:

- *pan* şi *tilt* reprezintă mişcarea translaţională în plan orizontal şi respectiv vertical,
- *div* reprezintă mişcarea de tip "zoom",
- *rot* reprezintă mişcarea de rotaţie,
- *hyp* reprezintă ceea ce autorii numesc "flux hiperbolic", ce corespunde situaţiilor de predominanţă a mişcării de obiecte.

Aceştia sunt definiţi în concordanţă cu modelul afin de mişcare, în felul următor:

$$pan = a_1 \tag{3.34}$$

$$tilt = a_4 \tag{3.35}$$

$$div = \frac{1}{2}(a_2 + a_6) \tag{3.36}$$

$$rot = \frac{1}{2}(a_5 - a_3) \tag{3.37}$$

$$hyp = \frac{1}{4}(|a_2 - a_6| + |a_3 + a_5|) \tag{3.38}$$

Diversele tipuri de mişcări ale camerei video, respectiv "tracking", "tilting", rotaţie şi "zoom", sunt mai departe determinate pe baza filtrării cu un anumit prag a acestor vectori de parametri.

O abordare similară ce foloseşte parametrizarea modelului afin de mişcare a fluxului MPEG este propusă în [Kramer 05]. Algoritmul propus are o performanţă de trei, până la patru ori, mai rapidă decât timpul real[9].

Metoda propusă în [Lee 02b] detectează tipul de mişcare al camerei video comparând diversele modele de mişcare prezente în secvenţă cu un anumit set de modele predefinite. Într-o primă etapă, câmpul de mişcare este extras la nivel de imagine din fluxul MPEG. Vectorii de mişcare astfel obţinuţi

[9]prin performanţă în timp real a unei metode de analiză a conţinutului secvenţelor de imagini, înţelegem o rată de prelucrare de aproximativ 25 de imagini pe secundă (în standardul European). Metoda respectivă poate fi astfel aplicată în timp ce secvenţa este vizualizată fără a perturba continuitatea acesteia.

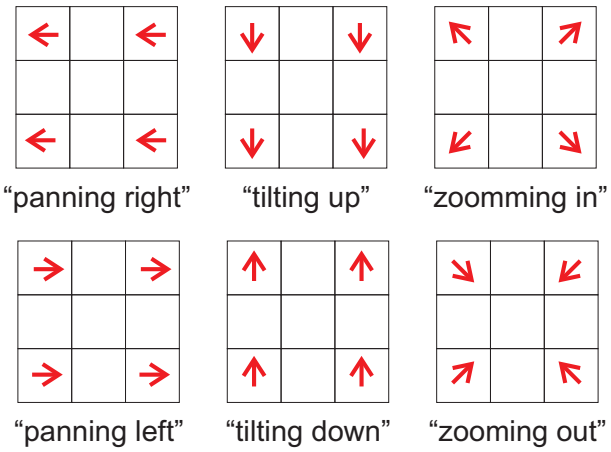


Figura 3.13: Exemple de modele predefinite de mişcare ale camerei video propuse în [Lee 02b] (săgeţile indică orientarea vectorilor de mişcare).

sunt împărţiţi la nivel de bloc de pixeli în două categorii: vectori de mişcare ce aparţin obiectelor din scenă, şi respectiv, vectori de mişcare ce aparţin fundalului imaginii. Folosind aceste informaţii, mişcările camerei video sunt determinate în funcţie de dispunerea acestora în nouă regiuni disjuncte ale imaginii (vezi Figura 3.13). Similaritatea mişcării dintre diversele regiuni ale imaginii este exprimată ca distanţă între histogramele de fază a vectorilor de mişcare conţinuţi în regiunea respectivă.

Principalul avantaj al metodelor de analiză a mişcării globale a camerei video ce folosesc domeniul comprimat este în primul rând timpul de calcul. Metodele din această categorie permit obţinerea de performanţe superioare prelucrării în timp real. Totuşi, precizia câmpului de mişcare este direct proporţională cu gradul de compresie al fluxului video folosit. Deseori, vectorii de mişcare obţinuţi direct din fluxul video modelează eronat mişcarea reală prezentă în secvenţă. Pentru a ameliora precizia, o soluţie constă în decomprimarea datelor până la un anumit nivel de detaliu şi re-estimarea mai precisă a mişcării folosind informaţia spaţio-temporală.

3.2.2 Analiza mişcării în domeniul spaţio-temporal

Metodele existente de analiză a mişcării camerei video ce folosesc informaţia spaţio-temporală sunt foarte variate. Astfel, întâlnim o vastă diversitate de tehnici de detecţie precum:

- metode ce folosesc modele predefinite de mişcare,

- metode de analiză a volumelor spaţio-temporale,
- metode de clasificare cu reţele neuronale,
- metode ce folosesc descompunerea "wavelet"[10], etc.

De exemplu, [Akutsu 92] propune pentru detecţia mişcării camerei video compararea repartiţiei vectorilor de mişcare din imagine cu modele predefinite, comparaţie ce este realizată de această dată în spaţiul transformatei Hough[11]. O abordare similară ce foloseşte modele predefinite de mişcare este propusă în [Ionescu 07b]. Aceasta vizează detecţia mişcărilor translaţionale, de rotaţie şi de tip "zoom-in/zoom-out", precum şi analiza a două situaţii particulare de mişcare, care sunt absenţa mişcării şi respectiv discontinuitatea mişcării. Spre deosebire de metodele existente, modelele de mişcare propuse în [Ionescu 07b] iau în calcul posibilitatea estimării eronate a vectorilor de mişcare precum şi situaţiile de confuzie generate de aceasta (în care mişcări diferite pot furniza modelele de mişcare similare). Un exemplu este ilustrat în Figura 3.14. Informaţia de mişcare este obţinută în urma unei estimări pe blocuri de pixeli. Aceasta este comparată cu modelele predefinite pe baza unui set de reguli de decizie. Regulile sunt aplicate la nivel de vectori de orientare medii, ce sunt estimaţi în nouă regiuni disjuncte ale imaginii (vezi Figura 3.14).

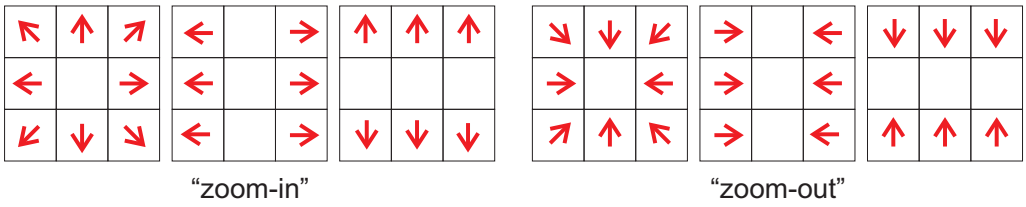


Figura 3.14: Exemplu de modele de mişcare pentru mişcarea camerei video de tip "zoom" [Ionescu 07b] (unui anumit tip de mişcare îi pot corespunde în realitate mai multe modele de mişcare).

[10] o "undişoară" ("wavelet") reprezintă o funcţie matematică ce este folosită la descompunerea unei anumite funcţii sau a unui semnal continuu, în componente frecvenţiale ce sunt apoi studiate la o rezoluţie ce corespunde scalei acestora. Funcţiile "wavelet" sunt copii scalate şi translatate ale unei forme de undă finită sau cu atenuare rapidă, numită şi funcţie "wavelet" de bază.

[11] transformata Hough este o tehnică generală de identificare a orientării şi a poziţiei anumitor tipuri de forme în imaginile digitale (de exemplu: linii, cercuri, elipse, etc.). Formele sunt determinate ca intersecţii ale unor curbe sau plane în spaţiul transformatei Hough, ce este determinat prin parametrizarea acestora.

O abordare inedită este propusă în [Ngo 00] unde caracterizarea mişcării camerei video şi a obiectelor este realizată pe baza analizei de volume spaţio-temporale extrase din imagini. O secvenţă de imagini poate fi văzută ca fiind un volum 3D în care primele două dimensiuni sunt date de dimensiunile spaţiale (x, y) iar a treia dimensiune este timpul t. Dintr-un alt punct de vedere, acest volum poate fi reprezentat ca fiind un ansamblu de straturi temporale 2D, fie orizontale (plan (x, t)), fie verticale (plan (y, t)). În spaţiul astfel format, diversele tipuri de mişcări ale camerei video vor fi indicate prin prezenţa modelelor orientate (vezi Figura 3.15). Caracterizarea acestora este realizată prin calculul a ceea ce autorii numesc histograme de tensori[12], unde tensorii sunt furnizaţi de derivata parţială după cele trei axe, ot, oX şi oY.

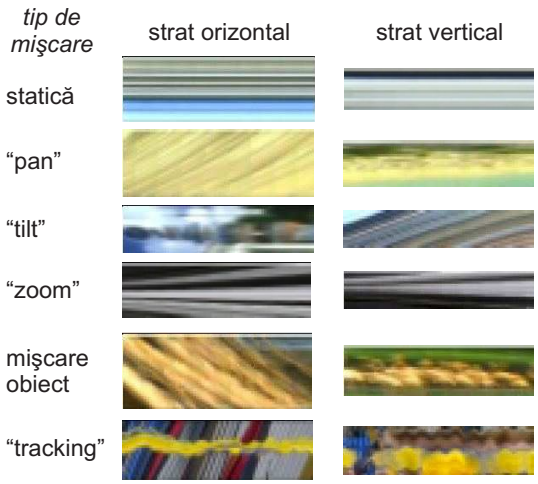


Figura 3.15: Modele orientate pentru diverse tipuri de mişcări [Ngo 00].

Raportat la metodele de analiză ce folosesc domeniul comprimat, analiza mişcării camerei video în domeniul spaţio-temporal al imaginii, are o complexitate de calcul mai semnificativă, dar rezultatele obţinute sunt mai precise. Domeniul spaţio-temporal are avantajul de a furniza o mai mare diversitate de informaţii decât coeficienţii MPEG.

3.3 Concluzii

În acest capitol am realizat o trecere în revistă a metodelor de analiză şi prelucrare a uneia dintre informaţiile definitorii ale unei secvenţe de imagini, şi anume *conţinutul de mişcare*. Metodele existente folosesc ca punct de

[12] vezi explicaţia de la pagina 85.

plecare pentru analiză estimarea câmpului de mişcare al pixelilor din imagine. Acesta este fie disponibil "a priori", cum este cazul fluxului MPEG, fie este determinat pe baza informaţiei spaţio-temporale.

În funcţie de precizia dorită, metodele de estimare a mişcării se împart în două categorii. Pe de o parte sunt metodele ce propun un câmp vectorial dens, calculat la nivel de pixel, metode ce sunt în general bazate pe estimarea fluxului optic în imagine. În acest caz, complexitatea de calcul este direct proporţională cu precizia câmpului rezultat. Pe de altă parte sunt metodele ce realizează estimarea la nivel de blocuri de pixeli. De-a lungul timpului, acestea s-au dovedit a furniza cel mai bun compromis între complexitatea de calcul şi precizia câmpului vectorial furnizat. Din acest motiv, metodele de estimare pe blocuri de pixeli au fost preferate de majoritatea standardelor de compresie video existente, precum MPEG, H.261, etc., unde timpul de prelucrare este critic.

Calitatea analizei conţinutului de mişcare al unei secvenţe de imagini, putem spune că este dependentă direct de precizia şi de calitatea metodei de estimare a mişcării folosite. În acest sens, metodele de estimare a mişcării sunt alese în funcţie de tipul aplicaţiei. De exemplu, pentru segmentarea obiectelor de interes din scenă, ce sunt reprezentate în imagine cu regiuni restrânse de pixeli, este preferabil un câmp de mişcare dens, care să ofere maximum de informaţie. Pe de altă parte, în cazul analizei globale a mişcării, precum analiza mişcării camerei video, este preferabilă o metodă mai rapidă, cu un nivel de detaliu mai scăzut, precum estimarea la nivel de blocuri de pixeli. De asemenea, în cazul aplicaţiilor în care timpul de prelucrare este limitat sau critic, se poate opta pentru recuperarea informaţiei de mişcare direct din fluxul video, dacă acesta este disponibil.

Principalele aplicaţii ale estimării mişcării în secvenţele de imagini sunt pe de-o parte analiza mişcării locale, ce este folosită de regulă pentru a segmenta şi caracteriza proprietăţile şi traiectoria anumitor obiecte de interes în scenă. Pe de altă parte, estimarea mişcării este folosită pentru caracterizarea mişcării globale a scenei, precum detecţia diverselor mişcări ale camerei video, sau la un nivel semantic superior, a anumitor tehnici de filmare.

În concluzie, informaţia de mişcare este un parametru deloc neglijabil pentru analiza conţinutului unei secvenţe de imagini, fiind însuşi motivul pentru care acestea există. Spre deosebire de alte surse de informaţie vizuală, precum imaginile statice, mişcarea face ca informaţia furnizată de secvenţele de imagini să fie mai aproape de realitate.

CAPITOLUL 4

Analiza de culoare

Rezumat: *Informaţia de culoare joacă un rol important în percepţia informaţiei vizuale. Aceasta ne permite înţelegerea proprietăţilor fizice ale obiectelor ce ne înconjoară, precum şi interacţia cu acestea prin senzaţiile de culoare ce ne sunt transmise. În ciuda faptului că informaţia fundamentală a unei secvenţe de imagini este dată de conţinutul de mişcare, din punct de vedere fiziologic, sistemul vizual uman este mult mai sensibil la schimbările de culoare. În acest capitol ne vom focaliza pe descrierea diverselor modalităţi de reprezentare şi de caracterizare a conţinutului de culoare, atât din punct de vedere sintactic cât şi perceptual.*

Unul dintre cele mai importante simţuri, poate chiar cel mai important, este vederea. Simţim, explorăm şi înţelegem lumea înconjurătoare folosindu-ne de percepţia vizuală. Fiecărui obiect sau entităţi cu care interacţionăm îi creăm mai întâi o imagine mentală a culorilor sale reprezentative, astfel: cerul este albastru, pădurea este verde, nisipul este galben, şi aşa mai departe.

Acest mecanism ne facilitează recunoaşterea şi identificarea obiectelor ce sunt similare. Mai mult, anumite culori individuale sau aranjamente de culori ne crează senzaţii particulare, ca de exemplu: albastrul ne dă senzaţia de rece, portocaliu ne transmite o senzaţie de cald, negru şi alb crează un contrast vizual, roşu în abundenţă dă o senzaţie de disconfort, etc. Astfel, pe

lângă informaţia vizuală furnizată de culoarea în sine, o importanţă ridicată în percepţia lumii înconjurătoare o au diversele relaţii ce pot exista între culori.

Inspirate de lumea reală, cercetările din domeniile vederii asistate de calculator şi a prelucrării de imagini încearcă să reproducă aceste simţuri umane pentru a dezvolta sisteme capabile de a furniza o înţelegere şi interpretare automată a informaţiei vizuale. Culoarea, în particular, a fost exploatată intensiv de mai bine de trei decade, pentru a descrie percepţia vizuală a imaginilor [Smeulders 00].

În cazul sistemelor de indexare automată după conţinut, informaţia de culoare a fost exploatată de sine stătător, aproape exclusiv în sistemele de indexare a imaginilor fixe [Bimbo 99] [Smeulders 00], aceasta fiind definitorie pentru conţinutul static spaţial al imaginii. În domeniul sistemelor de indexare după conţinut a secvenţelor de imagini, în care intervine şi dimensiunea temporală, nu există multe studii care să se focalizeze pe caracterizarea în termeni de distribuţie de culoare a conţinutului secvenţei. Acest lucru se datorează în principal faptului că doar culoarea în sine, în cele mai multe cazuri, nu este suficientă pentru înţelegerea conţinutului dinamic al secvenţei.

Astfel, majoritatea metodelor existente se folosesc de informaţia de culoare, fie în colaborare cu alte informaţii, precum mişcarea, distribuţia de plane, textură, etc., fie pentru a caracteriza anumite proprietăţi locale, la nivel de imagine, ale obiectelor de interes. Cu toate acestea, raportat la principala informaţie furnizată de o secvenţă de imagini, şi anume informaţia de mişcare, putem spune că din punct de vedere fiziologic, sistemul uman de percepţie vizuală este mai sensibil la schimbări de culoare decât la prezenţa mişcării.

Informaţia de culoare, în cazul secvenţelor de imagini, este conţinută la nivelul fiecărui cadru. Spre deosebire de imaginile statice, aceasta are o evoluţie temporală dată de evoluţia conţinutului din imagine. Astfel, metodele existente de analiză a culorii în secvenţele de imagini folosesc ca punct de plecare tehnici specifice analizei imaginilor statice ce sunt apoi extinse la dimensiunea temporală. De asemenea, pentru a evidenţia anumite proprietăţi ale culorilor, acestea sunt reprezentate în diverse spaţii de culoare. Spaţiile de culoare au fost concepute special în funcţie de necesităţile de prelucrare.

În cele ce urmează, vom face o trecere în revistă a diverselor tehnici folosite în secvenţele de imagini pentru a analiza conţinutul de culoare. În acest sens, mai întâi vom prezenta diversele modalităţi existente de reprezentare a culorilor, precum şi tehnicile de caracterizare a percepţiei culorii la nivel de imagine.

4.1 Spaţiile de culoare

La nivel de imagine, fiecare culoare este reprezentată ca fiind un punct într-un spaţiu de referinţă, de regulă tridimensional, care constituie ceea ce numim *spaţiul de culoare*. Ansamblul culorilor ce pot fi reprezentate într-un anumit spaţiu de culoare poartă numele de *gamut de culoare*. Particularizând la nivel de imagine, echivalentul gamutului de culoare este *paleta de culoare* a imaginii, aceasta fiind definită ca totalitatea culorilor, diferite, utilizate de imagine. Astfel, fiecare imagine, în funcţie de conţinutul acesteia, are o anumită paletă de culoare. Paleta de culoare poate fi *general valabilă* (unică) pentru un set finit de imagini sau *adaptivă*, şi astfel specifică fiecărei imagini în parte.

Conceperea diverselor spaţii de reprezentare a culorilor existente a fost motivată de însăşi existenţa sistemului vizual uman, astfel fiecare culoare fiind reprezentată în general după trei componente. În funcţie de natura acestora, spaţiile de culoare existente se împart în următoarele categorii [Trémeau 04]:

- **sisteme de culori primare**: acestea reprezintă culorile în funcţie de trei culori alese arbitrar, numite şi culori primare,
- **sisteme pe bază de luminanţă-crominanţă**: acestea au proprietatea de a separa informaţia de luminanţă (intensitate luminoasă) de componentele de crominanţă (culoare),
- **sisteme perceptuale**: acestea sunt sisteme uniforme din punct de vedere al percepţiei vizuale a culorilor,
- **sisteme de axe independente**: acestea au ca scop reprezentarea culorilor într-un spaţiu în care componentele de culoare sunt decorelate.

Totuşi, în ciuda existenţei unei diferenţieri pe categorii a sistemelor de culori existente, în mod uzual anumite sisteme pot fi încadrate ca aparţinând mai multor categorii deodată. Cum fiecare categorie în parte prezintă caracteristici sau proprietăţi ce sunt necesare în anumite etape de prelucrare, s-a căutat conceperea de sisteme de culoare, putem spune hibride, care să reunească aceste proprietăţi pentru a răspunde simultan acestor cerinţe. Se poate observa că nu există un sistem universal de culoare care să fie general valabil pentru orice aplicaţie, mai mult, în multe situaţii, nici unul dintre sistemele existente nesatisfăcând cerinţele de prelucare [Trémeau 04].

În cele ce urmează vom face o trecere în revistă a particularităţilor sistemele de culoare reprezentative din fiecare dintre categoriile enumerate anterior.

4.1.1 Sisteme de culori primare

Sistemele din această categorie folosesc pentru a reprezenta culorile principiul *trivarianţei vizuale*. Astfel, fiecare culoare existentă poate fi reprodusă vizual identic, în anumite condiţii de observare, prin amestecul matematic, în proporţii unice, a trei culori numite şi *culori primare*. Aceste culori primare sunt alese arbitrar şi au proprietatea că nici una dintre ele nu poate fi reprodusă ca un amestec al celorlalte două [Young 02].

Astfel, fiecare culoare C poate fi exprimată pe baza principiului trivarianţei vizuale ca fiind dată de ecuaţia:

$$C = p_1 \cdot C_1 + p_2 \cdot C_2 + p_3 \cdot C_3 \tag{4.1}$$

unde C_i, $i \in \{1, 2, 3\}$, reprezintă cele trei culori primare iar coeficienţii p_i reprezintă ponderea acestora la definirea culorii C.

Din această categorie de sisteme, putem menţiona ca reprezentative spaţiile de culoare RGB, CMY şi XYZ.

Spaţiul de culoare RGB

Spaţiul de culoare RGB este un spaţiu aditiv ce a fost inspirat de modalitatea fiziologică de reprezentare a culorilor în sistemul vizual uman. Fiecare culoare existentă, C, este exprimată ca un amestec aditiv a trei culori primare, şi anume roşu ("Red"), verde ("Green") şi respectiv albastru ("Blue"), astfel:

$$C = p_1 \cdot R + p_2 \cdot V + p_3 \cdot B \tag{4.2}$$

Totuşi, această ecuaţie caracterizează doar nuanţa culorii rezultate din sinteza aditivă şi în nici un caz intensitatea sa luminoasă [Délibéré 89]. Pentru a caracteriza luminanţa culorii C este necesar să cunoaştem luminanţa fiecărei dintre cele trei culori primare folosite, R, G şi B. În funcţie de aceasta, există mai multe sisteme bazate pe cele trei culori primare, astfel: sistemul de culori primare CIE definit luând ca referinţă iluminantul E[1], sistemul de culori primare NTSC definit luând ca referinta iluminantul C sau sistemul de culori primare EBU ("European Brodcasting Union") ce foloseşte ca referinţă iluminantul D_{65} [Trémeau 04].

[1]sursele de lumină sunt caracterizate de repartiţia spectrală a energiei sau altfel spus prin cantitatea de energie emisă pe un anumit interval de lungime de undă. Anumite surse de lumină ce corespund unor situaţii uzuale au fost normalizate de CIE ("Commission Internationale de l'Eclairage") sub numele de iluminanţi. Astfel, de exemplu iluminantul E corespunde unei lumini de energie egală, iluminantul C corespunde unei lumini medii de zi cu o temperatură de culoare de aproximativ $6770K$, iluminantul D_{65} corespunde în mare unei lumini ce provine de la un cer albastru, cu 3/5 nori albi, măsurată în jurul orei 10 dimineaţa în luna septembrie.

Sistemul RGB este utilizat cu precădere în majoritatea dispozitivelor "hardware" de reproducere a culorilor ce sunt destinate marelui public, precum aparatele foto digitale, monitoarele CRT şi LCD, imprimantele color, etc., şi stă la baza arhitecturii sistemelor de prelucrare actuale. Din această cauză, trecerea la alte spaţii de culoare se face în cele mai multe cazuri pornind de la valorile RGB pe baza unei transformări matematice.

Gamutul de culoare al spaţiului RGB este reprezentat de un cub în sistemul de coordonate R, G şi B (vezi Figura 4.1). În acesta, negrul se găseşte la originea sistemului iar nivelurile de gri se găsesc pe diagonala principală, fiind reprezentate de valori egale ale componentelor R, G şi respectiv B.

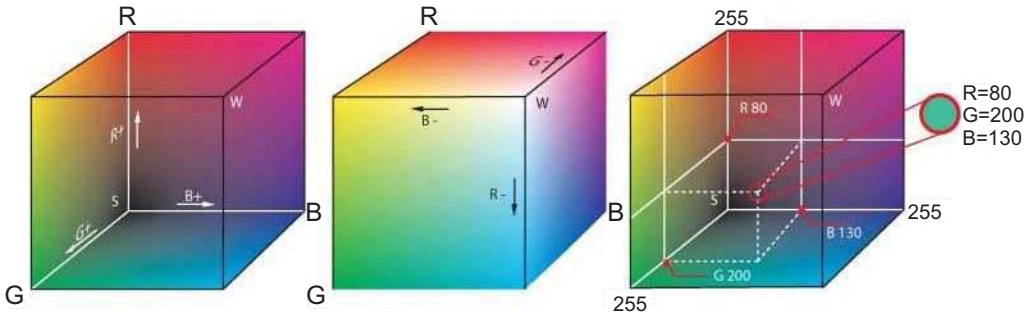


Figura 4.1: Cubul de culoare RGB (R-roşu, G-verde, B-albastru, W-alb, sursă imagine Wikipedia "http://en.wikipedia.org/wiki/RGB_color_space").

Spaţiul de culoare astfel obţinut este un spaţiu cu variaţie neuniformă în care distanţa Euclidiană dintre două culori nu corespunde distanţei perceptuale dintre acestea.

Spaţiul de culoare CMY

Spaţiul de culoare CMY, spre deosebire de RGB, este obţinut prin ceea ce se numeşte sinteză substractivă a culorilor primare turcoaz (Cyan), magenta şi galben (Yellow). Culorile primare folosite de cele două spaţii de culoare, RGB şi respectiv CMY, sunt complementare. Astfel, albastrul este complementar lui galben care el însuşi este obţinut ca un amestec de roşu şi verde, roşu este complementar lui turcoaz care este obţinut din albastru şi verde, iar verde este complementar lui magenta care este obţinut din roşu şi albastru (vezi Figura 4.2).

Trecerea de la spaţiul RGB la CMY este dată de următoarea ecuaţie:

$$\begin{bmatrix} C \\ M \\ Y \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{4.3}$$

unde (C, M, Y) reprezintă culoarea (R, G, B) în spaţiul CMY.

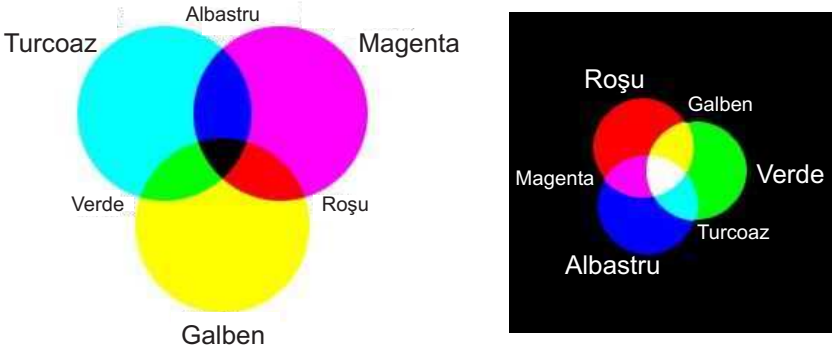


Figura 4.2: Comparaţie între amestecul substractiv (CMY) şi aditiv (RGB), sursă imagine "http://akaroosto.com/www.whoride.com/roosto/physics/sub_vs_add.html".

Spaţiul de culoare CMY, în varianta CMYK, la care se adaugă separat negrul (K-"Key"), este folosit cu predilecţie pentru imprimarea pe suport fizic, ca de exemplu în tipografii, deoarece modalitatea sa de constituire descrie însăşi procesul de imprimare. Adăugarea separată a negrului este motivată în principal de faptul ca negrul generat prin amestecul fizic al celor trei culori primare, C, M şi Y, nu oferă o calitate suficientă imprimării, cum este cazul imprimării detaliilor fine ale fonturilor de caractere.

Spaţiul de culoare XYZ

Sistemul XYZ este sistemul de referinţă colorimetrică definit de CIE ("Commission Internationale de l'Eclairage"). Acesta este determinat pornind de la spaţiul de culoare RGB printr-o transformare liniară, transformare ce ia în calcul nu numai coordonatele tricromatice ale fiecărei culori primare, R, G şi B, precum şi coordonatele tricromatice ale albului de referinţă W. Astfel, orice culoare (R, G, B) poate fi exprimată în coordonate (X, Y, Z) folosind următoarea ecuaţie:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 2.7690 & 1.7518 & 1.1300 \\ 1.0000 & 4.5907 & 0.0601 \\ 0.0000 & 0.0565 & 5.5943 \end{bmatrix} \times \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{4.4}$$

unde (2.7690, 1.0000, 0.0000), (1.7518, 4.5907, 0.0565) şi (1.1300, 0.0601, 5.5943) reprezintă coordonatele XYZ ale celor trei culori primare: roşu, verde şi respectiv albastru.

În funcţie de iluminantul de referinţă folosit, trecerea la spaţiul XYZ se face folosind matrice de transformări diferite, astfel:

- transformarea spaţiului NTSC RGB folosind iluminantul C, care este şi transformarea cea mai utilizată în domeniul prelucrării imaginilor color, este dată de:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.607 & 0.174 & 0.200 \\ 0.299 & 0.587 & 0.114 \\ 0.000 & 0.066 & 1.116 \end{bmatrix} \times \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{4.5}$$

- transformarea spaţiului CIE RGB folosind iluminantul A[2] este dată de:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.892 & 0.330 & 0.083 \\ 0.322 & 0.863 & 0.004 \\ 0.000 & 0.011 & 0.409 \end{bmatrix} \times \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{4.6}$$

- transformarea spaţiului CIE RGB folosind iluminantul C este dată de:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.166 & 0.125 & 0.093 \\ 0.060 & 0.327 & 0.005 \\ 0.000 & 0.004 & 0.460 \end{bmatrix} \times \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{4.7}$$

Spaţiul XYZ este definit astfel încât să respecte o serie de constrângeri, şi anume:

- orice culoare fizică monocromatică trebuie să fie caracterizată de valori tristimulus pozitive,
- componenta Y trebuie să fie o măsură a intensităţii luminoase,
- pentru o lumină albă, valorile tristimulus trebuie să fie egale.

Matricea transformării în spaţiul XYZ nu este o matrice unitară[3] şi astfel transformarea nu reprezintă o rotaţie a cubului RGB cu tot cu sistemul de coordonate, gamutul de culoare XYZ fiind în acest caz un paralelipiped înclinat [Vertan 08].

[2]iluminantul CIE standard de tip A corespunde unei lumini de interior tipice furnizate de o iluminare cu filament din tungsten ce are o temperatură de culoare de aproximativ 2856K.

[3]o transformare se numeşte unitară, dacă matricea transformării inverse este egală cu transpusa conjugată a matricei transformării directe.

Cum componenta Y reflectă intensitatea luminoasă, atunci informaţia cromatică a unei culori poate fi exprimată în funcţie de doi parametri derivaţi, notaţi x şi y, ce sunt daţi de relaţiile următoare:

$$x = \frac{X}{X + Y + Z} \tag{4.8}$$

$$y = \frac{Y}{X + Y + Z} \tag{4.9}$$

Spaţiul de culoare derivat, definit de cei doi parametri, x şi y, precum şi de componenta de intensitate luminoasă Y este cunoscut sub numele de CIE xyY şi este folosit la scară largă pentru specificarea culorilor. Diagrama de cromaticitate xy este prezentată în Figura 4.3.

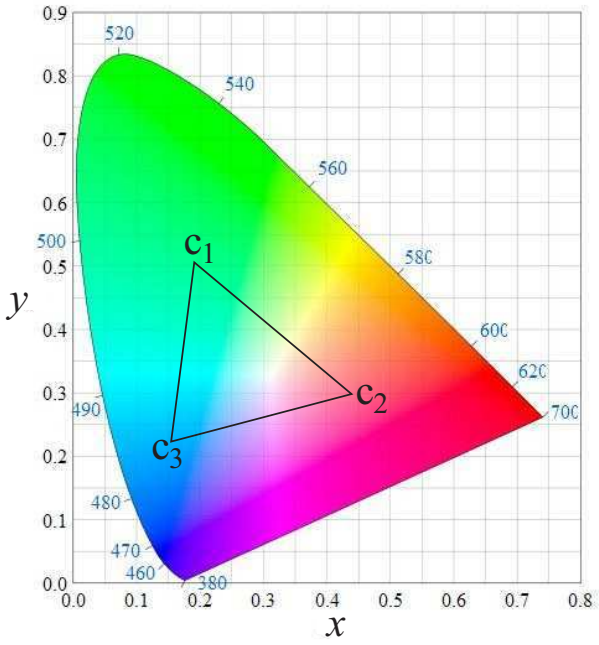


Figura 4.3: Diagrama de cromaticitate CIE (valorile de pe curba ce marchează marginea graficului reprezintă lungimi de undă exprimate în nanometrii, sursă Wikipedia "http://en.wikipedia.org/wiki/CIE_1931_color_space").

Reprezentată în acest fel, diagrama de cromaticitate pune în evidenţă o serie de proprietăţi interesante ale spaţiului de culoare XYZ, astfel:

- aceasta prezintă toate culorile vizibile de către o persoană normală, sau ceea ce se numeşte *gamutul sistemului vizual uman*. Conturul curbat al

gamutului corespunde astfel culorilor monocromatice (vezi lungimile de undă din Figura 4.3). Linia dreaptă din partea de jos a gamutului este numită linia culorilor purpurii. Culorile mai puţin saturate se găsesc în interiorul graficului având albul în centru.

- toate culorile cromatice vizibile au valorile X, Y, şi Z nenegative.
- alegând două culori diferite arbitrare din diagrama de cromaticitate, toate culorile ce pot fi formate cu acestea se regăsesc pe linia ce uneşte cele două puncte. Similar, în cazul a trei puncte diferite, toate culorile ce pot fi formate cu acestea se găsesc în interiorul triunghiului format de acestea (vezi Figura 4.3), şi aşa mai departe. În concluzie, gamutul de culoare reprezentat în diagrama de cromaticitate este o formă *geometrică convexă*.
- distanţa dintre culori, în diagrama de cromaticitate xy, nu corespunde distanţei perceptuale dintre culori.

4.1.2 Sisteme pe bază de luminanţă-crominanţă

Sistemele de culori din această categorie au proprietatea de a separa componentă de intensitate luminoasă de componentele cromatice. Din această categorie putem enumera sistemele de tip YC_bC_r şi sistemele antagoniste [Trémeau 04].

Spaţiile de culoare de tip YC_bC_r

Acest spaţiu de culoare a fost creat iniţial pentru a asigura compatibilitatea dintre televizoarele color şi cele alb-negru. Astfel, un semnal color de tip YC_bC_r putea fi vizionat şi pe un receptor alb-negru datorită separării componentei de luminanţă, Y, de cele cromatice, C_b şi C_r. Trecerea de la spaţiul RGB la spaţiul YC_bC_r se face printr-o simplă transformare liniară ai cărei coeficienţi diferă de la un standard de televiziune la altul.

Cele două valori de crominanţă, C_b şi respectiv C_r, ale unei anumite culori exprimată în coordonate RGB, pot fi calculate în funcţie de patru coeficienţi, a_1, a_2, b_1 şi b_2 (specifici standardului considerat) şi de informaţia de luminanţă Y, în felul următor:

$$C_b = a_1 \cdot (R - Y) + b_1 \cdot (B - Y) \tag{4.10}$$

$$C_r = a_2 \cdot (R - Y) + b_2 \cdot (B - Y) \tag{4.11}$$

În realitate, există mai multe sisteme de tip YC_bC_r, ca de exemplu sistemul YIQ ce corespunde standardului NTSC[4], sistemul YUV ce corespunde standardului PAL[5] şi sistemul YD_bD_r ce corespunde standardului SECAM[6]. Dintre acestea, sistemul cel mai frecvent utilizat în domeniul imagisticii color este sistemul YIQ.

Trecerea de la spaţiul de culoare RGB la spaţiul YIQ este dată de transformarea următoare:

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.273 & -0.322 \\ 0.212 & -0.522 & 0.315 \end{bmatrix} \times \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} \tag{4.12}$$

unde (R', G', B') reprezintă coordonatele (R, G, B) după corecţia de gamma[7] următoare:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} R^{1/2.2} \\ G^{1/2.2} \\ B^{1/2.2} \end{bmatrix} \tag{4.13}$$

În cazul în care se doreşte o caracterizare mai riguroasă a informaţiei de culoare, spaţiul YIQ poate fi descris şi în coordonate cilindrice [A.R.Weeks 95], astfel:

$$H = arctan\left(\frac{B' - Y}{R' - Y}\right) \tag{4.14}$$

$$S^2 = (R' - Y)^2 + (B' - Y)^2 \tag{4.15}$$

unde H reprezintă nuanţa de culoare ("Hue") iar S saturaţia culorii ("Saturation").

Spaţiile de culoare antagoniste

Spaţiile de culoare din această categorie sunt fundamentate pe teoria culorilor opuse elaborată de Young şi Hering. Aceasta afirmă că informaţia

[4]NTSC este prescurtarea pentru "National Television System Committee" şi reprezintă un standard de codare a semnalului TV analogic color pe 525 de linii şi la o cadenţă de 30 imagini pe secundă.

[5]vezi explicaţia de la pagina 9.

[6]SECAM este prescurtarea pentru "Séquentiel Couleur à Mémoire" şi reprezintă un standard de codare a semnalului TV analogic color pe 625 de linii şi la o cadenţă de 25 de imagini pe secundă.

[7]datorită faptului că sistemul vizual uman nu are un răspuns liniar la informaţia luminoasă, în cazul în care luminanţa este codată cu un număr redus de valori (de exemplu 256) este necesară utilizarea acestora în concordanţă cu proprietăţile sistemului vizual uman pentru a obţine o eficienţă optimală. Această operaţie este cunoscută sub numele de corecţie de gama. Corecţia de gama se poate exprima sub forma unei ecuaţii de tipul $S = E^{\gamma}$, unde E reprezintă semnalul căruia i se aplică corecţia iar S este semnalul corectat.

de culoare recepţionată de sistemul vizual uman este transmisă creierului sub forma a trei semnale, unul acromatic ce corespunde contrastului negru-alb, şi două semnale cromatice, unul corespunzând contrastului verde-roşu şi celălalt contrastului albastru-galben [Wyszecki 82].

Un exemplu este spaţiul de culoare AC_1C_2, introdus pentru prima oară în [Faugeras 79], ce este definit de următorul sistem de ecuaţii:

$$A = 22.6 \cdot (0.612 \cdot logL + 0.369 \cdot logM + 0.019 \cdot logS) \tag{4.16}$$

$$C_1 = 64(logL - logM) \tag{4.17}$$

$$C_2 = 10(logL - logS) \tag{4.18}$$

unde (L, M, S) reprezintă coordonatele culorii în sistemul primar LMS[8] iar (A, C_1, C_2) reprezintă noile coordonate în spaţiul AC_1C_2.

Trecerea la spaţiul LMS se poate realiza pornind de la spaţiul de culoare XYZ prin transformarea următoare:

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.15514 & 0.54312 & -0.03286 \\ -0.15514 & 0.45684 & 0.03286 \\ 0.0 & 0.0 & 0.00801 \end{bmatrix} \times \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{4.19}$$

În ciuda proprietăţilor interesante ale spaţiilor de culoare antagoniste, marea majoritate a acestora sunt puţin utilizate în imagistica color datorită faptului că sunt fie abordări experimentale în curs de validare, fie modele foarte simplificate [Trémeau 04].

4.1.3 Sisteme perceptuale

Sistemele perceptuale de reprezentare a culorilor sunt sisteme uniforme din punct de vedere al percepţiei vizuale. Astfel, în acestea distanţa matematică dintre două culori este proporţională cu distanţa perceptuală dintre acestea. Din această categorie putem enumera ca cele mai reprezentative: spaţiul de culoare $L^*a^*b^*$, spaţiile de culoare geometrice şi spaţiul de culoare al lui Munsell.

Spaţiul de culoare $L^*a^*b^*$ şi $L^*u^*v^*$

Datorită popularităţii acestuia, spaţiul de culoare $L^*a^*b^*$ poate fi considerat ca unul dintre sistemele de referinţă al CIE folosit pentru evaluarea distanţei

[8]sistemul de culori primare LMS este sistemul de referinţă în domeniul fiziologiei sistemului vizual. Acesta are avantajul de a folosi culori primare ce sunt în concordanţă directă cu semnalul de intrare folosit de sistemul vizual uman, fiind singurul sistem primar ce reproduce activitatea reală a celulelor cu con din retina umană.

dintre culori [Trémeau 04]. Trecerea la sistemul de coordonate $L^*a^*b^*$ se face de regulă pornind de la spaţiul XYZ pe baza unei transformări, de această dată, neliniare ce ia în calcul şi coordonatele tricromatice ale albului de referinţă, W, folosit.

Astfel, o culoare dată de tripletul (X, Y, Z) se poate exprima în coordonate $L^*a^*b^*$ pe baza ecuaţiilor următoare:

$$L^* = \begin{cases} 116 \cdot y^{1/3} - 16 & \text{dacă } y > 0.008856 \\ 903.3 \cdot y & \text{altfel} \end{cases} \tag{4.20}$$

$$a^* = 500 \cdot [f(x) - f(y)] \tag{4.21}$$

$$b^* = 200 \cdot [f(y) - f(z)] \tag{4.22}$$

unde:

$$x = \frac{X}{X_W}, \quad y = \frac{Y}{Y_W}, \quad z = \frac{Z}{Z_W} \tag{4.23}$$

(X_W, Y_W, Z_W) reprezintă coordonatele XYZ ale albului de referinţă iar funcţia $f()$ este dată de ecuaţia următoare:

$$f(t) = \begin{cases} t^{1/3} & \text{dacă } t > 0.008856 \\ 7.787 \cdot t + 0.137931 & \text{altfel} \end{cases} \tag{4.24}$$

În spaţiul $L^*a^*b^*$, fiind un spaţiu perceptual, distanţa perceptuală dintre două culori, $C_1 = (L_1^*, a_1^*, b_1^*)$ şi $C_2 = (L_2^*, a_2^*, b_2^*)$, poate fi evaluată direct pe baza distanţei Euclidiene dintre coordonatele acestora, astfel:

$$\Delta E_{C_1,C_2}^2 = (\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2 \tag{4.25}$$

unde $\Delta E_{C_1,C_2}$ quantifică diferenţa dintre culorile C_1 şi respectiv C_2 iar $\Delta L^* = L_1^* - L_2^*$, $\Delta a^* = a_1^* - a_2^*$ şi $\Delta b^* = b_1^* - b_2^*$.

Gamutul de culoare al spaţiului $L^*a^*b^*$ este reprezentat ca o sferă de culoare (vezi Figura 4.4) în care spre polul nord sunt figurate culorile deschise, finalizându-se cu alb, iar spre polul sud culorile întunecate, finalizându-se cu negru.

Datorită construcţiei particulare, spaţiul $L^*a^*b^*$ poate fi încadrat ca aparţinând la mai multe categorii de sisteme de reprezentare a culorilor, astfel acesta poate fi considerat ca un spaţiu antagonist, deoarece componenta L^* pune în evidenţă contrastul negru-alb, componenta a^* contrastul verde-roşu iar componenta b^* contrastul albastru-galben (vezi Figura 4.4). De asemenea, spaţiul $L^*a^*b^*$ se încadrează şi în categoria spaţiilor de culoare de tip YC_bC_r, deoarece componenta de luminaţă, L^*, este separată de cea de crominanţă dată de a^* şi respectiv b^*.

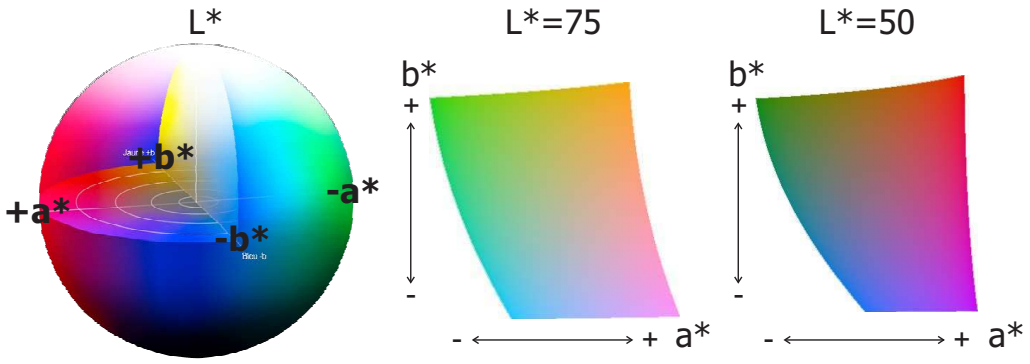


Figura 4.4: Sfera de culoare $L^*a^*b^*$ (sursă Wikipedia "http://en.wikipedia.org/wiki/Lab_color_space").

Prin trecerea spaţiului $L^*a^*b^*$ în coordonate cilindrice obţinem un nou spaţiu de culoare în care culorile sunt grupate în funcţie de nuanţă şi saturaţie. Acesta este cunoscut sub numele de spaţiul de culoare LCH şi este dat de relaţiile următoare:

$$H = arctan\left(\frac{b^*}{a^*}\right) \tag{4.26}$$

$$C^2 = a^{*^2} + b^{*^2} \tag{4.27}$$

unde H reprezintă nuanţa iar C, numit şi "chroma", oferă informaţii despre saturaţia culorii.

Tot în această categorie putem menţiona un spaţiu de culoare similar cu spaţiul $L^*a^*b^*$, şi anume spaţiul $L^*u^*v^*$. Cu toate că spaţiul $L^*u^*v^*$ nu este analog spaţiului $L^*a^*b^*$, pentru anumite intervale de culoare acesta furnizează rezultate similare [Agoston 87]. Dacă componenta L^* a spaţiului $L^*u^*v^*$ este calculată în acelaşi mod ca pentru $L^*a^*b^*$ (vezi ecuaţia 4.20), componentele u^* şi v^* sunt calculate în mod diferit, astfel:

$$u^* = 13 \cdot L^* \cdot (u' - u'_W) \tag{4.28}$$

$$v^* = 13 \cdot L^* \cdot (v' - v'_W) \tag{4.29}$$

unde

$$u' = \frac{4 \cdot X}{X + 15 \cdot Y + 3 \cdot Z} \tag{4.30}$$

$$v' = \frac{9 \cdot Y}{X + 15 \cdot Y + 3 \cdot Z} \tag{4.31}$$

iar (X_W, Y_W, Z_W) reprezintă coordonatele XYZ ale albului de referinţă folosit.

Spaţiile de culoare geometrice

Sistemele de culoare din această categorie folosesc o partiţionare uniformă a culorilor în funcţie de trei informaţii, şi anume: nuanţă, saturaţie şi luminanţă. Diferenţa dintre acestea constă în principal în unităţile folosite la partiţionare precum şi în dinamica fiecărei dintre cele trei axe de culoare. Sistemele geometrice au fost concepute în scopul de a clasa culorile pe baza percepţiei psiho-vizuale a acestora, culorile fiind reperate de această dată într-un sistem de coordonate geometrice. Spaţiile de culoare geometrice constituie reprezentări descriptive ale informaţiei de culoare [Trémeau 04].

Cele mai reprezentative spaţii din această categorie sunt spaţiile de culoare HSL (H-nuanţă, S-saturaţie, L-luminozitate) şi respectiv HSV (H-nuanţă, S-saturaţie, V-valoare). O alternativă similară, mai puţin standardizată, a spaţiilor HSL şi HSV o constituie spaţiile de culoare HSI şi respectiv HSB, unde I reprezintă intensitatea iar B strălucirea culorii ("brightness").

Atât spaţiul HSL cât şi HSV reprezintă culorile din punct de vedere matematic folosind o reprezentare cilindrică (vezi Figura 4.5). Astfel, nuanţa de culoare este dată de unghi, saturaţia culorii este dată de distanţa faţă de axa principală iar luminozitatea este dată de poziţia de-a lungul axei principale. Nivelurile de gri se vor găsi astfel pe axa centrală, pornind de la negru şi finalizând cu alb.

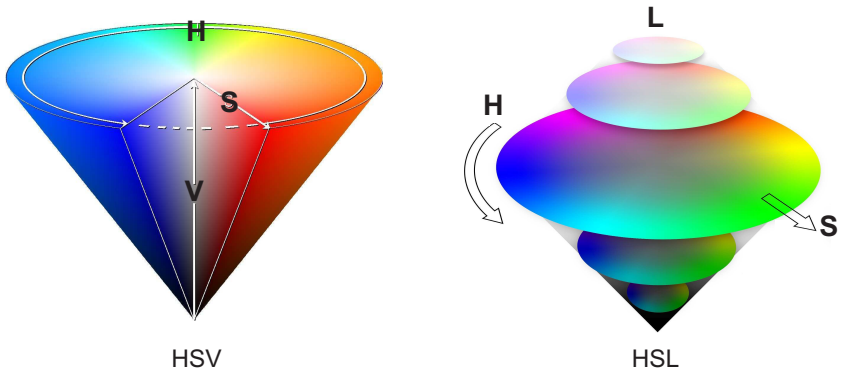


Figura 4.5: Spaţiile de culoare HSV (con invers) şi HSL (dublu con, sursă Wikipedia "http://en.wikipedia.org/wiki/HSV_color_space").

Cu toate că cele două tipuri de reprezentări sunt similare din punct de vedere al obiectivului urmărit, acestea diferă totuşi din punct de vedere al abordării folosite. Gamutul de culoare al sistemului HSV reprezintă conceptual un con invers în care negrul se găseşte în vârful conului iar culorile de

saturaţie maximă se găsesc pe cercul bazei. Pe de altă parte, gamutul de culoare al sistemului HSL reprezintă conceptual un con dublu în care albul se găseşte în vârful din Nord, negrul în vârful din Sud iar culorile de saturaţie maximă se găsesc pe cercul median.

Trecerea la spaţiul HSV se face pornind de la reprezentarea RGB pe baza următoarelor ecuaţii:

$$h_{HSV} = \begin{cases} 0 & \text{daca } max = min \\ 60 \cdot \frac{g-b}{max-min} + 0 & \text{dacă } max = r \text{ si } g \geq b \\ 60 \cdot \frac{g-b}{max-min} + 360 & \text{dacă } max = r \text{ si } g < b \\ 60 \cdot \frac{b-r}{max-min} + 120 & \text{dacă } max = g \\ 60 \cdot \frac{r-g}{max-min} + 240 & \text{dacă } max = b \end{cases} \tag{4.32}$$

$$s_{HSV} = \begin{cases} 0 & \text{dacă } max = 0 \\ \frac{max-min}{max} & \text{altfel} \end{cases} \tag{4.33}$$

$$v_{HSV} = max \tag{4.34}$$

unde r, g şi b reprezintă valorile normalizate între 0 şi 1 ale celor trei componente R, G şi respectiv B[9] iar max şi min reprezintă valoarea maximă şi respectiv minimă a componentelor normalizate r, g, b. Valorile astfel obţinute sunt normalizate între 0 şi 360 pentru componenta h şi între 0 şi 1 pentru componentele s şi v.

În ceea ce priveşte trecerea la spaţiul HSL, componenta normalizată h este dată de aceeaşi ecuaţie ca şi în cazul spaţiului HSV (vezi ecuaţia 4.32), în schimb componentele normalizate s şi l sunt date de ecuaţiile următoare:

$$s_{HSL} = \begin{cases} 0 & \text{dacă } max = min \\ \frac{max-min}{max+min} & \text{dacă } l \leq 0.5 \\ \frac{max-min}{2-(max+min)} & \text{dacă } l > 0.5 \end{cases} \tag{4.35}$$

$$l_{HSL} = \frac{1}{2}(max + min) \tag{4.36}$$

Principalul inconvenient al transformărilor bazate pe partiţionarea geometrică a spaţiului de culoare constă în reversibilitatea limitată a acestora. Astfel, de cele mai multe ori la revenirea în spaţiul iniţial de culoare, de regulă RGB, culorile sunt diferite de cele originale. Acest lucru se datorează faptului că marea majoritate a metodelor de transformare sunt bazate pe calculul valorilor minimale şi respectiv maximale ale coordonatelor RGB (vezi ecuaţiile 4.32).

[9] de regulă pentru valori RGB cuprinse între 0 şi 255, $r = \frac{R}{255}$, $g = \frac{G}{255}$ şi $b = \frac{B}{255}$.

Spaţiul de culoare al lui Munsell

Creat de profesorul Albert H. Munsell în prima decadă a secolului douăzeci, spaţiul de culoare ce îi poartă numele a fost conceput în principal pentru a fi utilizat în domeniul artei. Acesta se bazează pe observaţii subiective ale raportului dintre culori şi nu pe măsurarea directă a proprietăţilor de culoare [Bimbo 99]. Spaţiul obţinut este un spaţiu perceptual, uniform, bazat pe amestecul substractiv de culoare, în care culorile sunt reprezentate în funcţie de modul în care sunt percepute de obsevatori umani.

Fiecare culoare este reprezentată în funcţie de trei informaţii, şi anume: luminanţă ("Value"), nuanţă ("Hue") şi saturaţie ("Chroma"), într-un sistem de reprezentare de tip cilindric. Astfel, nuanţa este dată de unghiul în grade făcut de culoare în cercul perpendicular pe axa verticală a sistemului (axa nivelurilor de gri), saturaţia este măsurată radial ca distanţă faţă de axa verticală iar luminanţa este măsurată vertical de-a lungul axei principale. Fiecare dintre aceste trei axe sunt discretizate într-un număr limitat de intervale, ca de exemplu: luminanţa este exprimată cu valori întregi de la 0 (negru) la 10 (alb).

Repartizarea culorilor în acest mod nu este întâmplătoare ci este rezultatul măsurătorilor subiective ale percepţiei de culoare efectuate de Munsell. Pe fiecare dintre cele trei dimensiuni, culorile sunt alese pe cât posibil uniforme din punct de vedere al percepţiei, din această cauză "solidul" de culoare obţinut are o formă neregulată (vezi Figura 4.6).

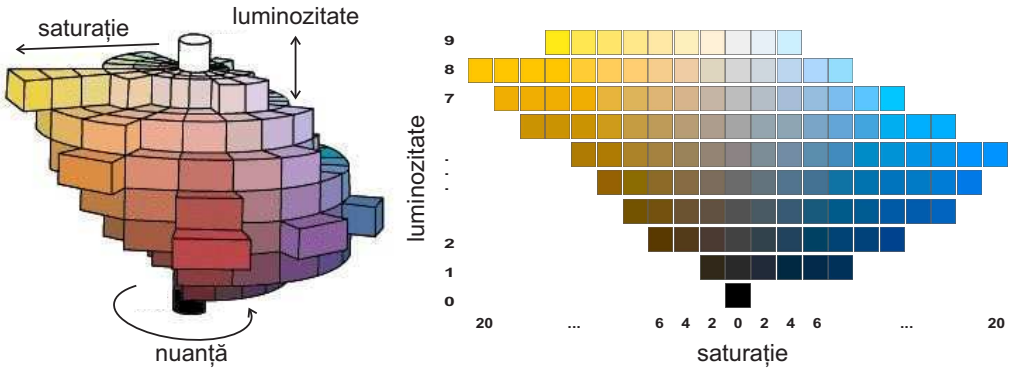


Figura 4.6: Spaţiul de culoare al lui Munsell ("solidul" de culoare şi o secţiune verticală a acestuia, sursă Computer Science Lab "http://www.computersciencelab.com/Direct3DTut1.htm").

Fiind un spaţiu perceptual, spaţiul de culoare al lui Munsell a stat la baza concepţiei spaţiilor de culoare perceptuale actuale, precum spaţiul $L^*a^*b^*$ sau

familia de spaţii HSV (vezi secţiunile anterioare).

4.1.4 Sisteme de axe independente

În funcţie de paleta de culoare a imaginii analizate şi de sistemul de reprezentare a culorilor folosit, se poate observa o corelaţie mai puternică sau mai slabă între diferitele componente de culoare. Astfel, pentru anumite etape de prelucrare este uneori mai "rentabil" să se folosească un sistem de reprezentare a culorilor în care componentele de culoare să fie complet decorelate[10], cum este cazul sistemelor de axe independente.

Principalul inconvenient al acestor tipuri de sisteme de reprezentare a culorilor constă în dependenţa lor de distribuţia de culoare considerată, astfel coeficienţii transformării fiind diferiţi de la o imagine la alta. Totuşi, în urma studiilor efectuate în această direcţie, s-a observat că în practică coeficienţii transformării pot fi similari de la o distribuţie de culoare la alta. Astfel, matricea de transformare nu este strict necesar să fie recalculată de la o imagine la alta, ci mai general, aceasta va fi specifică fiecărei aplicaţii în parte [Trémeau 04].

Dintre studiile semnificative realizate în această direcţie putem menţiona cele prezentate în [Ohta 80] unde sunt demonstrate următoarele ipoteze:

- pentru majoritatea imaginilor color, axa principală obţinută în urma descompunerii în componente principale (PCA - "Principal Component Analysis"[11]) se suprapune cu axa de luminanţă,
- pentru marea majoritate a imaginilor color, informaţia de culoare, fie în totalitate, fie parţial, este dată de prima şi de a doua componentă principală,
- pentru marea majoritate a imaginilor color, componentele de culoare sunt fie strâns corelate, fie relativ decorelate. Astfel, în timp ce componentele de culoare ale sistemelor RGB şi XYZ sunt puternic corelate, componentele sistemelor $L^*a^*b^*$ şi YC_bC_r sunt relativ decorelate,

[10] în statistică, corelaţia indică gradul de dependenţă liniară dintre două variabile aleatoare. Astfel, două variabile aleatoare sunt decorelate, sau independente, dacă nici una dintre ele nu oferă informaţii cu privire la valorile celeilalte, sau altfel zis, dacă acestea nu pot fi determinate una în funcţie de cealaltă.

[11] descompunerea în componente principale, sau PCA, este definită ca fiind transformarea liniară ortogonală a datelor de intrare într-un nou sistem de coordonate în care varianţa cea mai semnificativă a acestora este reprezentată pe prima coordonată, numită şi prima componentă principală, a doua varianţă este reprezentată pe a doua coorodonată, şi aşa mai departe. Astfel, transformarea PCA este teoretic optimală din punct de vedere al erorii pătratice.

- pentru marea majoritate a imaginilor color, sistemul de culoare decorelat obţinut pe baza transformatei Karhunen-Loeve[12] tinde să se confunde cu sistemul de axe $I_1I_2I_3$, ce este descris de următoarele ecuaţii:

$$I_1 = \frac{1}{3}(R + G + B) \tag{4.37}$$

$$I_2 = \frac{1}{2}(R - B) \tag{4.38}$$

$$I_3 = \frac{1}{4}(2 \cdot G - R - B) \tag{4.39}$$

  unde (R, G, B) reprezintă coordonatele RGB.

În concluzie, dezvoltarea diverselor sisteme de reprezentare a culorilor existente a fost motivată în principal de necesitatea de a evidenţia anumite proprietăţi de culoare, proprietăţi ce nu erau vizibile sau măsurabile în spaţiile de culoare clasice folosite de dispozitivele de prelucrare a imaginilor.

De exemplu, familia de sisteme de culoare HSV permit separarea nuanţei culorii de alte informaţii, precum saturaţia sau luminozitatea, distribuţia de culoare putând fi astfel interpretată pe baza culorilor primare din care aceasta a fost derivată. Un alt exemplu sunt spaţiile de tip $L^*a^*b^*$, în care culorile sunt reprezentate în funcţie de percepţia vizuală a acestora, permiţând astfel evaluarea similarităţii dintre culori pe baza calculului distanţei Euclidiene dintre acestea. Alte spaţii de culoare, precum familia de spaţii YC_bC_r permit separarea informaţiei de luminanţă de informaţia cromatică fiind eficiente pentru metodele de prelucrare ce sunt vulnerabile la fluctuaţii de intensitate luminoasă în imagine.

Astfel, alegerea adecvată a spaţiului de reprezentare a culorilor constituie o etapă premergătoare importantă în analiza şi prelucrarea conţinutului imaginilor sau a secvenţelor de imagini.

4.2 Conţinutul de culoare la nivel de imagine

La nivel de imagine, informaţia de culoare poate fi analizată folosind mai multe abordări. O primă metodă, şi cea mai des întâlnită, o constituie analiza *parametrilor de nivel scăzut* extraşi din imagine. Aceştia sunt de regulă măsuri statistice numerice ale proprietăţilor distribuţiei de culoare a imaginii globale sau a anumitor zone de interes din aceasta. Cum conceptul de culoare

[12] transformata Karhunen-Loeve discretă este cunoscută şi sub numele de descompunere în componente principale sau PCA.

implică analiza senzaţiei vizuale transmise, parametrii de nivel scăzut sunt deseori insuficienţi pentru a interpreta percepţia culorilor sau a tehnicilor de culoare prezente în imagine.

Un nivel semantic superior de descriere este atins atunci când sunt asociate descrieri *textuale culorilor*, precum numele culorii. Asociind un nume fiecărei culori, permite oricărei persoane să îşi creeze o imagine vizuală a culorii în cauză. Numele culorilor sunt astfel alese încât să fie reprezentative pentru proprietăţile perceptuale cele mai semnificative ale acestora, precum nuanţa, intensitatea sau saturaţia.

La nivel de imagine, culorile nu sunt informaţii vizuale individuale, ci dimpotrivă, acestea capătă sens în relaţie cu alte culori sau regiuni de culoare vecine spaţial. Astfel, caracterizarea conţinutului de culoare implică *analiza diverselor relaţii perceptuale* ce pot apărea între culori. O modalitate, inspirată din domeniul artei, domeniu ce a furnizat primele studii referitoare la percepţia relaţiilor de culoare, constă în folosirea roţilor de culoare ("color wheels"). Acestea constituie în esenţă o reprezentare grafică a unui set de culori primare ce sunt dispuse pe un cerc într-un mod perceptual uniform.

În cele ce urmează vom face o trecere în revistă a metodelor folosite de fiecare dintre direcţiile de studiu menţionate.

4.2.1 Analiza pe bază de histogramă

Una dintre metodele cele mai eficiente de reprezentare a conţinutului de culoare o constituie *histograma*. Aceasta este o măsură statistică ce contabilizează numărul de apariţii în imagine ale fiecărei culori dintr-o anumită paletă.

Din punct de vedere matematic, histograma imaginii I este dată de relaţia:

$$h(c) = \frac{1}{M \cdot N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \delta(I(i,j) - c) \tag{4.40}$$

unde $M \times N$ reprezintă dimensiunea imaginii exprimată în pixeli, c reprezintă indicele culorii curente din paleta de culoare folosită, P_c, iar funcţia delta $\delta(x)$ este dată de ecuaţia următoare:

$$\delta(x) = \begin{cases} 1 & \text{dacă } x = 0 \\ 0 & \text{altfel} \end{cases} \tag{4.41}$$

Definită în acest fel, histograma imaginii I are sens de densitate de probabilitate discretă, valoarea $h(c)$ reprezentând probabilitatea de apariţie a culorii c în imagine.

O altă formă de histogramă utilizată frecvent este *histograma cumulată* [Stricker 95]. Aceasta este calculată pe baza histogramei $h()$ în felul următor:

$$H(c) = \sum_{i=0}^{c} h(i) \tag{4.42}$$

unde culoarea c ia valori în paleta P_c. Definită în acest fel, histograma cumulată, $H(c)$, are sens de funcţie de repartiţie discretă, reprezentând probabilitatea ca, culoarea unui pixel din imagine să fie inferioară culorii c, unde relaţiile de ordine dintre culori sunt date de ordinea de apariţie a acestora în paleta de culoare P_c.

Histograma este eficientă în primul rând datorită invarianţei totale sau parţiale a acesteia la anumite transformări geometrice ale imaginii, precum translaţii, rotaţii (până la aproximativ 45 de grade), măriri ale imaginii (până la aproximativ 1.3 ori), precum şi la schimbări de rezoluţie (în cazul în care imaginile prezintă suficiente regiuni uniforme) sau la suprapuneri parţiale de obiecte în imagine [Bimbo 99]. Pe lângă invarianţă, histograma mai are avantajul de a avea o complexitate de calcul redusă, numărul de operaţii fiind dat de numărul de pixeli din imagine.

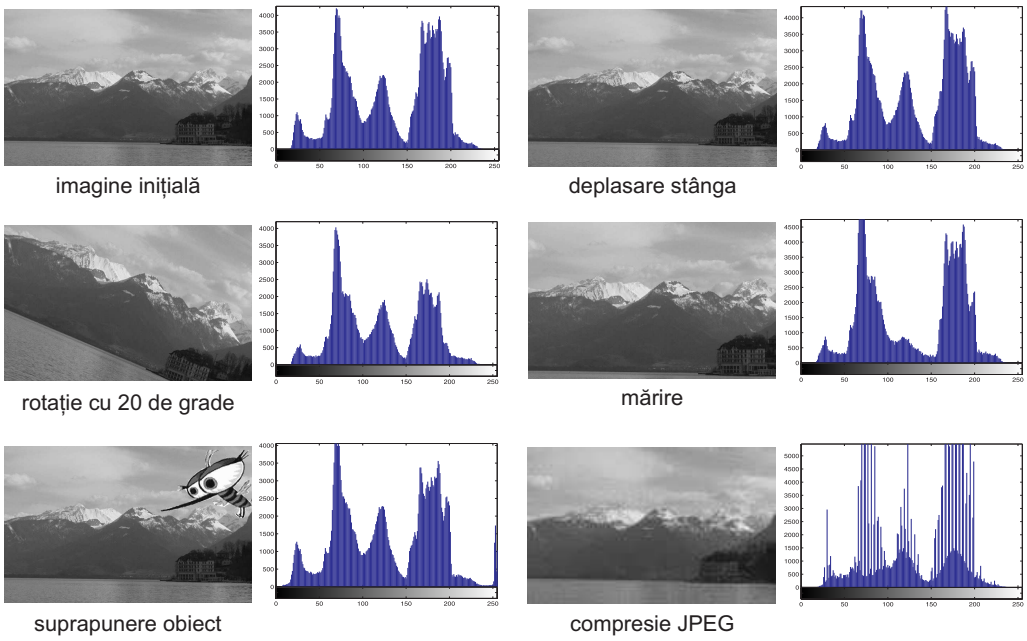


Figura 4.7: Exemple de histograme obţinute pentru diverse transformări ale imaginii: translaţie, rotaţie, mărire, suprapunere obiect şi compresie.

În Figura 4.7 am ilustrat câteva exemple de histograme obţinute în urma unor transformări ale imagini, pentru o imagine cu nivele de gri. În ciuda modificărilor, putem spune brutale, ce survin în imagine, histograma are tendinţa să păstreze distribuţia de moduri[13] a imaginii originale.

În ciuda multiplelor avantaje prezentate de histogramă, această formă de reprezentare a conţinutului de culoare din imagine prezintă şi o serie de limitări. În primul rând, histograma de culoare este foarte sensibilă la variaţiile importante de intensitate luminoasă din imagine, astfel două imagini identice din punct de vedere al conţinutului, dar "achiziţionate" în condiţii diferite de luminozitate vor furniza histograme de culoare foarte diferite. Pentru a îmbunătăţii invarianţa histogramei la fluctuaţiile de intensitate luminoasă, soluţia cea mai frecvent adoptată constă în folosirea unui alt spaţiu de culoare ce permite separarea informaţiei de luminanţă de cea cromatică [Bimbo 99], cum este cazul spaţiilor de culoare de tip YC_bC_r sau HSV (vezi Secţiunea 4.1). Astfel, histograma poate fi calculată doar pentru componentele de crominanţă, ce nu sunt afectate de variaţia intensităţii luminoase [Ionescu 08].

Un alt dezavantaj al histogramei este dat de faptul că pentru calculul acesteia nu se ţine cont de informaţia spaţială din imagine, mai exact, de poziţia spaţială a pixelilor în imagine. Astfel, două histograme identice pot corespunde în realitate cu două imagini complet diferite din punct de vedere vizual, dar care conţin aceleaşi culori şi cu o aceeaşi probabilitate de apariţie (vezi Figura 4.8).

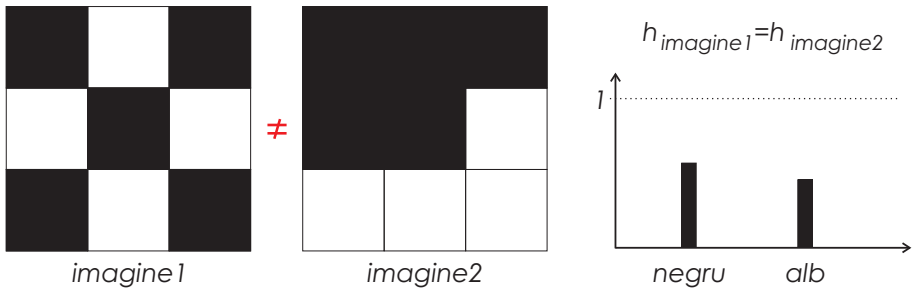


Figura 4.8: Exemplu clasic de două imagini diferite din punct de vedere vizual ce conduc la histograme absolut identice.

Acest lucru se datorează faptului că doi pixeli de aceeaşi culoare nu au

[13]un mod al histogramei este definit în general ca fiind plaja de valori din jurul unui maxim local, de regulă mărginită la stânga şi la dreapta de minime locale, ce indică prezenţa în imagine a uneia sau a mai multor regiuni semnificative cu o distribuţie de culoare similară.

obligatoriu şi aceeaşi semnificaţie vizuală, ci dimpotrivă, aceştia în realitate pot reprezenta informaţii spaţiale complet diferite, ca de exemplu pot fi pixeli de contur sau din regiuni uniforme de culoare.

Pentru a corecta acest lucru, *histogramele ponderate*, adaugă informaţii suplimentare legate de vecinătatea spaţială a pixelului curent analizat. Pe baza notaţiilor precedente, din ecuaţia 4.40, histograma ponderată este dată de ecuaţia următoare:

$$h_p(c) = \frac{1}{M \cdot N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} w(i,j) \cdot \delta(I(i,j) - c) \tag{4.43}$$

unde $w(i,j)$ reprezintă ponderea pixelului, de coordonate spaţiale în imagine (i,j), la valorea histogramei pentru culoarea c. Funcţia $w()$ este calculată pentru o anumită vecinătate a pixelului curent, iar valorile acesteia de regulă cresc odată cu gradul de neuniformitate al regiunii.

Printre funcţiile $w()$ cel mai frecvent folosite putem menţiona operatorul Laplacian, ce reprezintă derivata de ordinul doi în spaţiul bidimensional al imaginii $I()$:

$$\triangle I(i,j) = \nabla^2 I(i,j) = \frac{\partial^2 I(i,j)}{\partial i^2} + \frac{\partial^2 I(i,j)}{\partial j^2} \tag{4.44}$$

Acesta oferă informaţii referitoare la contururile prezente în imagine, conferind o pondere mai importantă pixelilor de contur, pentru care valoarea Laplacian-ului este importantă, precum şi o pondere semnificativ mai mică pixelilor din regiunile uniforme, pentru care valoarea Laplacian-ului este apropiată de zero.

O altă abordare constă în definirea funcţiei $w(i,j)$ ca fiind probabilitatea de reapariţie a culorii pixelului de coordonate spaţiale (i,j), în vecinătatea acestuia. Dacă probabilitatea este importantă, atunci pixelul face parte dintr-o regiune uniformă, şi vice-versa, dacă probabilitatea este redusă, atunci pixelul se poate afla într-o regiune de contur, poate fi un pixel izolat, etc.

O altă categorie de histograme ce ţin cont de repartiţia spaţială a pixelilor în imagine sunt *histogramele acumulative*. Pentru calculul acestora, imaginea este mai întâi divizată în mai multe regiuni iar pentru fiecare regiune este calculată o histogramă de culoare. Histograma acumulativă este obţinută prin acumularea aditivă a valorilor acestora, astfel:

$$\tilde{h}(c) = \sum_{b=0}^{N-1} f(h_b(c)) \tag{4.45}$$

unde b reprezintă indicele blocului curent, N reprezintă numărul total de blocuri din imagine iar $f()$ este de regulă o funcţie neliniară ce evidenţiază

repartiţia geometrică de culoare. Cu cât sunt acumulate mai multe culori, cu atât binii histogramei $\tilde{h}()$ sunt mai diferenţiabili. În ceea ce priveşte calitatea reprezentării conţinutului de culoare, raportat la histogramele clasice, histogramele ponderate tind să furnizeze rezultate mai precise în anumite aplicaţii, precum indexarea după conţinut a imaginilor [Ferecatu 01].

Pe lângă tipurile de histograme, putem spune, clasice, enumerate anterior, în literatura de specialitate a domeniului indexării de imagini întâlnim un caz particular de histogramă şi anume *histogramele fuzzy*. Spre deosebire de histogramele clasice în care culoarea pixelilor intervine doar la calculul valorii histogramei pentru un singur bin, în histogramele fuzzy se ia în calcul similaritatea dintre culori prin distribuirea unei anumite funcţii de apartenenţă fuzzy a fiecărui pixel către toţi binii histogramei. Astfel, histograma fuzzy a imaginii I este definită ca fiind:

$$F(I) = [f_1, f_2, ..., f_n] \tag{4.46}$$

unde

$$f_i = \sum_{j=1}^{N} \mu_{ij} \cdot P_j = \frac{1}{N} \sum_{j=1}^{N} \mu_{ij} \tag{4.47}$$

unde j reprezintă indicele pixelului curent din imagine, $j = 1, ..., N$ cu N numărul total de pixeli din imagine, P_j reprezintă probabilitatea de apariţie a pixelului j în imagine iar μ_{ij} reprezintă valoarea funcţiei de apartenenţă fuzzy a pixelului j pentru binul i din histogramă (pentru o descriere detaliată a conceptului de mulţime fuzzy vezi Capitolul 6).

Definită în acest fel, histograma fuzzy nu se limitează în a lua în calcul doar similaritatea dintre culorile din binii histogramei, ci şi disimilaritatea culorilor asociate aceluiaşi bin[14]. În domeniul indexării după conţinut a imaginilor, din punct de vedere al performanţelor, raportat la histogramele clasice, histogramele fuzzy se dovedesc a furniza rezultate mai bune cât şi o robusteţe mai ridicată la prezenţa zgomotului sau la fluctuaţiile de intensitate luminoasă din imagine [Han 02a].

4.2.2 Analiza pe baza denumirii culorilor

O altă modalitate de caracterizare a conţinutului de culoare constă în utilizarea *denumirilor culorilor*. Asocierea de denumiri textuale culorilor existente, permite oricăruia dintre noi, indiferent de cultură sau de domeniul

[14]de menţionat este faptul că în general termenul de bin al histogramei se referă nu numai la o singură culoare, cum este cazul unei palete fixe de culoare (vezi ecuaţia 4.40), ci la un interval de culoare. În acest caz, toate culorile din imagine ce sunt cuprinse în acest interval vor modifica valoarea histogramei doar pentru binul respectiv.

de activitate, să ne creăm o imagine vizuală a proprietăţilor culorii despre care este vorba. Denumirile culorilor sunt de regulă alese dintr-un dicţionar predefinit de nume de culori, iar asocierea între culoarea fizică şi numele acesteia se realizează pe baza unui sistem de denotare a culorilor ("Color Naming System"). Problema asocierii de denumiri textuale culorilor este o problemă ce a fost îndelung cercetată în literatura de specialitate. Pentru o sinteză a acesteia, cititorul se poate raporta la studiile prezentate în [Kay 03], [Benavente 04] sau [Lay 04].

În [Berlin 91] este prezentat un studiu cu privire la definirea conceptului de culoare elementară sau de bază. Astfel, numele culorilor de bază (primare) sunt definite ca fiind numele culorilor ce respectă simultan următoarele reguli:

- folosirea acestora nu este restrictivă doar pentru o anumită categorie aparte de obiecte, de exemplu, în acest sens, culoarea numită "măsliniu" ("olive") nu este o culoare de bază,
- sensul acestora nu poate fi predictibil prin înţelegerea proprietăţilor anumitor obiecte, de exemplu culoarea unei frunze nu este o culoare de bază,
- sensul acestora nu este inclus în numele unei alte culori,
- acestea au o constanţă a percepţiei, fiind percepute în acelaşi fel de persoane diferite.

Pe baza spaţiului de culoare al lui Munsell (vezi Secţiunea 4.1), [Berlin 91] merge mai departe şi defineşte ca fiind elementare, 11 culori ce sunt general valabile în cel puţin 20 de limbi diferite existente. Acestea sunt: *alb*, *negru*, *roşu*, *verde*, *galben*, *albastru*, *maro*, *roz*, *purpuriu*, *portocaliu* şi respectiv *gri*. Denumirea unui set de culori elementare stă la baza definirii descrierilor textuale pentru celelalte culori derivate existente. Dacă numele culorilor elementare sunt cvasi-similare în majoritatea culturilor existente, nu se poate spune acelaşi lucru şi pentru definirea frontierelor dintre diferitele categorii de culori, frontiere ce de regulă variază foarte mult de la o limbă la alta.

Sistemele existente de denumire a culorilor folosesc diverse tehnici pentru a obţine o anumită universalitate a denumirilor acestora. Printre metodele folosite putem enumera:

- modelizarea fuzzy a apartenenţei culorilor la anumite clase de culoare,
- denumirea culorilor în funcţie de reprezentarea fizică a acestora în termeni de intervale de lungimi de undă,

- asocierea denumirii culorilor dintr-un anumit dicţionar de culori, numit şi "Lookup Table", ce este definit "a priori" pe baza analizei manuale a diverselor spaţii de reprezentare a culorilor,

- folosirea de tehnici de "clustering" pentru definirea automată a claselor de culoare pe baza similarităţii perceptuale a acestora.

Aproape toate metodele existente de denumire a culorilor necesită pentru definirea dicţionarului de culori folosirea expertizei umane. Aceasta poate interveni, fie în totalitate, caz în care descrierile textuale sunt asociate manual culorilor, fie doar parţial prin definirea limitelor sau a anumitor parametri ce caracterizează fiecare categorie de culoare. De menţionat este faptul că până în prezent nu s-a dezvoltat o metodă complet automată de denumire a culorilor dintr-un anumit spaţiu de culoare [Benavente 04].

În marea majoritate a metodelor de analiză a conţinutului de culoare, se preferă folosirea dicţionarelor de nume de culori datorită faptului că în acest fel se elimină complexitatea de calcul specifică unei metode de denumire cvasi-automată, numele culorilor fiind disponibile direct din dicţionarul de culoare. Ca exemple de astfel de dicţionare de nume de culori putem menţiona:

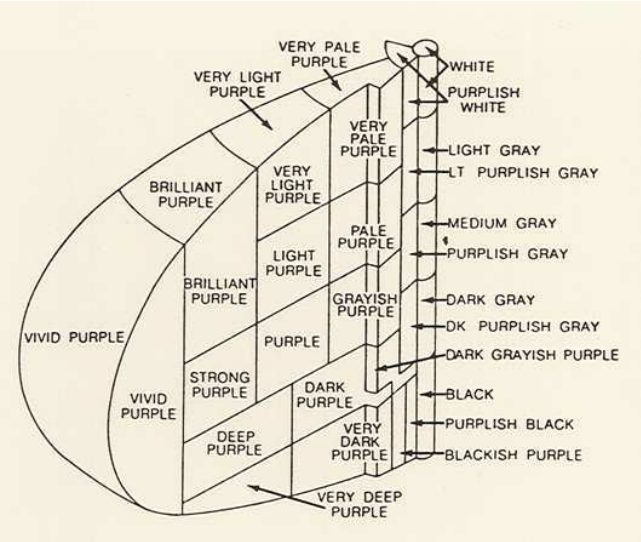


Figura 4.9: Sistemul standard de culori ISCC-NBS: exemplu de definire a culorilor purpurii prin partiţionarea spaţiului de culoare al lui Munsell.

- sistemul standard de culoare ISCC-NBS [Kelly 76] ("Inter Society Color Council" - "National Bureau of Standards") ce este definit pe baza sferei de culoare a lui Munsell (vezi Figura 4.9),

- dicţionarele de culoare "X11 Window System Distribution", "Netscape Color Names", "HTML-4 Color Names", "Two4U's Big Color Database", "Resene Paint Colours", "WebSafe" sau "CNS Color-Naming System" [CSAIL 06].

Datorită simplităţii şi a modului în care au fost alese culorile în dicţionarul de culori disponibil, un interes aparte în descrierea conţinutului de culoare îl prezintă paleta de culoare cunoscută sub numele de "WebSafe" sau "Webmaster" [Visibone 06] (vezi Figura 4.10).

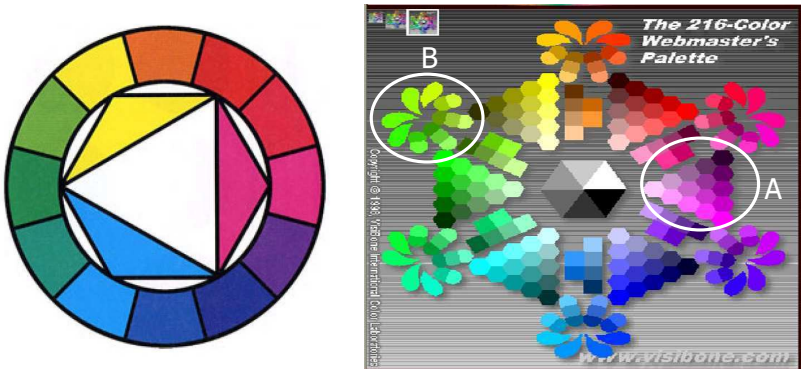


Figura 4.10: Analogia între roata de culoare a lui Itten (stânga, vezi Secţiunea 4.2.3) şi paleta de culoare "Webmaster" [Visibone 06] (dreapta, zona de tip A conţine variaţii ale unei culori elementare iar zona de tip B conţine amestecuri de culori elementare).

Aceasta prezintă o serie de avantaje ce constituie un atu pentru înţelegerea la nivel perceptual a conţinutului de culoare din imagine [Ionescu 08], şi anume:

- prezintă un bun compromis între numărul de culori disponibile (216) şi "bogăţia" de culoare furnizată: conţine 12 culori de bază şi anume: *portocaliu*, *roşu*, *roz*, *magenta*, *violet*, *albastru*, *"azure"*[15], *turcoaz*, *"teal"*[16], *verde*, *"spring"*[17] şi *galben* precum şi 6 niveluri de gri ce includ albul şi negrul (vezi Figura 4.10 unde culorile de bază enumerate sunt figurate în sensul acelor de ceasornic începând cu portocaliu la ora 12),

[15]culoarea "azure" este un albastru deschis, similar culorii cerului senin, ce se încadrează undeva între albastru şi turcoaz.

[16]culoarea "teal" este o combinaţie de verde şi albastru de saturaţie redusă, similară cu un turcuaz întunecat. Numele culorii vine de la denumirea unei anumite specii de raţe de apă dulce care prezintă în jurul ochilor tocmai această culoare.

[17]culoarea "spring" este o combinaţie de verde şi galben, şi reprezintă culoarea mugurilor plantelor la începutul primăverii.

- prezintă o diversitate de culoare ridicată: variaţii ale saturaţiei şi a intensităţii celor 12 culori elementare precum şi amestecuri de culoare,
- fiecare culoare din paletă este denumită în funcţie de nuanţă, saturaţie şi intensitate, facilitând astfel analiza percepţiei de culoare (un exemplu este prezentat în Tabelul 4.1),

Culoare	(R, G, B)	Hexazecimal	Denumire culoare
	(255, 255, 51)	FFFF33	*"Light Hard Yellow"*
	(204, 0, 102)	CC0066	*"Dark Hard Pink"*
	(204, 204, 204)	CCCCCC	*"Pale Gray"*

Tabelul 4.1: Exemplu de denumiri de culori din paleta "Webmaster".

- culorile sunt reprezentate în concordanţă cu roata de culoare a lui Itten (vezi Figura 4.10 şi explicaţia de la pagina 134), facilitând astfel analiza relaţiilor perceptuale dintre culori.

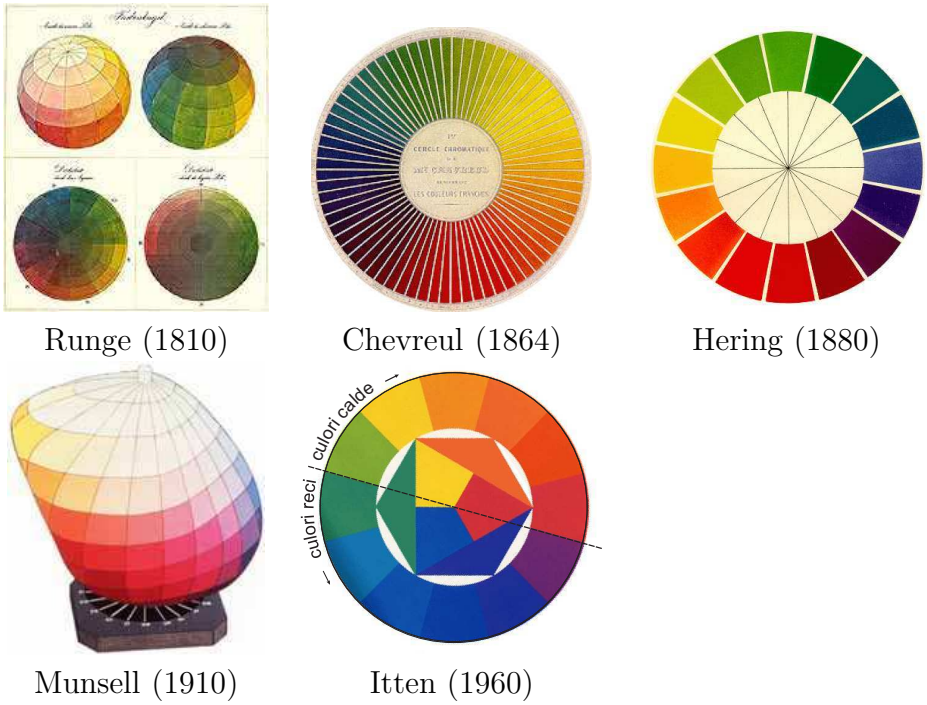
4.2.3 Analiza senzaţiei induse de culoare

O altă modalitate de caracterizare a conţinutului de culoare dintr-o imagine constă în analiza senzaţiei transmise de către acesta observatorului uman. În acest sens, o referinţă în domeniu sunt cercetările lui Itten [Itten 61], care în urma experienţei sale în calitate de pictor în curentul Bauhaus, defineşte în 1960 un set de *reguli formale* (limbaj de culoare) ce cuantificau într-un anumit număr de categorii, efectele vizuale produse la nivel perceptual de către diversele combinaţii de culori. Acestea sunt definite pe baza conceptelor folosite în domeniul artistic al picturii.

În domeniul artei, culoarea este deseori folosită pentru a descrie proprietăţile perceptuale ale diverselor obiecte din lumea înconjurătoare, precum şi ca reprezentare simbolică a culorii fizice. În acest sens, pentru a alege culorile potrivite şi pentru a crea noi amestecuri de culoare ce corespund senzaţiilor particulare ce vor fi transmise de opera lor, artiştii se folosesc de ceea ce numim "roţi de culoare" sau "color wheels" [Birren 69].

O *roată de culoare* este în esenţă un spaţiu de culoare particular în care relaţiile dintre culori sunt inspirate de teoria contrastelor şi a armoniei de culoare din domeniul picturii. Aceasta este construită pe baza unui număr arbitrar de culori elementare ce vor fi dispuse pe un cerc (roată de culoare) într-un mod perceptual liniar progresiv [Lay 04]. Definită în acest mod, o roată de culoare este o reprezentare bidimensională. Trecerea la o reprezentare tridimensională pe baza aceluiaşi principiu conduce la formarea a ceea

ce numim o *"sferă de culoare"*. De-a lungul istoriei, au fost propuse mai multe astfel de sisteme de reprezentare perceptuală a culorilor, dintre acestea menţionăm pe cele mai cunoscute ce poartă şi numele creatorilor lor, şi anume: sfera de culoare a lui Runge, roata de culoare a lui Chevreul, spaţiul culorilor opuse a lui Hering, solidul de culoare a lui Munsell, roata de culoare a lui Itten, etc. (vezi Figura 4.11).



Runge (1810) Chevreul (1864) Hering (1880)

Munsell (1910) Itten (1960)

Figura 4.11: Reprezentarea culorilor sub formă perceptuală pe baza roţilor şi a sferelor de culoare.

O astfel de reprezentare a culorilor este foarte utilă în cazul în care se doreşte studierea relaţiilor perceptuale dintre culori. În domeniul artei, conceptul de relaţie între culori, unde culorile sunt combinate pe baza unei roţi de culoare, este un studiu esenţial (vezi Josef Albers, Faber Birren, Johannes Itten, etc.).

Dacă luăm ca exemplu roata de culoare a lui Itten, care este şi una dintre cele mai cunoscute roţi de culoare, în aceasta culorile sunt aranjate cu un anumit scop, astfel: culorile considerate ca fiind calde se găsesc în prima jumătate a roţii, începând cu culoarea "spring", continuând cu galben şi finalizând cu magenta, în timp ce culorile considerate ca fiind reci se găsesc în cealaltă jumătate, pornind de la violet, continuând cu albastru şi finalizând

cu verde (vezi Figura 4.11). Mai mult, culorile ce sunt opuse din punct de vedere al percepţiei se găsesc poziţionate diametral opus (de exemplu albastru şi galben) în timp ce culorile considerate analoage sunt culori vecine pe roata de culoare (de exemplu galben şi portocaliu).

Astfel, folosind teoria de culoare dezvoltată de Itten, precum şi reprezentarea perceptuală a culorilor pe baza roţilor de culoare, putem caracteriza conţinutul vizual în termeni de *contrast* şi respectiv *concordanţă* de culoare. Itten [Itten 61] defineşte percepţia de culoare pe baza a *şapte contraste de culoare* ce sunt exemplificate în Figura 4.12, astfel:

- *contrastul de nuanţă*: acest contrast vizual este realizat prin juxtapunerea de diverse nuanţe de culoare. Cu cât acestea sunt mai diferite din punct de vedere perceptual, cu atât mai puternic este contrastul obţinut (distanţa dintre culori este evaluată folosind o roată de culoare). Un exemplu este illustrat în Figura 4.12.a.

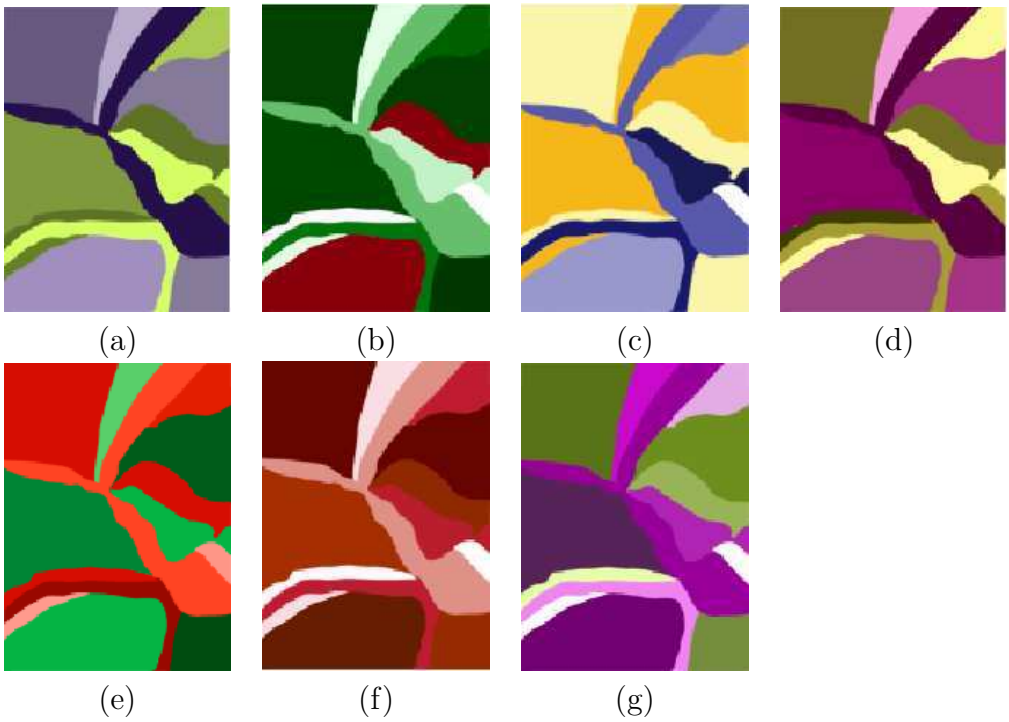


Figura 4.12: Cele şapte contraste ale lui Itten: (a) Contrastul de nuanţă, (b) Contrastul închis-deschis, (c) Contrastul cald-rece, (d) Contrastul de complementaritate, (e) Contrastul de simultaneitate, (f) Contrastul de saturaţie, (g) Contrastul de extensie (sursă imagini "http://www.worqx.com/color/itten.htm").

- *contrastul închis-deschis*: acest contrast este legat de gradul de percepţie al intensităţii luminoase. La extreme se găsesc negrul (absenţa luminii) şi albul (intensitatea maximă), iar între acestea sunt nivelele de gri şi nuanţele cromatice. Contrastul este realizat prin juxtapunerea atât a culorilor deschise cât şi închise (vezi Figura 4.12.b).

- *contrastul cald-rece*: acest contrast corespunde senzaţiei de căldură transmisă de anumite culori. În domeniul artei, culorile prezintă o anumită temperatură sau căldură. Astfel, galben, galben-portocaliu, portocaliu, roşu-portocaliu, roşu şi roşu-violet sunt considerate ca nuanţe calde, pe când galben-verde, verde, albastru-verde, albastru, albastru-violet şi violet sunt considerate ca nuanţe reci. Contrastul de culoare este realizat prin juxtapunerea atât a culorilor calde cât şi reci (vezi Figura 4.12.c).

- *contrastul de complementaritate*: acest contrast corespunde relaţiilor de complementaritate existente între culori. În practică, pe o roată de culoare (de exemplu roata de culoare a lui Itten, vezi Figura 4.11) perechile de culori complementare (opuse ca percepţie) sunt determinate de linia dreaptă ce trece prin centrul roţii şi care leagă două culori diametral opuse. Contrastul de complementaritate este astfel realizat prin folosirea de culori opuse din punct de vedere al percepţiei, obţinându-se astfel o anumită simetrie vizuală (vezi Figura 4.12.d).

- *contrastul de simultaneitate*: acest contrast se foloseşte de răspunsul asimetric al percepţiei umane la fenomenul culorilor opuse. Contrastul este realizat prin "vibrarea" percepţiei frontierelor dintre culori. Cu acest contrast se pot realiza o serie de iluzii optice interesante. Un exemplu este ilustrat în Figura 4.12.e.

- *contrastul de saturaţie*: acest contrast este realizat prin juxtapunerea de nuanţe pure saturate cu nuanţe diluate de saturaţie redusă. Acest contrast se dovedeşte însă a fi relativ, deoarece anumite culori pot apărea ca fiind mai saturate prin contrast dacă sunt alăturate unei culori mai puţin saturate, şi vice-versa. Un exemplu este ilustrat în Figura 4.12.f.

- *contrastul de extensie*: acest contrast este legat de proporţia în care sunt folosite culorile în imagine. Percepţia vizuală a unei culori este direct influenţată de gradul de luminanţă folosit, precum şi de suprafaţa spaţială ocupată de culoare în imagine. Contrastul de extensie este astfel realizat prin asocierea de culori regiunilor fizice din imagine ce au o suprafaţă proporţională cu ponderea perceptuală vizuală a culorii (vezi Figura 4.12.g).

Pe lângă cele şapte contraste definite de Itten, în domeniul artei întâlnim şi anumite configuraţii de culoare ce au efecte vizuale particulare. Aceste "*scheme de culoare*", ce exprimă relaţiile perceptuale dintre diferitele tipuri de contraste de culoare, sunt descrise deseori ca fiind modalităţile fundamentale de obţinere a armoniei de culoare în imagine [Birren 69]. Acestea sunt următoarele:

- *schema de culoare monocromatică*: implică armonizarea unei singure nuanţe de culoare şi este realizată pe baza contrastului de culoare închis-deschis. Astfel, pornind de la o anumită nuanţă de culoare se obţin mai multe variaţii ale acesteia în modul următor: nuanţa poate fi întunecată prin adăugarea de negru, se pot obţine mai multe tonalităţi ale acesteia prin adăugarea de gri sau nuanţa poate fi deschisă prin adăugarea de alb. Mai mult, pentru a se evita o posibilă monotonie a culorilor, se poate folosi şi contrastul de saturaţie prin varierea saturaţiei nuanţei folosite.

- *schema de culori adiacente*: implică armonizarea nuanţelor de culoare similare sau adiacente. Aceasta este realizată de regulă prin amestecul a maxim trei culori adiacente, ce sunt alese pe baza unei roţi de reprezentare a culorilor (ca de exemplu roata lui Itten). Diferenţa între acestea constă în faptul că una dintre cele trei culori va fi pusă în evidenţă prin folosirea sa în imagine într-o proporţie semnificativă faţă de celelalte nuanţe.

- *schema de culori complementare*: implică armonizarea echilibrului dintre culorile opuse ca percepţie sau complementare. Modalitatea cea mai simplă în care poate fi realizată este pe baza a două culori ce sunt complementare pe o roată de reprezentare a culorilor. O altă modalitate de realizare implică folosirea a două culori adiacente ce sunt contrastate de perechea de culori opuse acestora. Această alegere a culorilor corespunde unei alegeri în X pe o roată de culoare şi este cunoscută şi sub numele de *dublă complementaritate*. O *complementaritate divizată* este obţinută prin asocierea unei nuanţe de culoare cu două culori vecine nuanţei complementare, formându-se astfel o structură de tip Y.

Ca exemple de sisteme ce folosesc descrierea conţinutului de culoare la nivel perceptual pe baza conceptelor enumerate mai sus, putem menţiona sistemul QBIC al "State Hermitage Museum" [QBIC 03] şi sistemul PICASSO [Corridoni 99]. În sistemul QBIC, picturile sunt căutate într-o bază de date folosind culorile predominante, precum şi relaţiile existente între regiunile de culoare din pictură. Acestea din urmă sunt specificate de utilizator prin

desenarea interactivă a unei anumite scheme de culoare. Un alt exemplu este sistemul SoloArt propus în [Lay 04] unde conceptele artistice de culoare sunt extrase tot pentru indexarea după conţinut a picturilor. Relaţiile dintre culori sunt analizate de această dată într-un spaţiu perceptual, şi anume spaţiul de culoare LCH (vezi Secţiunea 4.1), pe baza căruia sunt determinate în mod automat tehnicile de culoare folosite: contrastele de culoare, cele şapte contraste ale lui Itten precum şi schemele de armonie a culorilor.

4.3 Conţinutul de culoare în secvenţele de imagini

După cum am menţionat deja în partea introductivă a acestui capitol, majoritatea tehnicilor de caracterizare a conţinutului de culoare se limitează în mod natural la analiza acestuia doar la nivel de imagine, fără a lua în consideraţie dimensiunea temporală. Astfel, majoritatea metodelor existente sunt destinate analizei imaginilor statice. Totuşi, putem menţiona o serie de abordări bazate mai mult sau mai puţin integral pe informaţia de culoare, abordări ce încearcă să descrie conţinutul temporal de culoare din secvenţele de imagini în diverse domenii de aplicaţie. Dintre acestea, ne vom limita la detalierea metodelor propuse în [Colombo 99] (secvenţe publicitare), [Detyniecki 03] (secvenţe de ştiri) şi [Ionescu 08] (secvenţe de animaţie).

Sistemul propus în [Colombo 99] foloseşte informaţia de culoare în relaţie cu alţi parametri, specifici secvenţelor de imagini, pentru a descrie din punct de vedere semantic conţinutul secvenţelor publicitare. Acesta propune două niveluri perceptuale de analiză, şi anume: "nivelul expresiv" şi "nivelul emoţional".

La *nivel expresiv*, secvenţele publicitare sunt clasificate în patru categorii perceptuale, şi anume: "tipul practic", "animat", "utopic" şi "critic". Clasificarea automată a acestora este realizată pe baza a mai mulţi parametrii de nivel scăzut, precum prezenţa tranziţiilor de tip "cut" şi "dissolve" (vezi Capitolul 2), apariţia şi dispariţia repetitivă a anumitor culori din imagine, prezenţa liniilor verticale şi orizontale în imagine precum şi prezenţa culorilor saturate sau nesaturate. La *nivel emoţional*, secvenţele publicitare sunt caracterizate din punct de vedere al acţiunii, suspansului, senzaţiei de calm, senzaţiei de relaxare, de veselie precum şi al entuziasmului transmis. Informaţia de culoare este utilizată în acest caz pentru a atinge anumite niveluri emoţionale. De exemplu, nivelul de acţiune al unei secvenţe de publicitate poate fi crescut prin folosirea culorilor roşu şi purpuriu. În mod similar, senzaţia de calm poate fi amplificată prin folosirea culorilor: albas-

tru, portocaliu, verde sau alb, sau poate fi diminuată prin prezenţa culorilor purpurii şi a negrului.

Fiecare categorie de descrieri semantice propusă este caracterizată de o anumită configuraţie a valorilor unui set de parametri de nivel scăzut, ce vor fi asociaţi secvenţelor publicitare pe baza reprezentării fuzzy. De exemplu, secvenţele publicitare din categoria "practică" sunt caracterizate de absenţa culorilor saturate, $\phi_{saturated} = 0$, unde ϕ reprezintă funcţia fuzzy de apartenenţă la simbolul ce modelizează conceptul de prezenţă a culorilor saturate în secvenţă (vezi Capitolul 6); de prezenţa liniilor orizontale şi verticale, $\phi_{hor/vert} = 1$, precum şi de prezenţa tranziţiilor de tip "dissolve", $\phi_{dissolves} = 1$.

Sistemul propus în [Detyniecki 03] foloseşte arbori de decizie fuzzy pentru a extrage informaţii din secvenţele de ştiri, procedeu cunoscut şi sub numele de "data mining"[18]. Arborii de decizie sunt folosiţi în acest caz la extragerea în mod automat a unui set de reguli ce vor fi folosite pentru clasificarea conţinutului anumitor segmente tematice din secvenţele de ştiri. Informaţia folosită în acest caz este exclusiv informaţia de culoare.

În primă fază de analiză, fiecare plan video al secvenţei este rezumat cu un anumit număr de imagini reprezentative pentru conţinutul acestuia (imagini cheie). Culorile fiecărei imagini reţinute sunt proiectate pe o paletă de culoare redusă (64 sau 256 culori) prin cuantificarea liniară a spaţiului RGB. Conţinutul de culoare la nivel de imagine este reprezentat mai departe folosind histograme color ce formează vectori de caracteristici de culoare în spaţiul determinat de paleta de culoare folosită. Mai departe, vectorii de caracteristici astfel obţinuti sunt folosiţi pentru a construi arborele de decizie fuzzy pe baza căruia se va realiza detectarea a două evenimente semantice din secvenţele de ştiri, şi anume: prezenţa de text încrustat în imagine ("inlays") şi prezenţa prezentatorului sau a jurnalistului.

Un sistem particular de analiză a conţinutului de culoare, ce ia în calcul dimensiunea temporală, este propus în [Ionescu 08]. Acesta foloseşte exclusiv informaţia de culoare pentru a detecta şi analiza tehnicile de culoare prezente în filmele artistice de animaţie. Metoda propusă se bazează pe ipoteza că aproape fiecare film de animaţie conţine o paletă de culoare particulară, paletă ce a fost aleasă în mod intenţionat de autor în momentul conceperii filmului pentru a transmite anumite sentimente sau pentru a exprima anumite concepte artistice.

Sistemul propus reduce în primă fază redundanţa temporală a secvenţei

[18] "data mining" reprezintă procesul de căutare în conţinutul unor colecţii foarte vaste de date, de regulă după criterii semantice, a anumitor informaţii relevante sau de noi "cunoaşteri" ("knowledge") în cazul în care natura conţinutul datelor precum şi relaţiile dintre acestea nu sunt cunoscute.

prin reţinerea a doar un procent, $p\%$, din imaginile fiecărui plan video. Mai departe, folosind un algoritm de reducere de culoare de tip "dithering"[19], metodă ce furnizează o fidelitate vizuală ridicată a imaginii rezultate, conţinutul temporal de culoare al secvenţei este reprezentat prin calcularea histogramei globale ponderate, h_{GW}, propusă în [Ionescu 05c]. Aceasta este dată de relaţia:

$$h_{GW}(c) = \sum_{i=0}^{M} \left[\frac{1}{N_i} \sum_{j=0}^{N_i} h^j_{shot_i}(c) \right] \cdot w_i \tag{4.48}$$

unde M reprezintă numărul total de plane video al secvenţei, N_i reprezintă numărul total de imagini reţinute pentru planul de indice i (reprezentând un procent $p\%$ din numărul total de imagini ale acestuia), $h^j_{shot_i}()$ reprezintă histograma color a imagini de indice j din planul video i, c reprezintă o culoare din paleta de culori folosită iar w_i reprezintă ponderea planului video i la calculul histogramei ponderate şi este dată de relaţia următoare:

$$w_i = \frac{N_{shot_i}}{N_{total}} \tag{4.49}$$

unde N_{shot_i} reprezintă numărul total de imagini ale planului video i iar N_{total} reprezintă numărul total cumulat de imagini ale tuturor planelor video din secvenţă. Astfel, cu cât planul curent analizat este mai lung, cu atât contribuţia acestuia la distribuţia globală de culoare a secvenţei este mai importantă.

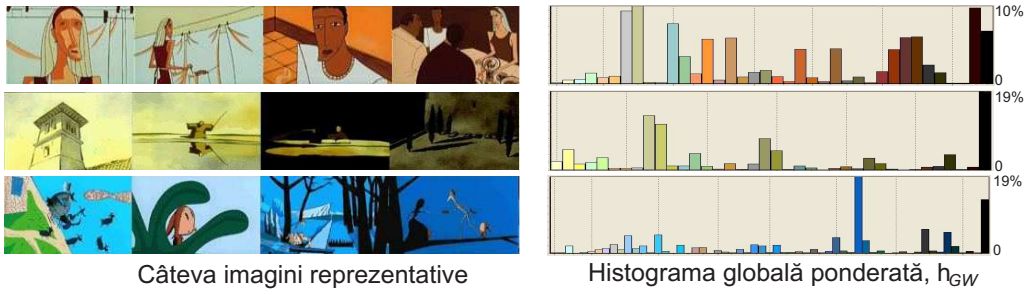


Figura 4.13: Exemple de histograme globale ponderate obţinute pentru $p = 15\%$ (sursă imagini [Folimage 06]).

[19] procesul de "dithering" constă în adăugarea intenţionată a unei anumite forme de zgomot necorelat pentru a distribui aleator eroarea de cuantizare în cazul procesului de cuantificare a valorilor unei funcţii. În cazul reducerii de culoare, eroarea obţinută prin substituirea culorilor este distribuită progresiv în vecinătatea acestora, evitând astfel formarea de artefacte în imagine.

Definită în acest fel, histograma ponderată redă procentul de apariţie temporală în secvenţă a fiecărei culori, fiind o măsură cantitativă a distribuţiei de culoare globală a secvenţei (vezi un exemplu în Figura 4.13).

Mai departe, conţinutul de culoare este caracterizat la nivel sintactic cu o serie de parametri statistici de nivel scăzut, precum procentul de culori închise şi deschise, procentul de culori calde şi reci, diversitatea de culoare, procentul de culori adiacente, etc., parametrii ce sunt extraşi pe baza histogramei ponderate folosind dicţionarului de nume de culori furnizat de paleta "Webmaster" (vezi Figura 4.10). Nivelul semantic de descriere este obţinut pe baza unui set de reguli de reprezentare fuzzy a acestor parametrii. Astfel, secvenţelor le sunt asociate descrieri simbolice, exprimate textual în felul următor: "culorile predominante sunt calde" sau "secvenţa prezintă un contrast de saturaţie", descrieri al căror grad de veridicitate este modelizat de gradul de apartenenţă fuzzy.

4.4 Concluzii

În acest capitol am discutat diversele modalităţi de analiză şi caracterizare a conţinutului de culoare, atât la nivel spaţial, în imagine, cât şi la nivel temporal, în secvenţe de imagini.

Un rol important în analiza de culoare îl are spaţiul de reprezentare al culorilor folosit. Acesta este de regulă ales în aşa fel încât să pună în evidenţă anumite proprietăţi de culoare ce vor facilita implementarea algoritmului de prelucrare dorit. Pornind de la spaţii de culoare, putem spune clasice, precum spaţiul RGB folosit în aproape toate dispozitivele hardware de reprezentare a culorilor (plăci grafice, ecrane, imprimante, etc.), spaţiile de culoare evoluează spre spaţii perceptuale sau inspirate de modalitea fizică de percepţie a culorii din sistemul vizual uman, precum spaţiul LCH. Dezvoltarea acestora este motivată în principal de necesitatea de caracterizare a informaţiei de culoare la un nivel perceptual, apropiat de cel uman, ţinta finală fiind atingerea unui nivel semantic de descriere a conţinutului de culoare din imagine, direcţie de studiu de foarte mare actualitate în acest moment.

Pentru a înţelege automat conţinutul de culoare, în absenţa unei analize vizuale a acestuia, o modalitate foarte eficientă constă în folosirea numelor atribuite culorilor. Prin asocierea unei descrieri textuale fiecărei culori, ne putem face o imagine mentală asupra proprietăţilor fizice ale acesteia. Problema care apare, având în vederea diversitatea putem spune infinită de culoare, este modalitatea de denumire a culorilor existente în natură. Soluţia adoptată este realizarea pe baza expertizei manuale a anumitor spaţii de culoare a unor dicţionare de nume de culori. Acestea, în funcţie de aplicaţie, pot

conţine o paletă mai vastă sau mai redusă de culoare precum şi un nivel de detaliu al descrierii textuale variabil. Dicţionarele de culoare sunt preferate metodelor de denumire automată a culorilor, pe de-o parte datorită preciziei mult mai bune a acestora (fiind constituite de operatori umani) dar şi datorită faptului că nu necesită operaţii de calcul, numele culorilor fiind disponibile în mod direct prin accesarea dicţionarului.

Cunoaşterea percepţiei conţinutului de culoare la nivel de pixel nu este suficientă pentru înţelegerea acestuia. În transmiterea informaţiei vizuale, datorită modului în care este construit sistemul vizual uman, un rol definitoriu îl au relaţiile existente între diversele regiuni de culoare din imagine. De exemplu, percepţia unei regiuni dintr-o imagine în care culorile pixelilor au o distribuţie similară unei table de şah, unde în loc de alb se află de exemplu o culoare arbitrară deschisă (efect ce poate fi obţinut în urma unei reducerii de culoare de tip "dithering"), este de regiune întunecată, în ciuda faptului că analizate individual, procentul culorilor deschise este egal cu cel al culorilor închise (negru în acest caz). Acest lucru se datorează faptului că la nivel vizual culorile sunt interpretate în corelaţie cu regiunile vecine de culoare. O modalitate eficientă de analiză a relaţiei de culoare este pe baza metodelor folosite în domeniul artei, domeniu ce a dat naştere primelor studii de percepţie de culoare. În artă, relaţiile dintre culori sunt analizate fie prin proiecţia acestora pe roţi de reprezentare a culorilor, fie folosind teoria contrastelor şi a armoniei de culoare.

Ca tendinţă generală a sistemelor existente, putem menţiona că informaţia de culoare este folosită cu predilecţie pentru caracterizarea conţinutului static al imaginilor fixe, cum este cazul indexării după conţinut a acestora. În domeniul indexării după conţinut a secvenţelor de imagini, metodele bazate exclusiv pe culoare sunt foarte puţine. Acest lucru se datorează în principal faptului că informaţia de culoare, analizată individual, nu are suficientă putere discriminatorie pentru a oferi o caracterizare globală a conţinutului. Din această cauză, în marea majoritate a metodelor de analiză existente, culoarea este folosită alături de alte informaţii definitorii pentru secvenţele de imagini, precum mişcarea şi structura temporală.

CAPITOLUL 5

Rezumarea automată de conţinut

Rezumat: *Accesarea informaţiei utile dintr-o colecţie vastă de secvenţe de imagini este de regulă o operaţie dificilă ce necesită un timp important. Acest lucru se datorează în principal volumului foarte mare de date ce trebuiesc vizualizate. O soluţie constă în reducerea redundanţei informaţionale prin generarea automată de reprezentări compacte de conţinut, numite şi rezumate. În funcţie de rezultatul obţinut, există două categorii distincte de rezumat: rezumatele statice ce furnizează o colecţie de imagini reprezentative ale secvenţei şi rezumatele dinamice ce furnizează o colecţie de pasaje ale secvenţei ce sunt asamblate tot sub forma unei secvenţe. În acest capitol vom discuta diversele metode şi tehnici de generare automată a rezumatelor din cele două categorii, precum şi aplicaţiile acestora. De asemenea, vom discuta un punct critic al metodelor de rezumare automată ce îl constituie problematica evaluării calităţii şi pertinenţei acestora.*

În general, bazele sau colecţiile de secvenţe de imagini conţin la ordinul a mii de secvenţe. Cum un sigur minut video, la o cadenţă de 25 de imagini pe secundă (standardul PAL[1]), este echivalentul a 1500 de imagini statice, putem doar să ne imaginăm numărul gigantesc de imagini conţinute într-o astfel de colecţie de date. Pentru a avea acces la informaţie, utilizatorul

[1]pentru o descriere a standardelor de televiziune vezi explicaţiile de la pagina 116.

în cele mai multe cazuri, are nevoie să vizualizeze conţinutul secvenţelor. Problema nu este una dificilă în cazul în care ar fi vorba de doar câteva secvenţe, dar vizualizarea a mii de secvenţe este un lucru aproape imposibil de realizat. Una dintre soluţiile adoptate constă în folosirea *rezumatelor de conţinut*.

Un rezumat al unei secvenţe de imagini poate fi definit în linii mari ca fiind *o reprezentare compactă a conţinutului acesteia* [Li 01]. Mai riguros, rezumatul unei secvenţe de imagini reprezintă *o colecţie, de dimensiuni reduse, de imagini fixe (colecţie de imagini) sau în mişcare (colecţie de segmente), ce redă conţinutul secvenţei în aşa fel încât partea esenţială a acestuia să fie conservată iar utilizatorul să fie informat rapid şi concis* [Pfeiffer 96].

Interesul în a dispune de o reprezentare compactă a secvenţei nu se rezumă doar la reducerea timpului necesar căutării şi navigării în conţinutul bazei de date. Rezumatul de conţinut poate fi folosit şi pentru a reduce timpul de calcul în anumite metode de analiză şi prelucrare a secvenţelor de imagini, prin reducerea volumului de date ce trebuiesc prelucrate. Mare majoritate a metodelor existente nu folosesc integral conţinutul secvenţei, ci numai informaţia furnizată de un anumit număr sau grup de imagini reprezentative ("imagini cheie"). Acestea sunt alese astfel încât, pentru prelucrarea vizată, informaţia necesară din secvenţă să nu fie alterată. De exemplu, pentru a calcula şi analiza distribuţia globală de culoare a unei secvenţe de imagini, folosirea unei singure imagini pentru fiecare plan video oferă o precizie similară folosirii tuturor imaginilor din secvenţă [Ionescu 05c].

Din punct de vedere al procesului de generare, un rezumat poate fi construit *manual*, *semi-automat* (intervenţia umană este parţială, de regulă folosită ca validare) sau în mod *automat*. Având în vedere volumul mare de date conţinute chiar la nivelul unei singure secvenţe, metodele de generare manuală, şi chiar semi-automată, sunt evitate datorită implicării unui număr important de resurse umane în procesul de selecţie a conţinutului reprezentativ al secvenţei. Tendinţa actulă este de automatizare completă a procesului de rezumare pentru a putea fi astfel folosit în timp real la indexarea conţinutului bazelor de secvenţe de imagini.

După cum reiese din însăşi definiţia conceptului de rezumat, este posibilă generarea a două categorii distincte de rezumat, şi anume:

- pe de-o parte sunt **rezumatele în imagini** sau *rezumatele statice*: acestea reprezintă un fel de "storyboard"[2] simplificat al secvenţei şi sunt

[2]un "storyboard" reprezintă o modalitate de organizare grafică sub forma unei serii de ilustraţii sau imagini, prezentate similar unei benzi desenate ("comics"), a momentelor importante din conţinutului unui document video, film, etc. Acesta are ca scop previzualizarea conţinutului şi precede procesul de creare propriu-zisă.

la bază o colecţie de imagini reprezentative pentru conţinutul secvenţei. În literatura de specialitate acestea sunt cunoscute sub numele de "video summaries".

- pe de altă parte sunt **rezumatele în mişcare** sau *rezumatele dinamice*: acestea reprezintă o colecţie de segmente ale secvenţei, fiind ele însele nişte secvenţe de imagini, dar de o durată mult inferioară secvenţei iniţiale. Uzual, dacă este vorba de rezumarea unui document video, în rezumatul dinamic este prezentă şi informaţia audio. Rezumatele dinamice sunt cunoscute în literatura de specialitate sub numele de "video skims".

Aceste două modalităţi de rezumare a conţinutului unei secvenţe de imagini prezintă fiecare o serie de avantaje şi dezavantaje. Astfel, în ceea ce priveşte rezumatele în imagini, principalele avantaje ale acestora pot fi sintetizate cu următoarele:

- pot fi *generate rapid* deoarece nu iau în calcul decât informaţia vizuală (sunetul şi textul nu sunt prezente),
- pot fi *vizualizate foarte uşor*, fiind doar un ansamblu de imagini ce nu necesită sincronizarea sau temporizarea datelor (de exemplu, sincronizarea sunetului şi a imaginii),
- pot facilita construirea *imaginilor de tip "mozaic"*[3] [Aner 01],
- pot fi uşor de imprimat pe un suport fizic pentru a ţine loc, de exemplu, de *"storyboard"* al secvenţei,
- permit *reducerea complexităţii de calcul* pentru anumite metode de analiză ce pot fi aplicate direct acestora.

Pe de altă parte, rezumatele dinamice prezintă şi ele o serie de avantaje fundamentale, astfel:

- acestea *au mai mult sens* decât rezumatele statice deoarece conţin informaţie temporală de mişcare, informaţie ce este pierdută în rezumatele statice,
- un rezumat dinamic este *mai bogat în informaţie*, acesta putând conţine şi alte informaţii precum sunetul,

[3] vezi exlicaţia de la pagina 150.

- vizualizarea rezumatelor dinamice este mult mai *naturală şi atractivă*: este mult mai interesant pentru utilizator să vizualizeze, de exemplu, reclama unui nou film în curs de apariţie ("trailer"), decât să vizualizeze o succesiune sacadată de imagini statice din acesta [Li 01].

În ciuda faptului că rezumatele dinamice oferă un conţinut informaţional mult mai bogat decât rezumatele statice, acestea implică în cele mai multe situaţii o complexitate de calcul mult mai importantă, precum şi un proces de genereare mult mai laborios (sincronizare imagine şi sunet, respectarea continuităţii şi a coerenţei vizuale, etc.).

În acest punct, analizând avantajele şi dezavantajele celor două tipuri de rezumate, am fi tentaţi să alegem doar unul dintre ele ca fiind soluţia optimală a problemei de rezumare de conţinut. În practică, *ambele tipuri de rezumate sunt necesare* într-o aplicaţie de indexare după conţinut, deoarece fiecare dintre acestea este adaptat unei cerinţe diferite. Astfel, rezumatul în imagini permite o reprezentare rapidă şi concisă, în doar câteva imagini, a conţinutului vizual, ideală în cazul în care utilizatorul doreşte doar să "răsfoiască" conţinutul bazei de date, în timp ce rezumatul dinamic permite o reprezentare rapidă şi concisă a conţinutului dinamic al secvenţei, permiţând utilizatorului ca într-un timp relativ scurt să-şi facă o idee asupra acţiunii secvenţei.

Mai mult, cu toate că nu este cea mai bună strategie, fiecare tip de rezumat poate fi generat pornind de la celălalt. De exemplu, un rezumat dinamic poate fi construit pe baza unui rezumat static prin concatenarea unor segmente de o anumită durată ce conţin imaginile rezumatului static. Similar, un rezumat static poate fi generat dintr-un rezumat dinamic prin păstrarea doar a anumitor imagini reprezentative sau chiar prin sub-eşantionarea temporală a acestuia.

Pentru un studiu bibliografic complet al literaturii de specialitate din această direcţie de studiu, cititorul se poate raporta la lucrările [Li 01] sau [Truong 07]. În cele ce urmează vom face o trecere în revistă a particularităţilor metodelor folosite de fiecare dintre cele două categorii de rezumat.

5.1 Construcţia rezumatelor statice

După cum am menţionat în partea introductivă a acestui capitol, un rezumat static este la bază o colecţie de imagini fixe ce sunt considerate ca fiind reprezentative pentru conţinutul secvenţei. Acestea sunt numite şi "imagini cheie". Din punct de vedere formal, rezumatul static al secvenţei S, notat

$R_{img}(S)$, poate fi definit în felul următor:

$$R_{img}(S) = \{image_1, image_2, ..., image_N\} \tag{5.1}$$

unde $image_i$ reprezintă imaginea cheie de indice i, cu $i = 1, ..., N$ iar N reprezintă numărul total de imagini din rezumat sau dimensiunea rezumatului.

Valoarea parametrului N are un rol important asupra calităţii rezumatului. Dacă N este cunoscut *"a priori"*, atunci dimensiunea rezumatului va fi folosită drept constrângere iniţială a algoritmului de extragere a imaginilor cheie. În acest caz, dimensiunea rezumatului va fi aceeaşi pentru toate secvenţele analizate. Pe de altă parte, dacă N nu este fixat iniţial, atunci valoarea acestuia va fi determinată automat de algoritmul de calcul. În acest caz dimensiunea rezumatului va fi adaptată la conţinutul fiecărei secvenţe în parte (de exemplu, un conţinut bogat în acţiune va fi reprezentat cu mai multe imagini decât un conţinut static).

5.1.1 Clasificarea metodelor existente

Din punct de vedere al modului în care sunt extrase imaginile cheie, metodele existente de calcul al rezumatelor statice pot fi clasificate în următoarele categorii [Li 01]:

- *extragerea imaginilor prin eşantionare*,
- *extragerea imaginilor la nivel de plan*,
- *extragerea imaginilor la nivel de segment video*,
- *alte abordări*.

Extragerea imaginilor cheie prin eşantionare

Extragerea prin eşantionare constă în selectarea imaginilor cheie, fie în mod aleator, fie ca fiind uniform distribuite temporal în secvenţă [Taniguchi 95]. Avantajul unei astfel de abordări constă în primul rând în simplitatea şi rapiditatea acesteia, complexitatea de calcul fiind în acest caz total neglijabilă.

Totuşi, rezumatul obţinut nu este neapărat reprezentativ pentru conţinutul secvenţei deoarece acesta nu este luat în calcul la momentul generării. De exemplu, folosind această strategie este foarte probabil ca anumite plane video de scurtă durată, dar reprezentative pentru acţiunea secvenţei, să nu fie reprezentate în rezumat, în timp ce anumite plane de lungă durată, cu un conţinut informaţional redundant, să fie reprezentate în rezumat cu mai multe imagini cheie similare.

Extragerea imaginilor cheie la nivel de plan

O metodă mai elaborată constă în extragerea imaginilor cheie la nivelul planelor video.

Într-o variantă simplificată, selectarea imaginilor cheie se poate realiza prin alegerea în mod arbitrar a unei singure imagini pentru fiecare plan video, de exemplu: prima imagine, imaginea de mijloc, o imagine aleatoare, ultima imagine, etc. Această strategie dă rezultate vizuale satisfăcătoare în cazul în care conţinutul secvenţei prezintă o variabilitate medie sau redusă, oferind un bun compromis între calitate şi complexitatea de calcul. Pe de altă parte, complexitatea de calcul în acest caz este dată în principal doar de segmentarea în plane. Luând în calcul faptul că de regulă orice metodă de analiză a conţinutului video necesită segmentarea temporală a secvenţei, se poate presupune că în momentul calculării rezumatului, aceasta este deja disponibilă. În aceste condiţii, putem considera complexitatea de calcul a rezumatului ca fiind neglijabilă.

Totuşi, în cazul unui conţinut bogat în acţiune, o singură imagine nu este întotdeauna suficientă pentru a reprezenta conţinutul vizual la nivel de plan, existând riscul ca aceasta să fie o imagine de tranziţie a unui efect vizual sau a unei mişcării rapide. Faptul că imaginile cheie sunt alese la nivel de plan video, va asigura totuşi că acestea nu sunt imagini de tranziţie ale tranziţiilor video prezente în secvenţă (de exemplu: "fades", "dissolves", etc., vezi Secţiunea 2.1).

Pentru a ilustra cele menţionate anterior, în Figura 5.1 am comparat rezultatele obţinute cu strategiile simplificate de extragere a imaginilor cheie la nivel de plan cu o metodă de rezumare ce ia în calcul conţinutul vizual al planului, şi anume alegerea imaginii mediane[4] ca imagine cheie. Se poate observa că strategiile de alegere a imaginilor cheie ce nu iau în calcul analiza conţinutul planului furnizează rezultate satisfăcătoare pentru un conţinut cvasi-constant al planului (vezi primul plan din Figura 5.1), în timp ce pentru un conţinut complex (vezi al doilea plan din Figura 5.1) imaginea mediană este mai pertinentă pentru acţiunea globală a planului. De notat este faptul că anumite studii au demonstrat că din punct de vedere statistic, uneori o strategie de alegere aleatoare poate furniza rezultate similare sau chiar superioare unei strategii adaptate conţinutului. Această ipoteză are sens în virtutea faptului că prin definiţie un plan video prezintă o omogenitate a conţinutului (temporală, spaţială şi de acţiune), astfel că teoretic oricare dintre imaginile acestuia este reprezentativă pentru conţinut.

O altă metodă, ce ţine cont de această dată de conţinutul vizual al

[4]imaginea mediană a unui anumit plan este definită ca fiind imaginea cea mai apropiată, în sensul unei măsuri de distanţă, de ansamblul celorlalte imagini din plan [Chanussot 98].

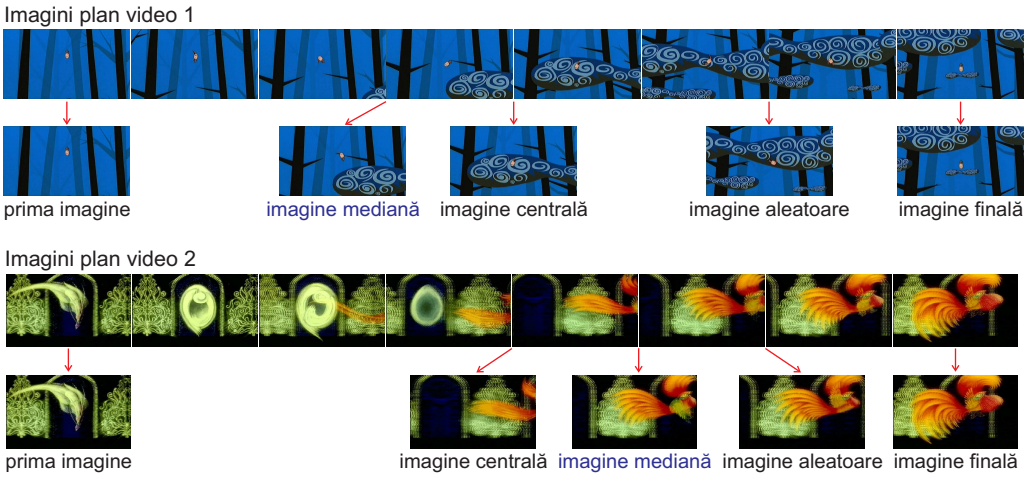


Figura 5.1: Exemplu de rezumare cu o singură imagine pentru o mişcare de obiecte (primul plan), precum şi pentru o schimbare complexă a scenei (al doilea plan, axa orizontală este axa temporală, sursă imagini [Folimage 06]).

planelor video, constă în extragerea imaginilor cheie pe baza *analizei statistice a distribuţiei de culoare*. Această strategie a fost motivată în principal de eficienţa şi robusteţea histogramelor de culoare în reprezentarea globală a conţinutului de culoare la nivel de imagine (vezi Secţiunea 4.2.1). Un exemplu este metoda propusă în [Zhang 97] în care diferenţele dintre histogramele color ale imaginilor succesive sunt filtrate cu un anumit prag pentru a extrage imaginile cele mai diferite din punct de vedere vizual. O altă abordare este propusă în [Zhuang 98] unde imaginile cheie sunt extrase folosind un algoritm de clasificare nesupervizata în care măsura de disimilaritate dintre imagini este dată de distanţa dintre histogramele de culoare.

O abordare interesantă inspirată de modalitatea de calcul a medianului vectorial este propusă în [Ott 07]. Metoda propusă rezumă conţinutul fiecărui plan video cu un număr de imagini ce este adaptat variabilităţii conţinutului vizual al acestuia. Alegerea acestora este realizată în funcţie de numărul de moduri[5] ale unei histograme de distanţe cumulate, unde distanţa cumulată a unei imagini este definită ca fiind suma distanţelor dintre histograma acesteia şi histogramele tuturor celorlalte imagini ale planului.

O limitare a metodelor din această categorie este dată de faptul că nu iau în calcul conţinutul de mişcare. Astfel, imaginile selectate, chiar dacă sunt diferite între ele, este posibil să aparţină unei aceleaşi mişcări, ca de exemplu a unei mişcări globale a camerei video. Mai mult, aceste metode sunt

[5]vezi explicaţia de la pagina 127.

dependente de alegerea unui anumit număr de praguri folosite la definirea similarităţii dintre imagini.

Abordările bazate pe *analiza mişcării* sunt concepute pentru a adapta numărul de imagini cheie extrase la dinamica temporală a scenei. Un exemplu este metoda propusă în [Wolf 96] unde imaginile cheie sunt extrase ca fiind minime locale ale unei anumite funcţii temporale ce măsoară activitatea de mişcare a secvenţei.

Acest gen de abordare prezintă şi el o serie de dezavantaje. În primul rând este dificil să determinăm importanţa unui segment al secvenţei folosind doar criterii de mişcare. Mai mult, ca şi în cazul precendent, este posibil cu această strategie să obţinem imagini de tranziţie, de regulă înceţoşate, specifice unei mişcări foarte rapide.

Un alt tip de abordare, care se foloseşte de asemenea de informaţia de mişcare, este construcţia *imaginilor de tip "mozaic"*. Acestea sunt imagini panoramice ce reprezintă întregul conţinut al unui plan video sau chiar al unei porţiuni a secvenţei cu o singură imagine [Irani 95] (vezi Figura 5.2). Avantajul imaginilor "mozaic" constă în faptul că acestea înglobează întreaga dinamică a scenei în doar o singură imagine. Pe de altă parte, acestea nu au sens decât pentru pasaje ale secvenţei ce conţin o mişcare globală a camerei video, neputând fi astfel generate pentru pasaje statice sau pasaje ce conţin mişcări complexe ale scenei. Pentru a compensa acest neajuns, de regulă este adoptată o strategie hibridă: imaginile "mozaic" sunt determinate doar în situaţiile în care acest lucru este posibil, rămânând ca în celelalte cazuri conţinutul planelor video să fie reprezentat în mod clasic cu imagini cheie.

Figura 5.2: Exemplu de imagine de tip "mozaic" obţinută pentru o deplasare la dreapta a camerei video (sursă "http://scien.stanford.edu/2002projects/ee392j/Roman_Gilat/").

Pe lângă metodele enumerate anterior, putem menţiona şi o serie de abordări mixte ce folosesc colaborarea mai multor surse de informaţie din secvenţă, precum culoarea şi mişcarea, profitând astfel de avantajele fiecăreia dintre acestea. Un exemplu în acest sens este metoda din [Doulamis 00a],

unde imaginile cheie sunt extrase ca fiind puncte similare ale curbei formate de vectorii de caracteristici ai fiecărei imagini. Aceştia sunt obţinuţi pe baza unui proces de segmentare aplicat, atât la nivel de culoare cât şi de mişcare.

În general metodele mixte, cu toate că sunt mai dificil de implementat datorită complexităţii de calcul mai ridicate, oferă rezultate mai bune decât celelalte metode datorită diversităţii surselor de informaţii folosite.

Extragerea imaginilor cheie la nivel de segment

Una dintre constrângerile strategiei de extragere a imaginilor cheie la nivel de plan video este dată de lipsa de scalabilitate a unei astfel de abordări. În cazul secvenţelor de imagini cu o durată semnificativă, numărul de imagini obţinut cu o astfel de abordare este ridicat, ajungând uşor la ordinul sutelor, ceea ce face nerentabilă vizualizarea integrală de către utilizator a acestora.

Soluţia la această problemă constă în folosirea unităţilor structurale de nivel ierarhic superior planelor video (vezi Secţiunea 2.1), numite generic şi *segmente video*: scene, episoade, anumite evenimente sau chiar secvenţa în totalitate. Astfel, în funcţie de aplicaţie, se poate opta pentru un nivel de granularitate, de la ridicat (număr mai mare de imagini, de exemplu obţinut prin extragerea la nivel de plane sau scene), până la un nivel redus (număr redus de imagini, de exemplu prin extragerea la nivel de episoade sau evenimente).

O astfel de metodă ce foloseşte un nivel structural superior planelor video este propusă în [Sun 00]. Într-o primă etapă, întreaga secvenţă este segmentată în mod uniform, în ceea ce autorii numesc "unităţi de lungă durată" (segmente). Pentru fiecare segment astfel obţinut este calculată o măsura de schimbare prin evaluarea distanţei dintre histogramele primei şi respectiv ultimei imagini a segmentului. Segmentele din întreaga secvenţă sunt ordonate şi clasificate pe baza valorilor acestei măsuri, în două clase, şi anume: clasa A ce va conţine valorile de schimbare redusă şi clasa B ce va conţine valorile de schimbare importantă. Mai departe, imaginile cheie sunt extrase în felul următor: pentru segmentele din clasa A, prima şi ultima imagine a segmentului vor fi selectate ca imagini cheie, iar pentru segmentele din clasa B, toate imaginile segmentului vor fi selectate ca imagini cheie. Dacă rezumatul obţinut satisface numărul de imagini dorit, atunci algoritmul se încheie. În caz contrar, rezumatul obţinut va ţine loc de secvenţă, iar procesul de selecţie este repetat. Limitarea acestei metode este dată de măsura de similaritate folosită, diferenţa dintre prima şi ultima imagine a segmentului nefiind suficientă pentru a evalua variabilitatea conţinutului acestuia. Astfel, metoda va neglija conţinutul anumitor segmente importante pentru care imaginea de început şi de sfârşit sunt similare, şi respectiv va reprezenta

cu un număr important de imagini redundante segmentele cu un conţinut omogen, dar care încep şi se termină cu imagini relativ diferite.

Metoda propusă în [Doulamis 00b] introduce o strategie fuzzy pentru extragerea imaginilor cheie. În primă fază, fiecare imagine a secvenţei este segmentată folosind algoritmul RSST ("Recursive Shortest Spanning Tree", vezi [Kwok 04]). Atributele de culoare şi de mişcare obţinute în urma segmentării sunt clasificate într-un număr predefinit de clase folosind o clasificare fuzzy (vezi Secţiunea 7.1.2). Imaginile cheie sunt extrase mai departe prin minimizarea unei funcţii de cost, determinată cu un algoritm genetic pe baza unei măsuri a intercorelaţiei dintre date. În acest caz, principalul inconvenient al acestei metode este dat de complexitatea de calcul ridicată.

Un alt exemplu este metoda propusă în [Gong 00] ce foloseşte descompunerea în valori singulare sau SVD ("Singular Value Decomposition"[6]). Pentru întreaga secvenţă este creată mai întâi o matrice A de vectori de caracteristici. Aceştia sunt obţinuţi pentru un anumit set, suficient de vast, de imagini din secvenţă. Algoritmul SVD este aplicat matricei A. Pe baza acestuia se obţine pe de-o parte rafinarea spaţiului de caracteristici ce va facilita o mai bună clasificare a imaginilor similare din punct de vedere vizual, dar şi determinarea unei metrici ce va permite măsurarea cantitativă a schimbărilor vizuale din fiecare clasă de imagini. Mai departe, în spaţiul de caracteristici obţinut după aplicarea SVD, clasa cu conţinutul cel mai puţin variabil este luată ca referinţă, iar valoarea de schimbare a conţinutului vizual al acesteia în corelaţie cu distanţa dintre imagini sunt folosite ca praguri pentru clasificarea restului de imagini. În cele din urmă, imaginile cheie sunt alese ca fiind imaginile cu vectorii de caracteristici cei mai apropiaţi de centroizii claselor. Această abordare asigură o minimă redundanţă informaţională a rezumatului, precum şi posibilitatea ca numărul de imagini din rezumat să fie reglat de utilizator. Totuşi, o constrângere este dată de faptul că datorită clasificării, în rezumatul obţinut nu se va mai respecta ordinea temporală iniţială a secvenţei.

Alte abordări

Alte abordări de extragere a imaginilor cheie se folosesc de alte spaţii de reprezentare a informaţiei vizuale, precum spaţiul transformatei "wavelet"[7] [Campisi 99] sau se bazează pe localizarea în secvenţă a anumitor pasaje de

[6] în algebra liniară, descompunerea în valori singulare sau SVD, reprezintă descompunerea ("factorizarea") unei matrice pătratice de numere reale sau complexe la forma canonică (forma standard). Ca exemple de aplicaţii ce folosesc SVD putem menţiona: calcularea matricei pseudo-inverse, aproximarea unei matrice sau determinarea rangului.

[7] vezi explicaţia de la pagina 104.

interes, ca de exemplu prezenţa feţelor umane sau a culorii pielii [Dufaux 00]. Metoda propusă în [Dufaux 00] se foloseşte de analiza de mişcare precum şi de analiza activităţii spaţiale, ce integrează detecţia feţelor umane şi a prezenţei culorii pielii în imagine, pentru a rezuma conţinutul secvenţei cu o singură imagine. În acest caz, sunt considerate ca reprezentative imaginile ce conţin personaje sau portrete.

O altă abordare ce se foloseşte de numărul de obiecte prezente în scenă pentru a localiza pasajele importante ale secvenţei, este propusă în [Kim 00a]. O imagine cheie este extrasă de fiecare dată când numărul de obiecte prezente în imaginea curentă este diferit de numărul de obiecte din imaginea ulterioară temporal.

Principalul inconvenient al metodelor bazate pe localizarea anumitor pasaje de interes din secvenţă este dat de faptul că acestea sunt dependente şi adaptate fiecărui domeniu de aplicaţie, sau în unele situaţii, chiar tipului de secvenţă folosit, neavând astfel un caracter generic.

5.1.2 Mecanismul de extragere a imaginilor cheie

Metodele existente de construcţie a rezumatelor statice folosesc o serie de mecanisme de selectare a imaginilor cheie din secvenţă. Acestea, în funcţie de strategia folosită, se împart în următoarele categorii [Truong 07]:

- selectarea imaginilor pe baza *schimbării suficiente* a conţinutului,
- selectarea imaginilor pe baza *egalizării varianţei temporale*,
- selectarea imaginilor astfel încât acestea să asigure o *acoperire maximală* a conţinutului,
- selectarea imaginilor cu metode de *clasificare*,
- selectarea imaginilor pe baza *corelaţiei minimale* dintre acestea,
- selectarea imaginilor astfel încât *eroarea de reconstrucţie* a secvenţei să fie minimală,
- selectarea imaginilor prin *simplificarea curbei de caracteristici*,
- selectarea imaginilor pe baza *localizării evenimentelor* importante din secvenţă.

În cele ce urmează, vom prezenta particularităţile fiecărei dintre aceste abordări aşa cum au fost prezentate în [Truong 07].

Schimbarea suficientă de conţinut

Această abordare analizează imaginile secvenţei în mod secvenţial. O imagine analizată este selectată ca imagine cheie dacă conţinutul vizual al acesteia prezintă o schimbare importantă, relativ la imaginile cheie extrase anterior.

În general, metodele ce folosesc criteriul schimbării suficiente de conţinut aleg o nouă imagine cheie, de indice r_{i+1} în rezumat, în funcţie de imaginea cheie extrasă anterior, de indice r_i, pe baza relaţiei următoare:

$$r_{i+1} = argmin|_t\{C(I_t, I_{r_i}) > \varepsilon, i < t \leq N\} \tag{5.2}$$

unde I_t reprezintă imaginea de indice temporal t în pasajul analizat al secvenţei (de exemplu, acesta poate fi un plan video, o scenă, etc.), N reprezintă numărul total de imagini conţinute în acesta, $C()$ este o funcţie ce măsoară schimbarea conţinutului vizual iar ε reprezintă pragul ce defineşte o schimbare de conţinut ca fiind semnificativă. De regulă, prima imagine a rezumatului, de indice r_1, este aleasă ca fiind prima imagine a pasajului analizat. Dintre cele mai frecvent folosite funcţii $C()$ putem menţiona: diferenţa dintre histogramele de culoare, funcţia de energie cumulativă sau schimbarea proprietăţilor geometrice ale obiectelor, etc.

Unul dintre principalele avantajele ale acestei abordări este dat de faptul că rezumatul static obţinut va avea un număr variabil de imagini, fiind adaptat la conţinutul secvenţei şi reflectând evoluţia dinamică a acesteia, indiferent de durata sau de tipul conţinutului de acţiune. Pe de altă parte, principalul inconvenient este dat de faptul că imaginile cheie sunt extrase fără a ţine cont de conţinutul secvenţei ce precede imaginea cheie curentă, ceea ce face ca rezumatul obţinut să nu ofere o acoperire maximală a conţinutului secvenţei [Truong 07].

În [Xiong 97] este propusă o extensie a acestei abordări către metoda numită "Seek and Spread", ce nu restricţionează alegerea noii imagini cheie în funcţie de imaginea cheie curentă din rezumat. Dacă b_i reprezintă indicele imaginii de referinţă curente, cu $b_1 = 1$, atunci r_i, indicele imaginii cheie, şi b_{i+1}, indicele noii imagini de referinţă, sunt calculate în felul următor:

$$r_i = argmin|_t\{C(I_t, I_{b_i}) > \varepsilon, t > b_i\} \tag{5.3}$$

$$b_{i+1} = argmin|_t\{C(I_t, I_{r_i}) > \varepsilon, t > r_i\} \tag{5.4}$$

unde imaginea cheie I_{r_i}, din rezumat, este imaginea considerată ca reprezentativă pentru segmentul din secvenţă delimitat de indicii $[b_i, b_{i+1}-1]$. Măsura $C()$ este calculată în acest caz ca fiind distanţa Euclidiană dintre o serie de vectori de caracteristici obţinuţi pe baza descompunerii "wavelet". Totuşi, în ciuda îmbunătăţirii aduse, această metodă nu este scalabilă iar rezumatul obţinut este asimetric.

Un alt exemplu este metoda propusă în [Rasheed 03]. Rezumatul obţinut cu aceasta este mai concis şi se bazează pe extragerea unei imagini cheie de fiecare dată când imaginea curentă analizată diferă suficient faţă de toate imaginile cheie extrase anterior, astfel fiind luat în calcul, chiar dacă la un nivel de detaliu redus, şi conţinutul precedent imaginii curente din rezumat.

Egalizarea varianţei temporale

În această abordare, numărul de imagini cheie al rezumatului static este cunoscut "a priori". Principiul metodei de egalizare a varianţei temporale constă în alegerea imaginilor cheie (indice r_i în rezumat) ca fiind reprezentative pentru anumite segmente ale secvenţei (intervale $[b_i, b_{i+1}-1]$) ce prezintă o varianţă temporală similară.

Valorile indicilor r_i şi b_i vor fi determinate independent, în doi paşi. Din punct de vedere teoretic, frontierele b_i ale segmentelor secvenţei sunt selecţionate astfel încât să satisfacă egalitatea următoare:

$$\nu(b_1, b_2) = \nu(b_2, b_3) = ... = \nu(b_k, b_{k+1}) \tag{5.5}$$

unde $\nu(b_i, b_{i+1})$ reprezintă varianţa temporală a segmentului $[b_i, b_{i+1} - 1]$, segment ce este reprezentat în rezumat cu imaginea cheie I_{r_i}.

Cum în realitate egalitatea exactă a valorilor varianţei este aproape imposibil de obţinut, [Sun 00] propune transformarea acesteia într-o problemă de optimizare, astfel:

$$\{b_1, b_2, ..., b_{k+1}\} = argmin|_{b_i} \sum_{i=1}^{k} \sum_{j=1}^{k} |\nu(b_i, b_{i+1}) - \nu(b_j, b_{j+1})| \tag{5.6}$$

Funcţia de varianţă temporală, $\nu()$, poate fi aproximată cu o măsură a schimbării cumulative a conţinutului dintre imagini sau chiar cu diferenţa dintre prima şi ultima imagine a fiecărui segment. Această ipoteză permite soluţionarea ecuaţiei 5.6 într-un mod recursiv.

După localizarea segmentelor ce îndeplinesc criteriul egalităţii varianţei temporale, imaginile cheie pot fi extrase folosind mai multe strategii. De exemplu, în [Fauvet 04] imaginile cheie sunt imaginile de început şi de mijloc ale fiecărui segment. O abordare mai complexă este propusă în [Lee 02a], unde imaginea I_{r_i} este selectată ca imagine cheie reprezentativă pentru conţinutul segmentului $[b_i, b_{i+1} - 1]$ prin minimizarea diferenţei cumulative următoare:

$$r_i = argmin|_{b_i \leq t < b_{i+1}} \sum_{j=b_i}^{b_{i+1}-1} D(I_j, I_t) \tag{5.7}$$

unde $D()$ reprezintă un anumit operator de distanţă între imagini.

În general, metodele din această categorie prezintă o complexitate de calcul superioară metodelor ce folosesc schimbarea suficientă a conţinutului vizual, drept criteriu de selecţie al imaginilor cheie, dar, în revanşă, rezumatul obţinut este de această dată optimal din punct de vedere global.

Acoperirea maximală a conţinutului

Această strategie este cunoscută şi sub numele de metoda de selecţie a imaginilor cheie pe baza criteriului de fidelitate. Principiul metodei constă în a determina pentru fiecare imagine din secvenţă, sau dintr-un anumit pasaj al acesteia, a unei liste de imagini pe care aceasta le poate substitui, listă ce poartă numele de acoperire a imaginii.

Dacă notăm cu $C_i(\varepsilon)$ lista imaginilor din pasajul V al secvenţei ce pot fi substituite cu imaginea cheie I_i pe baza unei valori de toleranţă ε, atunci, setul optimal al imaginilor cheie din V, determinat fără constrângeri de debit (număr de imagini), este dat de ecuaţia următoare:

$$\{r_1, r_2, ..., r_k\} = argmin|_{r_i}\{k/C_{r_1}(\varepsilon) \cup C_{r_2}(\varepsilon) \cup ... \cup C_{r_k}(\varepsilon) = V\} \quad (5.8)$$

unde r_i reprezintă indicele imaginii cheie din rezumat.

Dacă alegem drept constrângere numărul de imagini din rezumat, atunci problema selecţiei imaginilor cheie poate fi văzută din două perspective. Pe de-o parte, putem căuta valoarea minimală a lui ε pentru care toate imaginile pot fi reprezentate cu cel puţin o singură imagine cheie, astfel:

$$\{r_1, r_2, ..., r_k\} = argmin|_{r_i}\{\varepsilon/\ C_{r_1}(\varepsilon) \cup C_{r_2}(\varepsilon) \cup ... \cup C_{r_k}(\varepsilon) = V\} \quad (5.9)$$

sau pe de altă parte, putem căuta un grup de imagini ce pot fi reprezentative pentru cea mai mare parte a imaginilor disponibile, astfel:

$$\{r_1, r_2, ..., r_k\} = argmin|_{r_i}\{|C_{r_1}(\varepsilon) \cup C_{r_2}(\varepsilon) \cup ... \cup C_{r_k}(\varepsilon)|\} \quad (5.10)$$

Astfel, în procesul de construcţie al rezumaului apar două etape distincte, şi anume: determinarea *acoperirii imaginilor* în secvenţă şi respectiv *optimizarea selecţiei* făcute.

În [Yahiaoui 01] acoperirea unei imagini este dată de numărul de extrase din secvenţă ("excerpts"), de dimensiune fixă, ce conţin cel puţin o imagine similară cu imaginea curent analizată. Imaginile rezumatului sunt selectate mai departe folosind procedeul programării dinamice. Un alt exemplu este metoda propusă în [Rong 04] ce face analogie între procesul de selecţie al imaginilor cheie din secvenţă şi procesul de extragere a cuvintelor pe baza

metodei TF-IDF ("Term-Frequency Inverse Document Frequency"[8]) folosit în sistemele de căutare a textului. Astfel, o imagine a secvenţei va fi selectată ca fiind imagine cheie reprezentativă pentru un anumit plan video, dacă aceasta are o acoperire redusă pentru restul conţinutului secvenţei.

Unul dintre principalele avantaje ale metodelor din această categorie este dat de faptul că nu se impune ca acoperirea unei imagini să fie un segment continuu din punct de vedere temporal, cum este cazul metodelor ce folosesc ca criteriu schimbarea suficientă de conţinut sau egalitatea varianţei temporale. Acest lucru, va permite obţinerea unui rezumat mai concis. Pe de altă parte, complexitatea de calcul creşte datorită calculului pentru fiecare pereche de imagini, a unei măsuri de similaritate.

Clasificarea imaginilor

Metodele de selecţie a imaginilor cheie din această categorie se folosesc de metode de clasificare automată a datelor pentru a regrupa imaginile în funcţie de similaritatea conţinutului acestora (pentru un studiu detaliat al metodelor de clasificare existente cititorul se poate raporta la Capitolul 7).

Principiul clasificării constă în reprezentarea fiecărei imagini cu un vector de atribute (caracteristici) ce sunt extrase la nivel de imagine, ca de exemplu: histograme de culoare, parametri de textură, etc. Acestea formează în spaţiul de caracteristici un nor de puncte. Pe baza evaluării unei anumite măsuri de similaritate între vectorii de atribute, imaginile (punctele din spaţiul de caracteristici) vor fi mai departe alocate unei anumite clase. Imaginile cheie ale rezumatului static vor fi alese din clasele considerate ca reprezentative pentru conţinutul secvenţei.

Procesul de constituire a claselor de imagini implică în general patru etape de analiză, şi anume:

- *o etapă de pre-analiză*: aceasta are ca scop să reducă redundanţa datelor de intrare şi astfel să îmbunătăţească eficienţa algoritmului de clasificare. Printre metodele cele mai cunoscute de decorelare a datelor, putem menţiona analiza în componente principale sau PCA[9] [Gibson 02] ce se foloseşte de descompunerea în vectori primi. O altă strategie constă în a reduce, înaintea clasificării, numărul de imagini

[8]metoda TF-IDF presupune calculul unei măsuri statistice folosită la evaluarea importanţei unui anumit cuvânt într-un document textual dintr-o anumită colecţie sau corpus. Importanţa acestuia va creşte proporţional cu numărul de apariţii în document. Variaţii ale metodei TF-IDF sunt folosite frecvent în motoarele de căutare pentru a calcula relevanţa şi rangul unui document relativ la cererea de căutare formulată de utilizator.

[9]vezi explicaţia de la pagina 123.

din secvenţă folosind de exemplu imaginile cheie obţinute cu una dintre metodele de rezumare prezentate anterior.

- *o etapă de clasificare propriu-zisă*: aceasta efectuează repartiţia imaginilor în clase. Printre metodele cel mai frecvent folosite putem enumera următoarele: clasificarea secvenţială, metoda "Hierarchical Complete Link", clasificarea Fuzzy C-Means [Yu 04] sau folosirea modelelor Gaussiene mixte (GMM - "Gaussian Mixture Models") [Gibson 02].

- *o etapă de filtrare a claselor obţinute*: în urma procesului de împărţire în clase este foarte probabil ca anumite clase, de exemplu, afectate de prezenţa zgomotului, să nu conţină informaţii relevante. Pentru aceasta, este necesară o etapă de triere a claselor. Astfel, putem considera ca relevante doar clasele cu o dimensiune (număr de imagini) superioară dimensiunii medii, sau, pentru a evita prezenţa "artefactelor", putem selecta doar clasele ce conţin cel puţin o secvenţă continuuă de imagini, cu o durată superioară unui anumit prag (de exemplu 9 secunde [Truong 07]).

- *o etapă de selecţie a imaginilor cheie*: soluţia cea mai intuitivă şi cea mai frecvent folosită constă în alegerea ca imagini cheie a imaginilor ai căror vectori de caracteristici se află cel mai aproape de centroidul clasei [Yu 04]. În acest fel, avem siguranţa că imaginile selectate sunt cele mai reprezentative pentru conţinut, deoarece, centroidul clasei are sens de vector mediu în spaţiul n-dimensional de caracteristici.

Una dintre principalele limitări ale selecţiei imaginilor cheie pe baza clasificării imaginilor secvenţei este dată de faptul că în general, informaţia semantică vizuală a secvenţei implică o varianţă intra-clasă, în acelaşi timp importantă cât şi redusă, ceea ce face dificilă separarea în clase. Mai mult, în acest caz rezumatul obţinut în urma clasificării nu va respecta evoluţia temporală a secvenţei.

Corelaţia minimală a imaginilor cheie

Metodele din această categorie consideră că imaginile rezumatului static trebuie să fie cât mai puţin corelate între ele. Totuşi, în practică, deseori se ia în calcul doar corelaţia dintre imagini vecine. Principiul de selecţie a imaginilor cheie pe baza corelaţiei minimale poate fi formulat în felul următor:

$$\{r_1, r_2, ..., r_k\} = argmin|_{r_i}\{Corr(I_{r_1}, I_{r_2}, ..., I_{r_k})\} \tag{5.11}$$

unde r_i reprezintă indicele imaginii cheie în rezumat iar $Corr()$ reprezintă operatorul de corelaţie.

În [Doulamis 98] valoarea corelaţiei pentru ansamblul imaginilor cheie este redefinită pentru a ţine cont de contribuţia corelaţiei dintre fiecare pereche de imagini, astfel:

$$Corr^2(I_{r_1}, I_{r_2}, ..., I_{r_k}) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Corr^2(I_{r_i}, I_{r_j}) \tag{5.12}$$

unde $Corr(I_i, I_j)$ reprezintă funcţia de corelaţie dintre vectorii de caracteristici ai imaginilor de indice i şi j. Pentru ca soluţia ecuaţiei 5.11 să fie una optimală, se pot folosi o serie de metode, ca de exemplu căutarea logaritmică, căutarea stohastică sau algoritmii genetici [Doulamis 00b].

O soluţie aproximativă este propusă în [Liu 02c]. Complexitatea de calcul este redusă la ordinul $O(n \cdot log(n))$, unde n reprezintă numărul de imagini analizate, pe baza algoritmului "Greedy"[10]. La început se porneşte cu toate cele n imagini ale secvenţei drept imagini cheie, ceea ce constituie nivelul n de analiză. Mai departe, se trece iterativ de la un nivel de analiză la altul. La trecerea de la nivelul de analiză curent, t, la un nivel inferior, $t-1$, imaginile ce prezintă o corelaţie minimală în raport cu imaginile vecine sunt eliminate din rezumat.

Criteriul corelaţiei minimale asigură un nivel redus de redundanţă a imaginilor din rezumat. Totuşi, acest criteriu este sensibil la prezenţa variaţiilor importante din secvenţă (numite în literatura de specialitate şi "outliers"[11]). Pentru a diminua acest efect, soluţia propusă în [Porter 03] constă în a reprezenta secvenţa sub forma unui graf[12]. În acesta, fiecare imagine corespunde unui nod iar corelaţia dintre imagini, calculată în acest caz în funcţie de informaţia de mişcare, reprezintă legăturile dintre noduri. Setul optimal de imagini cheie va fi dat de drumul cel mai scurt ce uneşte prima şi ultima imagine a grafului.

Eroarea minimală de reconstrucţie

Metodele din această categorie se folosesc pentru construcţia rezumatului de o anumită metrică, numită şi SRE (eroarea de reconstrucţie a secvenţei),

[10] algoritmul "Greedy" este un algoritm general de soluţionare a unei probleme de calcul. Aceasta este abordată în mai multe faze. Pentru fiecare fază este adoptată o soluţie optimală local, fără a se ţine cont de implicaţiile ulterioare ale acesteia. La finalizarea algoritmului, este de aşteptat ca optimul local să conveargă spre optimul global. În caz contrar, soluţia obţinută este considerată ca fiind suboptimală.

[11] în statistica matematică, un "outlier" reprezintă o observaţie a cărei reprezentare numerică este foarte diferită de restul valorilor datelor considerate.

[12] în terminologia matematică, un graf este definit ca fiind o colecţie de puncte şi linii ce interconectează anumite subseturi ale acesteia (posibil vide). Punctele unui graf sunt numite şi noduri în timp ce liniile ce le conectează se numesc muchii.

ce măsoară capacitatea rezumatului de a reproduce conţinutul original al secvenţei. Folosirea măsurii SRE se dovedeşte a fi eficientă în două situaţii, şi anume: dacă numărul de imagini al rezumatului este cunoscut "a priori" sau dacă se doreşte ca imaginile din rezumat să păstreze evoluţia temporală din secvenţă.

Dacă considerăm o funcţie de interpolare a imaginilor, $F_{int}(t, R)$, ce permite reconstrucţia unei imagini la momentul t din unitatea V a secvenţei (de exemplu: un plan video, un segment, etc.) pornind de la imaginile rezumatului R, atunci mărimea SRE, notată $\varepsilon(V, R)$, este dată de relaţia următoare:

$$\varepsilon(V, R) = \sum_{i=1}^{n} D(I_i, F_{int}(i, R)) \tag{5.13}$$

unde n reprezintă numărul total de imagini din unitatea V a secvenţei iar $D()$ este un operator de distanţă între imagini.

Imaginile cheie ale rezumatului sunt selectate folosind drept criteriu ipoteza că un rezumat optimal trebuie să aibă un scor SRE minim, astfel:

$$\{r_1, r_2, ..., r_k\} = argmin|_{r_i}\{\varepsilon(V, R), 1 \leq r_i \leq n\} \tag{5.14}$$

unde r_i reprezintă indicele imaginii cheie din rezumat iar k reprezintă numărul de imagini al rezumatului ce este fixat "a priori".

Dintre metodele existente ce folosesc măsura SRE putem exemplifica metodele propuse în [Liu 02b] sau [Li 04].

În [Liu 02b] rezumatul static este generat folosind metoda "Heap-Based Greedy"[13]. Algoritmul de selecţie a imaginilor cheie porneşte cu toate cele n imagini ale secvenţei ca imagini cheie. Acesta constituie nivelul n de analiză. Mai departe, imaginile cheie sunt eliminate iterativ la trecerea de la un nivel de analiză la altul. Astfel, imaginile cheie din nivelul de analiză $t-1$ sunt determinate pe baza imaginilor cheie din nivelul superior t prin eliminarea imaginilor ce minimizează scorul SRE.

Metoda propusă în [Li 04] încearcă să determine un rezumat optimal pe baza metodei de programare dinamică. Aceasta foloseşte drept constrângere egalitatea:

$$b_i = r_i \tag{5.15}$$

unde r_i reprezintă indicele imaginii cheie din rezumat ce este reprezentativă pentru segmentul secvenţei definit de indicii $[b_i, b_{i+1} - 1]$. Cu alte cuvinte, fiecare imagine cheie este determinată în funcţie de imaginile ce o succed.

[13] algoritmul "Heap-Based Greedy" este un algoritm "Greedy" (vezi explicaţia anterioară de la pagina 159) ce se foloseşte de acumularea valorilor într-o coadă, reducând astfel complexitatea de calcul.

Rezumatul optimal este obţinut cu o căutare prin înjumătăţire ("bisection search"[14]).

Simplificarea curbei de caracteristici

Similar metodelor de clasificare, metodele din această categorie reprezintă secvenţa sub forma unei curbe de caracteristici în spaţiul multidimensional format de vectorii de atribute ai imaginii. Una dintre dimensiunile acestui spaţiu va fi dată de axa temporală a secvenţei. Principiul simplificării curbei de caracteristici constă în izolarea unui set de puncte considerate ca fiind reprezentative pentru forma curbei şi care păstrează o bună aproximare a acesteia. Aceste puncte, nu este obligatoriu să fie distribuite uniform după axa temporală.

Ca exemplu de astfel de metode putem menţiona abordările propuse în [Latecki 01] şi [Calic 02a], unde selecţia imaginilor cheie este realizată cu ajutorul algoritmul "Discrete Contour Evolution" de analiză al evoluţiei contururilor din imagine. Acesta este aplicat curbei de evoluţie temporală a diferenţelor dintre histogramele imaginilor din secvenţă. Principiul folosit constă în înlocuirea iterativă a fiecărei perechi de segmente consecutive ale curbei de caracteristici ce prezintă un scor de relevanţă minim, cu segmentul ce reuneşte primul nod al primului segment cu ultimul nod al celui de-al doilea segment din pereche. Procesul se opreşte în momentul în care numărul dorit de imagini al rezumatului este atins. În [Calic 02a] imaginile cheie astfel obţinute sunt rafinate mai departe prin reţinerea doar a imaginilor ce corespund punctelor de minim local ale curbei de caracteristici simplificate.

Rezumatul obţinut cu această strategie are proprietatea de a respecta evoluţia temporală a secvenţei.

Localizarea evenimentelor importante

Metodele din această categorie selectează imaginile cheie ca fiind imagini reprezentative ale anumitor pasaje ale secvenţei ce prezintă un interes semantic (de exemplu: anumite evenimente, pasaje de acţiune, prezenţa personajelor, etc.).

Metoda propusă în [Dufaux 00] rezumă întreaga secvenţă cu o singură imagine cheie. Algoritmul propus implică două etape. În prima etapă sunt selectate din secvenţă doar planele video considerate ca importante. Acestea

[14] căutarea de tip "bisection search" este un algoritm simplu de căutare ce se bazează pe localizarea soluţiei prin înjumătăţirea iterativă a intervalului de căutare. La fiecare iteraţie, intervalul curent de căutare devine jumătatea intervalului din iteraţia anterioară în care este cel mai probabil să se afle soluţia.

sunt planele ce îndeplinesc următoarele cerinţe: durata lor este importantă, au un conţinut de mişcare semnificativ, o activitate spaţială importantă iar în acestea apar personaje umane. Mai departe, în a doua etapă, imaginea cheie este aleasă folosind criterii similare cu cele menţionate anterior, cu excepţia faptului că pentru a evita ca imaginea selectată să fie o imagine înceţoşată (datorită efectului de "motion blur"[15]) sau să conţină "artefacte"[16], aceasta nu va fi aleasă dintr-un pasaj de mişcare, ci mai degrabă dintr-un pasaj cu o activitate spaţială importantă.

Un alt exemplu este metoda propusă în [Liu 03] unde extragerea imaginilor cheie este realizată pe baza conţinutului de mişcare. Pentru aceasta, planele video sunt mai întâi descompuse în segmente ce conţin o mişcare continuă din punct de vedere al acceleraţiei şi respectiv deceleraţiei mişcării. Imaginile cheie sunt alese, mai departe, ca fiind imaginile ce fac joncţiunea între un segment în care mişcarea este accelerată şi un segment cu mişcare decelerată. Dacă planul video analizat nu prezintă un conţinut de mişcare, atunci este aleasă doar o singură imagine cheie şi anume imaginea de început a planului.

Metoda propusă în [Calic 04] selectează imaginile la nivel de plan video prin localizarea anumitor evenimente de interes din interiorul planului. Acestea sunt marcate de comportamentul particular al anumitor regiuni din imagine. Astfel, un eveniment este detectat dacă o regiune din imagine dispare din scenă, dacă aceasta se combină cu o alta sau dacă aceasta devine plan principal al scenei.

Una dintre limitările unei astfel de abordări este dată de faptul că localizarea evenimentelor considerate ca importante pentru conţinut este guvernată de reguli eurisitice. Acestea sunt de regulă definite pe baza observării proprietăţilor unui set redus de date, fiind astfel mai mult sau mai puţin adaptate unui anumit domeniu de aplicaţie.

5.2 Construcţia rezumatelor dinamice

Un rezumat dinamic, sau "video skim", reprezintă o colecţie de segmente sau pasaje ale secvenţei. Acesta constituie el însuşi o secvenţă de imagini, dar, de o durată mult inferioară secvenţei iniţiale.

Din punct de vedere formal, rezumatul dinamic $R_{seq}(S)$ al secvenţei S poate fi exprimat astfel:

$$R_{seq}(S) = seg_1 \cup seg_2 \cup ... \cup seg_M \tag{5.16}$$

[15] vezi explicaţia de la pagina 81.
[16] vezi explicaţia de la pagina 101.

unde seg_i reprezintă segmentul video de indice i în rezumat, cu $i = 1, ..., M$, iar M reprezintă numărul total de segmente prezente în rezumat.

Spre deosebire de mecanismele de rezumare în imagini, generarea unui rezumat dinamic este o operaţie mult mai laborioasă, având o complexitate de calcul ridicată datorată în principal necesităţii de analiză şi înţelegere a conţinutului secvenţei la un nivel semantic superior. Dacă în cazul rezumării statice, unitatea de analiză de bază o constituie imaginea, în cazul rezumării dinamice, aceasta este segmentul video (sub-secvenţă de imagini).

După cum am menţionat şi în partea introductivă a acestui capitol, rezumatele dinamice pot fi generate pornind de la rezumatele statice. O metodă imediată constă în înlocuirea fiecărei imagini cheie din rezumatul static cu un interval de imagini din secvenţă, ca de exemplu intervalul centrat pe imaginea cheie. Rezumatul astfel obţinut este o reprezentare compactă a conţinutului secvenţei, dar calitatea acestuia este dependentă de calitatea rezumatului static. Datorită limitărilor metodelor de rezumare în imagini (vezi secţiunea anterioară) este foarte probabil ca rezumatul dinamic obţinut astfel să nu fie reprezentativ pentru conţinutul de mişcare al secvenţei. Din această cauză, în cele mai multe situaţii este de preferat ca rezumatul să fie extras direct din secvenţă cu metode specifice.

O altă strategie rapidă de generare a unui rezumat dinamic constă în utilizarea segmentării în plane a secvenţei. Astfel, un posibil rezumat poate fi constituit prin reprezentarea fiecărui plan video cu un anumit pasaj al acestuia. Rezumatul astfel obţinut este o foarte bună sinteză a conţinutului global al secvenţei. Pe de altă parte, acesta va conţine atât pasajele importante cât şi cele mai puţin relevante din secvenţă, ceea ce face ca durata rezumatului să fie în cele mai multe situaţii foarte ridicată (de regula, în condiţiile asigurării unei continuităţi vizuale a rezumatului se obţine în medie o compresie de 1/3 a duratei secvenţei iniţiale).

Un rezumat dinamic eficient din punct de vedere al duratei precum şi al calităţii informaţiei furnizate, necesită în mod ideal o analiză de nivel semantic a conţinutului secvenţei, similară modului de percepţie uman. Datorită limitărilor metodelor de analiză semantică existente, ce nu au ajuns la un nivel de dezvoltare suficient pentru a facilita o astfel de analiză, metodele existente de rezumare dinamică tind să simplifice problema rezumării prin adoptarea de condiţii particulare. Din această cauză, metodele existente sunt în mare parte specifice domeniului de aplicaţie. Astfel întâlnim metode de rezumare distincte pentru: secvenţele sportive [Coldefy 04], secvenţele documentare [Yu 03], materialele video personale [Zhao 03], secvenţele de animaţie [Ionescu 06b], etc. Dificultatea analizei semantice este simplificată în acest caz prin folosirea expertizei domeniului respectiv.

Din punct de vedere al popularităţii metodelor de rezumare, literatura

de specialitate este mult mai bogată în tehnici de rezumate statică, decât în tehnici de rezumare dinamică. Acest lucru se datorează în principal complexităţii de calcul ridicate a acestora din urmă [Li 01]. Pentru un studiu complet al literaturii de specialitate, cititorul se poate raporta la lucrările [Lecce 99], [Truong 07] sau [Benoit 07].

5.2.1 Informaţia conservată de rezumat

O primă constrângere la generarea unui rezumat dinamic o constituie felul în care informaţia din secvenţă va fi reprezentată în rezumat. Aceasta depinde în general de aplicaţia vizată, căreia îi este destinat rezumatul, şi va determina modalitatea de generare a acestuia. În funcţie de modul de reprezentare al conţinutului secvenţei, metodele existente de rezumare dinamică se împart în trei categorii [Truong 07]:

- metode ce generează rezumate ce propun o acoperire *integrală* a conţinutului secvenţei,
- metode ce generează rezumate ce reproduc doar *anumite pasaje* ale unor evenimente importante pentru conţinutul secvenţei,
- metode ce produc rezumate *personalizate*, generate în funcţie de cerinţele utilizatorului.

Acoperirea integrală a conţinutului secvenţei

Metodele ce rezumă secvenţa astfel încât conţinutul acesteia să fie acoperit în totalitate de rezumat, au ca scop de a informa utilizatorul cu privire la conţinutul global al secvenţei [Sundaram 02] [Gong 03]. În acest caz, înţelegerea conţinutului iniţial al secvenţei nu este aproape deloc alterată de rezumat.

Acest tip de rezumat este util în cazul în care utilizatorul doreşte informaţii mai detaliate despre conţinutul dinamic al secvenţei, dar totuşi nu dispune de timpul necesar pentru a vizualiza secvenţa în totalitate. Putem spune că un astfel de rezumat este un compromis realizat asupra duratei rezumatului în favoarea completitudinii informaţiei furnizate de acesta.

Redarea evenimentelor importante ale secvenţei

Metodele de construcţie a rezumatelor dinamice cel mai frecvent întâlnite sunt cele care reproduc anumite evenimente importante ale secvenţei, fiind numite şi "video highlights". Acestea sunt de regulă specifice fiecărui domeniu de aplicaţie.

În funcţie de tipul evenimentelor redate de rezumat, putem menţiona următoarele abordări [Truong 07]:

- [Xiong 03]: rezumatul este constituit pe baza detecţiei evenimentelor ce antrenează reacţii particulare din partea publicului, ca de exemplu aplauze sau aclamaţii,
- [Coldefy 04]: rezumatul este construit pe baza pasajelor secvenţei ce provoacă o reacţie de entuziasm a naratorului,
- [Pan 01]: rezumatul este construit folosind pasajele din secvenţă ce sunt evidenţiate de producător prin tehnici de montaj specifice, ca de exemplu o frecvenţă ridicată a tranziţiilor de tip "cut", prezenţa textului sau reluarea anumitor scene ale secvenţei,
- [Radhakrishnan 04]: rezumatul este constituit pe baza detecţiei evenimentelor ce corespund unor anumite modele predefinite,
- [Yu 03]: rezumatul este construit pe baza pasajelor secvenţei ce sunt preferate de utilizator (de exemplu, vizualizate de mai multe ori).

O categorie aparte de "video highlights" o constituie rezumatele de tip "movie trailer". Acestea sunt rezumate publicitare, destinate de regulă promovării unui nou film, ce redau un colaj al pasajelor celor mai captivante şi mai reprezentative ale secvenţei. Localizarea automată a acestora în secvenţă este o operaţie dificilă, din această cauză se foloseşte expertiza domeniului de aplicaţie. De exemplu, este evident că pentru a construi rezumatul de tip "trailer" al unui meci de fotbal, evenimentele căutate vor fi momentele de gol precum şi aclamaţiile publicului.

Ca exemplu de metodă de generare a unui astfel de rezumat, putem menţiona metoda propusă în [Ionescu 06b] ce se foloseşte de analiza distribuţiei temporale a tranziţiilor video pentru a localiza pasajele de acţiune în filmele artistice de animaţie. În primă fază, activitatea temporală a secvenţei este reprezentată grafic sub forma unui semnal continuu temporal, de amplitudine arbitrară 1, ce trece prin valoarea 0 la apariţia unei tranziţii video (vezi Figura 5.3). Acesta constituie nivelul de analiză macro, la nivel de segment.

Un pasaj de acţiune este definit mai departe ca fiind un segment al secvenţei pentru care numărul de schimbări de plan survenite în ferestre temporale de durată $T = 5s$, este superior vitezei medii de schimbare de plan, $\bar{v}_T$. Valoarea lui $\bar{v}_T$ este estimată pentru întreaga secvenţă folosind

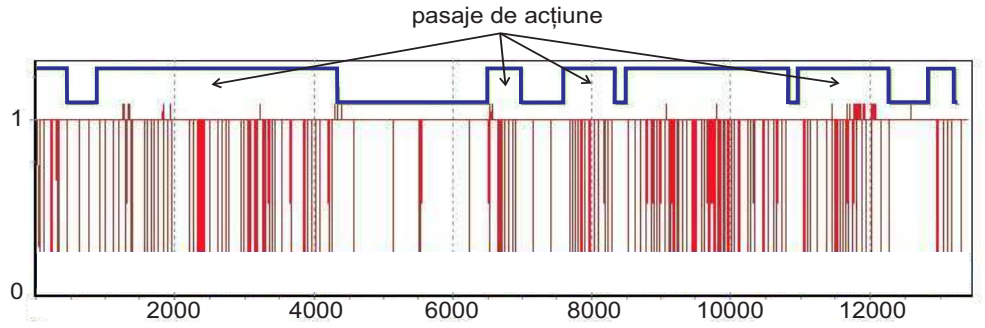


Figura 5.3: Exemplu de localizare a pasajelor de acţiune pe baza analizei frecvenţei schimbărilor de plan (axa oX este axa temporală, liniile roşii verticale reprezintă tranziţiile video iar linia albastră reprezintă un semnal binar ce ia valoarea 1 pentru un pasaj de acţiune şi valoarea 0 în rest, secvenţă "François le Vaillant" [Folimage 06]).

relaţia următoare:

$$\bar{v}_T = E\{\zeta_T\} = \sum_{t=1}^{T \cdot 25} t \cdot f_{\zeta_T}(t) \tag{5.17}$$

unde $f_{\zeta_T}()$ reprezintă densitatea de probabilitate a variabilei aleatoare discrete $\zeta_T()$ ce reprezintă numărul de schimbări de plan survenite în fereastra temporală de durată T, şi este dată de ecuaţia:

$$f_{\zeta_T}(t) = \frac{1}{N} \sum_{i=1}^{N} \delta(\zeta_T(i) - t) \tag{5.18}$$

unde N reprezintă numărul total de ferestre de analiză de durată T secunde, $\delta(t) = 1$ pentru $t = 0$ şi 0 altfel, iar i reprezintă indicele temporal al ferestrei de analiză curente pentru care se calculează valoarea lui $\zeta_T()$ (vezi Figura 5.3).

Rezumatul de tip "movie trailer" este constituit mai departe pe baza segmentelor de acţiune folosind o analiză intra-plan. Fiecare plan video dintr-un segment de acţiune este rezumat cu o sub-secvenţă a cărei durată este proporţională cu nivelul de activitate vizuală al imaginilor acestuia. Nivelul de activitate este evaluat pe baza calculului unei histograme de distanţe cumulate [Ionescu 06b].

Personalizarea conţinutului rezumatului

O altă categorie de rezumate dinamice sunt cele care folosesc personalizarea conţintului. Acestea sunt generate în funcţie de specificaţiile utilizatorului

cu privire la conţinutul secvenţei ce urmează să fie prezentat în rezumat. Pentru aceasta, utilizatorul îşi alege, fie un model predefint de conţinut, fie îşi exprimă criterile de selecţie sub forma unei cereri de tip "query". Ca exemple de astfel de metode de rezumare putem menţiona metoda propusă în [Lu 03], destinată rezumării secvenţelor de ştiri. În aceasta, utilizatorul poate opta pentru un rezumat care să prezinte pasajele secvenţei ce conţin feţe umane, pasajele de dialog, pasajele ce conţin o mişcare a camerei video de tip "zoom-in/zoom-out" sau pasajele ce conţin text încrustat în imagine. Un alt exemplu este metoda propusă în [Li 03] unde evenimentele vizate pentru a face parte din rezumat sunt scenele de dialog dintre două sau mai multe personaje, precum şi scenele hibride.

În cazul rezumatelor personalizate, procesul de selectare a segmentelor secvenţei ce vor face parte din rezumat este simplificat. Rezumatul obţinut nu va conţine decât informaţia care corespunde cerinţelor utilizatorului. Acest tip de rezumat poate fi văzut ca un rezumat semi-automat, deoarece în procesul de generare este necesară intervenţia utilizatorului. Rezumatul obţinut cu o astfel de strategie are un caracter generic redus, fiind dificil de aplicat într-un context necunoscut.

5.2.2 Procesul de generare a rezumatului dinamic

În general, procesul de constituire a unui rezumat dinamic implică o serie de etape de prelucare. Acestea pot fi sintetizate în următoarele [Truong 07]:

- *descompunerea* secvenţei în segmente (plane video, scene, etc.),
- *selecţia segmentelor* ce vor fi folosite la constituirea rezumatului,
- *reducerea dimensiunii* segmentelor prin reducerea redundanţei informaţionale sau pe baza anumitor constrângeri de conţinut,
- *integrarea multimodală* a segmentelor,
- *construcţia* propriu-zisă a rezumatului.

În funcţie de aplicaţie, este posibil ca anumite metode de rezumare să folosească doar unele dintre cele cinci etape enumerate mai sus, sau ca anumite etape să fie efectuate simultan.

Descompunerea în segmente

Prima etapă, indispensabilă procesului de constituire a rezumatului dinamic, constă în *descompunerea secvenţei în segmente*. Un segment poate fi un plan

video, o scenă sau un pasaj al secvenţei ce conţine un eveniment considerat drept important pentru conţinut. Noţiunea de segment nu se limitează doar la informaţia vizuală, descompunerea în segmente fiind efectuată, după caz, şi pentru celelalte modalităţi ale unui document video, precum sunetul (descompunere în segmente audio) şi textul (descompunere în cuvinte sau propoziţii).

Selecţia segmentelor

Etapa următoare descompunerii în segmente o reprezintă *selecţia segmentelor* ce vor fi folosite pentru constituirea rezumatului. Felul în care aceasta este realizată va influenţa atât *coerenţa* rezumatului cât şi *gradul de acoperire* al conţinutului secvenţei, şi va determina *contextul aplicaţiei* căreia îi este destinat rezumatul.

Ca exemple de tehnici de selecţie a segmentelor, putem menţiona metodele propuse în [Gong 03] şi [Ngo 03]. În [Gong 03] proprietăţile temporale şi spaţiale ale secvenţei sunt caracterizate pe baza descompunerii în valori singulare (SVD[17]) a unei matrice de caracteristici de culoare a imaginilor. Pe baza acesteia, planele video sunt selectate în funcţie de gradul de schimbare vizuală, de uniformitatea distribuţiei de culoare şi în funcţie de similaritate.

Metoda propusă în [Ngo 03] foloseşte un algoritm generalizat de detecţie a tranziţiilor de tip "cut" pentru a determina planele video, sau mai general, clasele de imagini din secvenţă. Acestea sunt analizate mai departe folosind un model al percepţiei umane de tip "Motion Attention Model". Valorile obţinute pentru gradul de atenţie sunt structurate sub forma unui graf şi sunt prelucrate folosind metode specifice analizei lanţurilor Markov[18]. Graful obţinut este folosit pentru a regrupa clasele de imagini similare în scene, iar gradul de atenţie este folosit drept criteriu de selecţie a segmentelor ce vor face parte din rezumat.

Reducerea dimensiunii segmentelor

În general, segmentele obţinute în urma etapei anterioare sunt redundante din punct de vedere al conţinutului, iar durata acestora tinde să fie prea ridicată pentru rezumat.

Etapa de *reducere a dimensiunii segmentelor* va asigura optimizarea acestora prin păstrarea doar a conţinutului esenţial, şi va conduce astfel la

[17] vezi explicaţia de la pagina 152.

[18] un lanţ Markov, numit astfel după numele matematicianului rus Andrey Markov, reprezintă o colecţie de variabile aleatoare, $\{X_t\}$, unde $t = 0, 1, ...$, ce au proprietatea că pentru o anumită stare prezentă a acestora, valorile viitoare sunt condiţional independente de cele anterioare.

obţinerea unui rezumat mai concis. Pentru aceasta, s-au propus diferite soluţii, ca de exemplu metoda propusă în [Cooper 02] ce foloseşte o matrice de autosimilaritate. Din fiecare segment este reţinut doar acel pasaj al cărui conţinut este cel mai reprezentativ pentru conţinutul segmentului întreg (din punct de vedere al similarităţii). Alte abordări folosesc metode de compresie a segmentelor prin eliminarea anumitor imagini sau pasaje redundante. Eliminarea acestora se face ţinând cont şi de informaţia audio, dacă aceasta este prezentă. Rezumatul obţinut trebuie să respecte limitele de coerenţă vizuală cât şi sonoră [Li 03].

Totuşi, datorită procesului de eliminare a informaţiei, această etapă poate introduce momente de discontinuitate vizuală în rezumat.

Integrarea multimodală

În general, segmentele extrase din secvenţă sunt *unimodale*. Acestea sunt, fie segmente vizuale, fie segmente audio sau segmente textuale. Etapa de *integrare multimodală* are ca scop tocmai fuzionarea şi sincronizarea tuturor acestor informaţii astfel încât rezumatul final să respecte evoluţia temporală a evenimentelor secvenţei. În acest scop, *alinierea segmentelor* joacă un rol important. Astfel, frontierele dintre segmente sunt ajustate astfel încât să fie conservat fluxul vizual precum şi coerenţa secvenţei. Mai mult, continuitatea sunetului este asigurată prin evitarea întreruperilor ce pot surveni, de exemplu, în mijlocul propoziţiilor.

În funcţie de modul în care este realizată integrarea audio-vizuală, putem distinge două categorii de rezumate dinamice, şi anume: *rezumatele sincrone* si *rezumatele asincrone* [Truong 07].

În rezumatele sincrone, informaţia vizuală este sincronizată cu cea audio folosind ca sistem de referinţă axa temporală a secvenţei. Acest tip de rezumat este mai adaptat filmelor, deoarece în momentul vizualizării, sunetul este în corespondenţă directă cu imaginile. În cazul în care segmentele vizuale şi audio au fost generate separat din cele două modalităţi ale secvenţei, integrarea acestora poate fi efectuată pe baza unui set de reguli de fuziune guvernate de operatorii clasici de conjuncţie ("*şi*") şi disjuncţie ("*sau*") [Erol 03]. De exemplu, putem selecţiona pasajele importante ale secvenţei ce sunt conţinute într-un segment audio "*sau*" într-unul vizual.

Pe de altă parte, rezumatele asincrone sunt mai adaptate în cazul rezumării secvenţelor documentare şi de ştiri, deoarece acestea tind să maximizeze acoperirea conţinutului secvenţei. Acest gen de rezumat este de regulă generat în primă fază folosind doar una dintre modalităţile secvenţei, urmând ca celelalte să fie adăugate ulterior. Un exemplu este metoda propusă în [Smith 98] unde rezumatul este extras folosind informaţia audio. Elementele

vizuale sunt adăugate ulterior pe baza unui set de reguli euristice determinate în funcţie de mişcarea obiectelor sau a camerei video, de prezenţa feţelor umane sau a textului încrustat în imagine.

Construcţia propriu-zisă a rezumatului

În ceea ce priveşte *construcţia rezumatului final*, strategia cel mai frecvent folosită constă în agregarea tuturor segmentelor obţinute în etapele anterioare folosind ca referinţă axa temporală. Un caz particular îl constituie rezumatele de tip "movie trailer", ce prezintă succint doar pasajele dinamice şi de acţiune ale secvenţei, în scopul de a capta atenţia utilizatorului. În acest caz, ordinea temporală nu este întotdeauna respectată.

5.3 Metodele de evaluare a rezumatelor

Pe lângă dificultăţile de analiză implicate de procesul de generare a rezumatelor de conţinut, o problemă la fel de importantă o constituie *evaluarea calităţii acestora*.

Întrebările care se pun sunt: "*Este rezumatul obţinut pertinent pentru conţinutul secvenţei ? Este acesta coerent pentru persoana care îl vizualizează ?*". În realitate, pentru o aceeaşi secvenţă se pot genera o multitudine de rezumate care să fie corecte din acest punct de vedere. Faptul că rezumatul a fost generat în conformitate cu anumite criterii obiective, inspirate de percepţia umană, nu asigură întotdeauna o coerenţă vizuală a acestuia.

În momentul de faţă, în literatura de specialitate, *nu există o metodologie validată* şi general valabilă de evaluare a calităţii rezumatelor automate de conţinut, cum sunt de exemplu erorile de tip "precision" şi "recall" pentru evaluarea detecţiei de caracteristici (vezi Secţiunea 2.2.4). Tendinţa este de a testa metodele propuse folosind strategii de evaluare mai mult sau mai puţin particulare, lucru ce nu încurajează un studiu comparativ şi competitiv al rezultatelor obţinute cu diverse metode.

Acest lucru se datorează în principal *dificultăţii* şi a *subiectivităţii* construirii de rezumate de referinţă, generate manual pe baza expertizei umane, care să constituie ceea ce numim "realitate de teren" sau "groundtruth"[19] pentru procesul de evaluare. Dispunând de aceste date de referinţă, rezu-

[19]termenul de "groundtruth" îşi are originea în domeniul cartografiei şi implică procesul de colectare de informaţii despre un anumit fenomen, prin observarea practică pe teren a acestuia. Datele obţinute constituie "realitatea de teren" folosită pentru calibrarea, validarea şi interpretarea observaţiilor sau a măsurătorilor de la distanţă a fenomenului în cauză sau a altor fenomene similare.

matele obţinute urmează să fie comparate cu acestea pentru a judeca calitatea lor. Chiar şi în cazul în care constituirea unui "groundtruth" ar fi posibilă, apare ca problemă subiectivitatea comparării a două rezumate, deoarece în multe situaţii chiar şi pentru o persoană este dificil să decidă dacă un anumit rezumat este mai bun decât un altul.

Totuşi în ciuda acestor dificultăţi de evaluare, în literatura de specialitate putem evidenţia o serie de strategii folosite pentru evaluarea atât subiectivă cât şi obiectivă a rezumatelor automate. Acestea sunt [Truong 07]:

- *analiza descriptivă* a rezultatelor obţinute,
- folosirea unei anumite *măsuri matematice* pentru evaluarea obiectivă,
- *testele de evaluare* efectuate de utilizatori.

5.3.1 Analiza descriptivă a rezultatului

Metoda cea mai simplă de evaluare, ce nu necesită compararea cu alte metode, o constituie analiza descriptivă a rezultatului. Rezumatele obţinute cu o anumită metodă sunt analizate manual pentru anumite secvenţe reprezentative, fiind justificate astfel avantajele şi performanţele metodei propuse.

O astfel de evaluare este propusă în [Yu 04] unde rezultatele obţinute sunt exemplificate şi analizate manual din punct de vedere al pertinenţei acestora. Alte abordări constau în explicarea şi ilustrarea, într-o manieră descriptivă, a avantajelor şi a superiorităţii metodei propuse în raport cu alte strategii de bază, cum ar fi alegerea aleatoare a imaginilor cheie sau eşantionarea uniformă a secvenţei [Vermaak 02] (vezi Figura 5.4, se observă eficienţa rezumatului propus în raport cu rezumatul de comparaţie prin prezenţa unui număr redus de imagini redundante).

Din punct de vedere global, o astfel de strategie de evaluare este foarte subiectivă, deoarece nu există dovezi că metoda propusă este eficientă şi aplicată altor secvenţe decât cele testate şi exemplificate. Evaluarea experimentală propusă în acest caz este insuficientă pentru o evaluare globală. Mai mult, acest tip de abordare este dificil de utilizat la evaluarea rezumatelor dinamice, deoarece volumul de date în acest caz este mult prea ridicat pentru a permite o analiză descriptivă.

5.3.2 Utilizarea unei măsuri matematice

Pentru evaluarea rezumatelor statice, măsura matematică folosită este de regulă o funcţie de fidelitate ce este estimată atât pentru imaginile cheie din rezumat cât şi pentru imaginile din secvenţă.

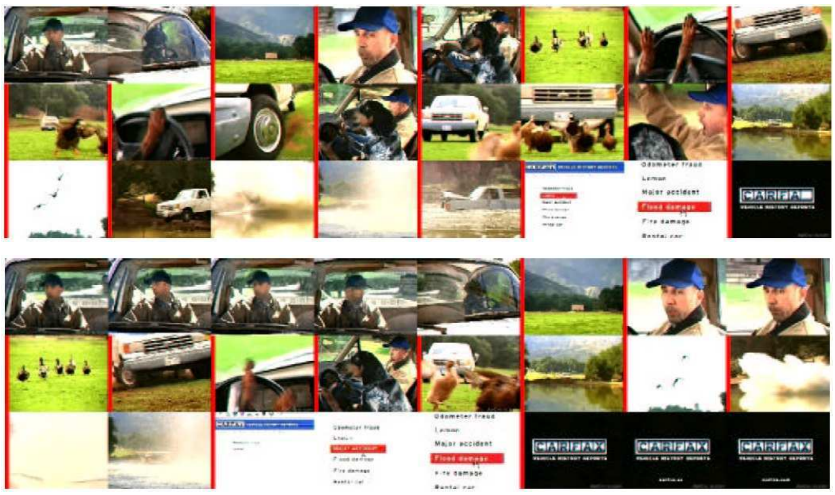


Figura 5.4: Exemplu de analiză descriptivă şi comparare a rezumatului obţinut cu criteriul "Bayes Information Criterion" propus în [Vermaak 02] (prima imagine), cu rezumatul obţinut prin eşantionarea uniformă a secvenţei (a doua imagine, liniile roşii marchează schimbările de plan).

Această strategie permite compararea rezultatelor obţinute cu mai multe metode. Pe de altă parte, similar evaluării prin descrierea rezultatului, această măsură nu oferă certitudinea că este o măsură pertinentă din punct de vedere al modului de evaluare uman. Ca exemplu, putem menţiona metoda propusă în [Liu 04] ce foloseşte ca măsură de evaluare distanţa Hausdorff[20] aplicată între erorile de reconstrucţie SRE[21] obţinute cu metodele de rezumare evaluate. Un alt exemplu este metoda propusă în [Liu 02c] ce foloseşte pentru evaluare conceptul de "imagine cheie bine distribuită". Aceasta este definită de autori ca fiind o imagine cheie ce nu este redundantă şi care nu aparţine unei tranziţii video. Evaluarea calităţii rezumatelor este realizată prin reprezentarea grafică a numărului de imagini "bine distribuite" în funcţie de numărul total al imaginilor cheie din rezumat.

În cazul particular al rezumatelor dinamice de tip "video highlight", ce sunt determinate pe baza anumitor pasaje ale secvenţei ce conţin evenimente de interes, evaluarea calităţii este relativ mai uşor de realizat deoarce se poate construi echivalentul unei "realităţi de teren". Aceasta este constituită prin

[20] numită după matematicianul Felix Hausdorff, distanţa ce îi poartă numele este definită între două mulţimi A şi B ca fiind distanţa maximă a mulţimi A faţă de cel mai apropiat punct din mulţimea B, astfel: $d_H(A,B) = max|_{a\in A}\{min|_{b\in B}\{d(a,b)\}\}$, unde $d()$ reprezintă o anumită metrică.

[21] vezi explicaţia de la pagina 159.

localizarea manuală în secvenţă a evenimentelor de interes, iar evaluarea constă în verificarea prezenţei acestora în rezumat [Xiong 03] [Ariki 03]. Ca măsură cantitativă a evaluării se pot folosi în acest caz erorile de tip "precision" şi "recall" (vezi Secţiunea 2.2.4) folosite la scară largă în toate metodele de detecţie.

Alte abordări ale procesului de evaluare încearcă să îmbunătăţească obiectivitatea acestuia prin folosirea unei "realităţi de teren". Aceasta este constituită de regulă de o serie de rezumate de referinţă, ce au fost create manual de specialişti pentru un set redus de secvenţe ce vor fi folosite la evaluare. De exemplu, în [He 99] evaluarea metodei propuse de rezumare a secvenţelor de conferinţe este realizată prin compararea rezultatelor cu o serie de rezumate create manual de autori, sau în [Miura 03] unde rezumatele propuse ale programelor televizate dedicate gastronomiei sunt evaluate folosind ca referinţă comentarile furnizate de producător. Totuşi, cu toate că putem vorbi de existenţa unei "realităţi de teren", aceasta nu poate fi unic determinată, fapt ce poate conduce la o evaluare total eronată. Ma mult, evaluarea devine subiectivă în momentul comparării rezumatului obţinut, cu rezumatul sau rezumatele de referinţă. După cum am menţionat şi în partea introductivă, compararea a două rezumate este dificilă chiar şi pentru operatorul uman.

Din punct de vedere global, metodele de evaluare ce folosesc măsuri matematice sunt mai obiective decât metodele ce descriu rezultatele obţinute. Acestea încearcă constituirea unei "realităţi de teren" prin expertiza secvenţelor folosite, lucru ce permite o evaluare competitivă raportată la alte metode de rezumare. Evaluarea calităţii este realizată prin compararea pe baza unei anumite măsuri de distanţă a rezumatului obţinut, cu rezumatele de referinţă din "realitatea de teren". De notat este faptul că în această abordare, analiza percepţiei umane nu intervine în procesul propriu-zis de evaluare.

5.3.3 Testele de evaluare

În cazul *testelor de evaluare*, sarcina de a decide asupra calităţii rezumatului îi revine operatorului uman.

Un test de evaluare implică de regulă un grup de persoane, specialiste sau non specialiste în domeniu, ce sunt desemnate pentru a vizualiza rezumatele obţinute cu o anumită metodă. Opinia personală a acestora asupra conţinutului şi a calităţii rezumatelor este de regulă analizată pe baza răspunsurilor la un anumit chestionar. Chestionarul în cauză este constituit în funcţie de cerinţele şi de constrângerile iniţiale impuse tipului de rezumat evaluat. În ciuda subiectivităţii aparente a acestei strategii de evaluare, testele de evaluare se dovedesc a fi cele mai apropiate de realitate, deoa-

rece evaluarea este realizată de însuşi "consumatorul produsului", şi anume utilizatorul.

Din punct de vedere practic, organizarea unei astfel de campanii de evaluare este dificilă în primul rând ca logistică (pregătirea filmelor, pregătirea protocolului de evaluare, strângerea unui număr suficient de evaluatori, etc.), dar şi din punct de vedere al timpului necesar vizionării de către evaluatori a rezumatelor propuse, cât şi a secvenţelor originale. Vizualizarea secvenţelor originale este foarte importantă pentru procesul de evaluare deoarece doar pe baza acestora evaluatorul poate să-şi facă o idee asupra conţinutului real al secvenţei.

Ca exemple de astfel de strategii de evaluare, putem menţiona în cazul rezumatelor statice metoda propusă în [Dufaux 00] unde fiecare imagine cheie a rezumatului este clasată de utilizator folosind trei niveluri de apreciere, şi anume: "bună", "corectă" sau "slabă". Pe baza acestor aprecieri este determinat un scor global pentru întregul rezumat. O abordare similară, dar care implică un test mai complex, este propusă în [Liu 03]. Imaginile cheie sunt analizate în acest caz în contextul planelor video din care provin. Pe baza acestora, utilizatorii atribuie un scor de satisfacţie cuantificat pe trei niveluri de apreciere: "bine", "acceptabil" sau "nesatisfăcător".

În cazul evaluării rezumatelor dinamice, strategia cel mai frecvent folosită constă în a cere utilizatorului direct opinia acestuia cu privire la calitatea globală a rezumatului propus. În [Ionescu 06b], o astfel de abordare este folosită pentru evaluarea rezumatelor de tip "movie trailer" în cazul filmelor artistice de animaţie. Evaluarea este efectuată în acest caz pe baza scorurilor atribuite răspunsurilor la întrebările unui chestionar inspirat de teoria sondajelor. O primă întrebare este relativă la calitatea rezumatului, şi anume: "Consideraţi că rezumatul "movie trailer" propus conţine părţile cele mai atractive ale filmului?". Răspunsului îi este atribuit un scor numeric cu valori de la 1 la 10, având semnificaţia: 1, 2="deloc", 3, 4="foarte puţine", 5, 6="unele dintre acestea", 7, 8="aproape toate" şi respectiv 9, 10="toate". O a doua întrebare este relativă la durata rezumatului, şi anume: "Cum vi se pare durata rezumatului propus?". Scorul atribuit răspunsurilor la această întrebare variază de la 0 la 4 şi are semnificaţia următoare: 0="foarte scurtă", 1="scurtă", 2="corectă", 3="lungă" şi respectiv 4="foarte lungă". Evaluarea calităţii rezumatului se face pe baza scorului mediu şi a varianţei acestuia obţinute pentru ansamblul evaluatorilor. Alte abordări mai elaborate încearcă să aprecieze în ce măsură rezumatele propuse pot fi utile pentru o serie de aplicaţii practice, precum navigarea şi căutarea într-o bază de filme [Ngo 03]. De asemenea, o altă strategie constă în încercarea de a evalua calitatea unui rezumat pe baza capacităţii utilizatorului de a identifica conţinutul original al secvenţei, pornind de la rezumatul propus [Erol 03].

Pe lângă avantajele prezentate din punct de vedere al relevanţei procesului de evaluare, testele de evaluare prezintă şi o serie de limitări [Ionescu 07a]. Acestea pot fi sintetizate cu următoarele:

- *timpul necesar* evaluării este ridicat, mai ales în cazul rezumatelor dinamice,
- *dificultatea realizării unui test comparativ* pentru mai multe metode de rezumare, datorată subiectivităţii evaluatorului pus în situaţia de a decide între două rezumate diferite,
- *subiectivitatea* modului de percepţie a fiecărui evaluator, ce este strict dependentă de formarea profesională a acestuia (de exemplu, un artist va avea o percepţie diferită faţă de un inginer),
- *dificultatea de vizualizare* a rezumatelor în imagini: modul de prezentare al acestora influenţează calitatea evaluării. De exemplu, prezentarea imaginilor cheie sub forma unui "slideshow" are ca efect crearea artificială a senzaţiei de secvenţă sacadată, pe când prezentarea acestora sub forma unei planşe are ca rezultat tendinţa de a neglija evoluţia temporală a secvenţei.

5.4 Concluzii

În acest capitol am discutat problematica rezumării automate a conţinutului secvenţelor de imagini în contextul general al indexării după conţinut. În funcţie de tipul rezumatului furnizat, tehnicile de rezumare de conţinut existente se împart în două mari categorii, şi anume: metode de *rezumare în imagini* şi metode de *rezumare în mişcare*.

Metodele de rezumare în imagini generează ceea ce numim rezumate statice. Un rezumat static reprezintă o colecţie de imagini statice, numite şi imagini cheie, ce sunt considerate ca fiind reprezentative pentru conţinutul secvenţei. Tehnicile existente variază de la metode simple ce aleg imaginile cheie, uniform, la nivel de plan video, până la metode adaptive ce selectează imaginile cheie în funcţie de nivelul de activitate al secvenţei. Cu toate că acest tip de reprezentare este foarte aproximativă, se dovedeşte totuşi a fi foarte eficientă în multe situaţii. Rezumatul în imagini are în general o complexitate redusă de calcul şi furnizează utilizatorului o imagine de ansamblu asupra conţinutului vizual al secvenţei. Rezumatul static este util în cazul navigării în baza de secvenţe, când utilizatorul poate fi informat instantaneu, prin simpla vizualizare a câtorva imagini cheie, asupra conţinutului secvenţei.

Mai mult, rezumatul în imagini permite reducerea redundanţei vizuale, necesară în multe metode de prelucrare specifice analizei video.

Pe de altă parte, metodele de rezumare în mişcare generează rezumate dinamice. Acestea sunt la bază o colecţie de pasaje ale secvenţei, ce sunt asamblate şi sincronizate pentru a forma ele însele o secvenţă. Alegerea pasajelor este realizată în acest caz în funcţie de tipul informaţiei ce va fi redată de rezumat. De exemplu, un rezumat de tip "movie trailer" va prezenta succint doar momentele de acţiune ale secvenţei, în timp ce un rezumat dinamic de tip "movie highlight" va prezenta doar anumite evenimente de interes din secvenţă. Spre deosebire de rezumatele statice, rezumatele dinamice au mai mult sens din punct de vedere al vizualizării, deoarece redau informaţia de mişcare a secvenţei. Mai mult, în funcţie de caz, acestea pot prezenta şi informaţia audio. Din această cauză, un dezavantaj al rezumatelor dinamice este dat de complexitatea de calcul superioară metodelor de extragere a rezumatelor statice.

De notat este faptul că cele două categorii de rezumate *nu sunt concurenţiale*, sau cu alte cuvinte, metodele dezvoltate nu au ca scop dovedirea superiorităţii uneia dintre categorii. Ambele tipuri de rezumate sunt necesare indexării după conţinut a secvenţelor de imagini. Rezumatele în imagini sunt uşor de generat şi vizualizat în detrimentul conţinutului de mişcare, în timp ce rezumatele dinamice sunt mai greu de generat şi necesită un timp mai ridicat pentru vizualizare în avantajul furnizării informaţiei dinamice şi audio a secvenţei.

O problemă delicată care apare este *evaluarea calităţii* unui rezumat. Aceasta este în cele mai multe cazuri subiectivă şi depinde în principal de modul de percepţie al fiecăruia dintre noi, dar şi de tipul conţinutului redat de rezumat. De exemplu, evaluarea unui rezumat ce prezintă conţinutul global al secvenţei nu poate fi realizată în acelaşi mod ca pentru un rezumat ce redă doar anumite evenimente importante ale acesteia, sau, evaluarea unui rezumat static nu poate utiliza aceleaşi criterii ca evaluarea unui rezumat dinamic, deoarece prezintă informaţii diferite. În acest moment, nu se poate spune că există o metodologie standardizată de evaluare a rezumatelor, existând mai multe metodologii mai mult sau mai puţin relevante de evaluare. Dintre acestea, cele mai pertinente se dovedesc a fi testele de evaluare, deoarece acestea implică în procesul de evaluare chiar utilizatorul căruia îi sunt destinate.

CAPITOLUL 6

Formalizarea fuzzy

Rezumat: *Descrierea matematică, cu valori numerice, a proprietăţilor de interes a datelor multimedia analizate, nu reprezintă o modalitate de descriere accesibilă publicului larg, ci mai degrabă specialiştilor din domeniu. Mai mult, aceste descrieri nu au nici un sens din punct de vedere al percepţiei umane. Tendinţa actuală a metodelor de analiză existente este tocmai de a dezvolta metode capabile să înţeleagă conţinutul datelor multimedia într-un mod similar modului în care percepem lumea înconjurătoare. Una dintre soluţiile larg acceptate o reprezintă formalizarea fuzzy pe baza conceptului de incertitudine. În acest capitol vom prezenta modul în care informaţiile de nivel scăzut, numerice, pot fi convertite în concepte semantice folosind inferenţa fuzzy, precum şi impactul acesteia asupra metodelor existente de analiză şi prelucrare a secvenţelor de imagini.*

După cum am menţionat în primul capitol al acestei lucrări, direcţia de studiu ce face obiectul numeroaselor cercetări actuale din domeniul sistemelor de indexare după conţinut, o constituie *analiza semantică* a percepţiei datelor.

Această direcţie, putem spune, aflată încă la începuturi, are ca obiectiv dezvoltarea de metode şi algoritmi de modelare a sistemului de percepţie uman, în încercarea de înţelegere automată a sensului datelor. Furnizarea unui nivel de descriere semantic, conferă sistemelor de prelucrare şi analiză,

în primul rând, mult mai mult sens, acestea devenind mai uşor accesibile publicului larg. De exemplu, descrierea distribuţiei de culoare a unei imagini cu parametri de nivel scăzut, precum $I = \{P_{(255,0,0)} = 50\%, P_{(0,0,255)} = 50\%\}$ (unde $P_{(c)}$ reprezintă procentul de apariţie al unei culori c), va fi accesibilă doar unei persoane avizate în domeniu, în timp ce descrierea la un nivel semantic perceptual, precum "imaginea contrastează culoarea roşie şi albastră", va permite oricăruia dintre noi să-şi creeze o imagine mentală a conţinutul de culoare al imagini în cauză.

În sistemele de indexare după conţinut a secvenţelor de imagini, descrierile semantice îşi găsesc un spectru de aplicabilitate mai larg [Ionescu 07a], astfel:

- permit *simplificarea navigării în baza de date*: acestea vin cu informaţii suplimentare despre conţinutul secvenţelor, facilitând astfel înţelegerea rapidă a acestuia de către utilizator. De exemplu, pe lângă rezumatele automate de conţinut, o secvenţă poate fi acompaniată şi de informaţii textuale, cum ar fi: genul acesteia (acţiune, dramă, documentar), tipul conţinutului (natură, oraş, studio), etc.

- permit *simplificarea căutării în baza*: fiind exprimate sub formă textuală, pot fi folosite ca indici de căutare în baza de date. Astfel, utilizatorul îşi poate formula cererea de căutare într-un limbaj natural, apropiat de limbajul uman, de exemplu "caută secvenţele de gol din meciul echipelor X şi Y" sau "caută filmele de ficţiune".

- constituie *un ajutor pentru specialişti*: descrierile semantice de conţinut pot însoţi descrierile de nivel scăzut pentru a furniza un pachet de informaţii complet cu privire la tehnicile folosite în secvenţă: structura temporală, conţinut de mişcare, conţinut de culoare, etc.

După cum se poate observa din cele enunţate anterior, modalitatea cea mai expresivă de exprimare a sensului semantic al datelor constă în *reprezentarea textuală a acestuia*. Dintre metodele de asociere de descrieri textuale de conţinut datelor numerice, o pondere importantă o au metodele ce se folosesc de conceptul de incertitudine. Una dintre acestea o reprezintă *formalizarea fuzzy* a conceptelor semantice textuale.

6.1 Introducerea conceptului de incertitudine

Printre schimbările spectaculoase ale paradigmelor existente în diversele domenii ştiinţifice, una dintre cele mai importante o constituie fundamentarea şi dezvoltarea *conceptului de incertitudine* a datelor.

Această schimbare majoră a modului de percepţie ştiinţific s-a materializat prin tranziţia modului tradiţional de gândire, care insista asupra faptului că noţiunea de incertitudine este o proprietate a datelor ce nu se doreşte să apară şi care trebuia evitată pe cât posibil, spre o viziune alternativă ce tolera tocmai această noţiune, fiind considerată de această dată inevitabilă şi foarte utilă pentru analiză.

În general, în momentul proiectării unui anumit sistem de analiză, întrebarea care se pune este: "*Cum trebuie gerat sistemul şi problemele asociate acestuia în cazul în care complexitatea proceselor ce trebuiesc modelizate depăşeşte cu mult posibilităţile noastre de prelucrare?*". Cu alte cuvinte, volumul informaţional disponibil este foarte ridicat pentru a putea fi controlat în totalitate iar înţelegerea proceselor este limitată. Soluţia în acest caz constă tocmai în introducerea noţiunii de incertitudine pentru situaţiile în care soluţia ce trebuie adoptată nu este deloc evidentă, ci mai degrabă incertă.

În momentul construcţiei unui anumit model, se încearcă întotdeauna să se maximizeze utilitatea acestuia. Acest obiectiv este strâns legat de relaţiile ce pot exista între cele trei categorii de caracteristici cheie ale unui model, şi anume: *complexitatea*, *credibilitatea* şi *incertitudinea* acestuia. Aceste relaţii nu sunt întotdeauna înţelese în totalitate. Ştim doar că incertitudinea, fie că este predictivă, prescriptivă, etc., joacă rolul esenţial pentru efortul de maximizare a utilităţii sistemului.

Totuşi în cele mai multe situaţii, dar nu întotdeauna, incertitudinea nu reprezintă un punct forte dacă aceasta este considerată independent de alţi parametri. Incertitudinea devine o informaţie preţioasă a sistemului dacă este analizată în raport cu alte caracteristici ale acestuia. În general, cu toate că aparent este paradoxal, cu cât se adaugă mai multă incertitudine în modelarea sistemului, cu atât complexitatea acestuia este redusă şi în efect credibilitatea modelului creşte [Klir 95].

În concluzie, conceptul de incertitudine este un instrument important pentru modelarea unui anumit sistem sau pentru soluţionarea unei anumite probleme. Acesta permite obţinerea de caracteristici "avantajoase" pentru modelul vizat, caracteristici ce vor conduce ulterior la maximizarea utilităţii acestuia relativ la scopul pentru care a fost creat.

Conceptul de incertitudine a fost materializat pentru prima dată în lucrările publicate de Lotfi A. Zadeh [Zadeh 65] (anticipat de filozoful Max Black în 1937). Acesta propunea o nouă teorie bazată pe reprezentarea datelor cu mulţimi fuzzy. Mulţimile fuzzy sunt mulţimi pentru care frontierele dintre date nu sunt exacte. Apartenenţa datelor, în acest caz, la o astfel de mulţime, nu mai este o problemă de confirmare sau negare, ci o problemă de *grad de apartenenţă*.

Dacă teoria probabilităţilor este fondată pe definirea a două valori logice de adevăr, şi anume *Adevărat* (1) şi *Fals* (0), în logica fuzzy, gradul de adevăr este formulat în felul următor:

dacă A reprezintă o mulţime fuzzy iar x este un obiect de interes, atunci propoziţia "x este inclus în A" nu este obligatoriu să fie Adevărată sau Falsă, lucru impus de logica booleană, ci aceasta poate fi adevărată într-un anumit grad.

Acest grad de adevăr este exprimat de regulă ca o valoare cuprinsă în intervalul $[0; 1]$, unde limitele acestuia reprezintă negaţia totală (limita inferioară, valoare de adevăr 0) şi respectiv, afirmaţia totală (limita superioară, valoare de adevăr 1).

Capacitatea mulţimilor fuzzy de a exprima tranziţia graduală între apartenenţa totală şi non apartenenţă, şi vice-versa, îşi găseşte o vastă utilitate în marea majoritate a domeniilor existente. Mulţimile fuzzy nu numai că propun o reprezentare discriminantă şi plină de sens a conceptului de incertitudine, ci şi o reprezentare pertinentă a conceptelor vagi ce sunt exprimate într-un *limbaj natural*.

Pentru a înţelege avantajul folosirii mulţimilor fuzzy la descrierea proprietăţilor anumitor procese, vom considera exemplul următor [Klir 95]: în loc să descriem prognoza meteo a zilei curente specificând procentajul exact de acoperire al cerului cu nori, P_{nori}, putem adopta o soluţie mai eficientă spunând că ziua va fi, fie "însorită", fie "cu un cer acoperit".

Această descriere este o descriere vagă şi puţin exactă, dar în cele mai multe cazuri este mult mai utilă decât prima modalitate de descriere. Sensul termenului de "însorit" nu este în totalitate arbitrar. O acoperire cu nori în procent de $P_{nori} = 100\%$ indică faptul că ziua nu este însorită, dar acelaşi lucru este valabil şi pentru $P_{nori} = 80\%$. Astfel, pentru a desemna o "zi însorită" vom considera o serie de valori intermediare pentru P_{nori}, de exemplu: $P_{nori} \in [10\%; 20\%]$. Problema care apare este cum alegem aceste frontiere? Dacă considerăm că un procent de acoperire de mai puţin de 25% corespunde unei "zi însorite", atunci o acoperire de 26% corespunde sau nu aceluiaşi caz? Este evident inacceptabil în această modalitate de descriere binară ca o singură valoare a lui P_{nori} să facă diferenţa între două concepte opuse: "zi însorită" şi respectiv "zi cu cer acoperit".

Pentru a soluţiona acest conflict, termenul de "zi însorită" necesită un anumit grad de incertitudine ce va fi obţinut prin introducerea unei tranziţii graduale între valorile lui P_{nori}, folosite pe de-o parte pentru a desemna noţiunea de "zi însorită" şi respectiv pentru conceptul opus. Acesta constituie exact principiul de bază al logicii fuzzy ce reprezintă o generalizare a logicii booleene.

6.2 Logica booleană şi logica fuzzy

În teoria clasică a mulţimilor, ce este bazată pe *logica booleană*, apartenenţa unui obiect la o anumită mulţime este exprimată în felul următor: dacă U reprezintă universul de discuţie iar A reprezintă o submulţime a acestuia, $A \subset U$, atunci pentru fiecare obiect x din U, apartenenţa acestuia la mulţimea A este definită de funcţia specifică α_A:

$$\alpha_A(x) = \begin{cases} 1 & \text{dacă } x \in A \\ 0 & \text{dacă } x \notin A \end{cases} \tag{6.1}$$

Astfel, în acest caz sunt posibile doar două situaţii, şi anume: fie x aparţine lui A, fie acesta nu aparţine.

Pe de altă parte, în *logica fuzzy* conceptul de apartenenţă se bazează pe noţiunea de incertitudine. Dacă presupunem că α_A reprezintă funcţia de apartenenţă fuzzy a unui obiect din universul de discuţie la mulţimea fuzzy A, atunci α_A este dat de relaţia următoare:

$$\alpha_A(x) \in [0; 1] \tag{6.2}$$

pentru toate valorile lui $x \in U$. Funcţia α_A exprimă în acest caz *gradul de apartenenţă*, sau cu alte cuvinte, valoarea de adevăr, unde valoarea 1 reprezintă apartenenţa sigură iar valoarea 0 non-apartenenţa sigură.

Pentru a ilustra diferenţa dintre cele două concepte, vom considera exemplul următor: dacă presupunem că un pacient bolnav de gripă prezintă o temperatura T a corpului ridicată, atunci conceptul de *"temperatură ridicată"* poate fi asociat valorilor măsurate ale lui T folosind, fie modelul clasic, fie modelul fuzzy (vezi Figura 6.1).

În primul caz al logicii clasice, decizia de apartenenţă este una netă (tranşantă), astfel, fie pacientul are o "temperatură ridicată" dacă $T > 39$ de grade (valoare de adevăr $\alpha(T) = 1$), fie nu are o "temperatură ridicată" în caz contrar ($\alpha(T) = 0$). În logica fuzzy, pacientul are o "temperatură ridicată" cu o valoare de adevăr 1 doar dacă $T > 41$ de grade. Pentru $T \in [37; 41]$ funcţia de apartenenţă este o funcţie liniară ce ia valori reale între 0 şi 1. Astfel, în acest caz, valoarea de adevăr a apartenenţei la conceptul de "temperatură ridicată" va fi una graduală. De exemplu, pentru $T = 39$ de grade putem afirma că pacientul are o "temperatură ridicată" cu o valoare de adevăr de numai 0.5 (vezi Figura 6.1).

Decizia netă nu este eficientă în acest caz deoarece pacientul va fi diagnosticat ca având o temperatură ridicată doar în cazul în care valoarea temperaturii T este deja foarte ridicată (de exemplu, $T > 39$). Mai mult, este cu totul artificial să spunem că pacientul are o temperatura ridicată

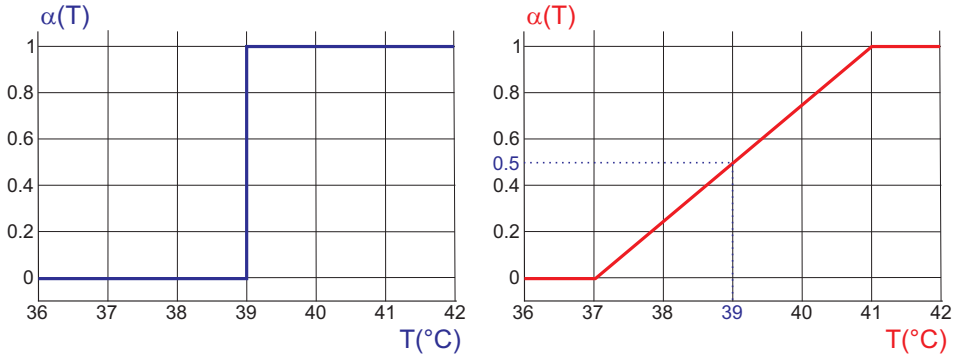


Figura 6.1: Reprezentarea funcţiei de apartenenţă $\alpha(T)$ a conceptului de *"temperatură ridicată"* în logica booleană (prima imagine) şi respectiv în logica fuzzy (a doua imagine).

daca $T > 39$ şi în cazul în care $T = 38.8$ acestă afirmaţie să nu fie valabilă. Avantajul reprezentării fuzzy este evident în cazul în care asociem conceptul de "febră ridicată" cu diagnosticul de gripă. Astfel, în logica clasică deducem că dacă pacientul nu are o "temperatură ridicată" atunci pacientul nu este bolnav de gripă, afirmaţie care nu este întotdeauna adevărată. Pe de altă parte, în logica fuzzy vom deduce că dacă pacientul prezintă o "temperatură ridicată" cu o valoare de adevăr $x \in [0; 1]$, atunci pacientul este bolnav de gripă cu o valoare de adevăr $\tilde{x} \in [0; 1]$, ce va fi calculată în funcţie de x.

Modul în care funcţia de apartenenţă fuzzy $\alpha()$ ia valori în intervalul $[0; 1]$ este dependent de aplicaţia vizată. Acest lucru reprezintă un avantaj important al acestui tip de abordare şi anume de a fi adaptabil la contexte diferite. De exemplu, definirea conceptului de "temperatură ridicată", nu poate fi realizată în acelaşi fel pentru două contexte diferite, precum prognoza vremii şi măsurarea temperaturii unui reactor nuclear. Dacă o temperatură de 40 de grade poate fi considerată ca fiind ridicată pentru condiţiile meteo, nu putem spune acelaşi lucru şi pentru reactorul nuclear. Pentru aceasta, cu toate că conceptul va fi formulat textual în acelaşi fel, funcţiile de apartenenţă $\alpha()$ vor avea variaţii diferite, acestea fiind determinate pe baza expertizei fiecărui context în parte.

În Figura 6.2 am ilustrat câteva dintre funcţiile de apartenenţă fuzzy cel mai frecvent folosite. În ciuda diferenţelor importante dintre acestea, putem identifica o serie de proprietăţi constructive comune:

- în general valoarea maximală este limitată la valoarea 1, în timp ce valoarea minimală este 0, ceea ce corespunde celor două grade de adevăr folosite în logica booleană,

- funcţia de apartenenţă nu trebuie să atingă de mai multe ori, în mod alternativ discontinuu, valoarea maximală,

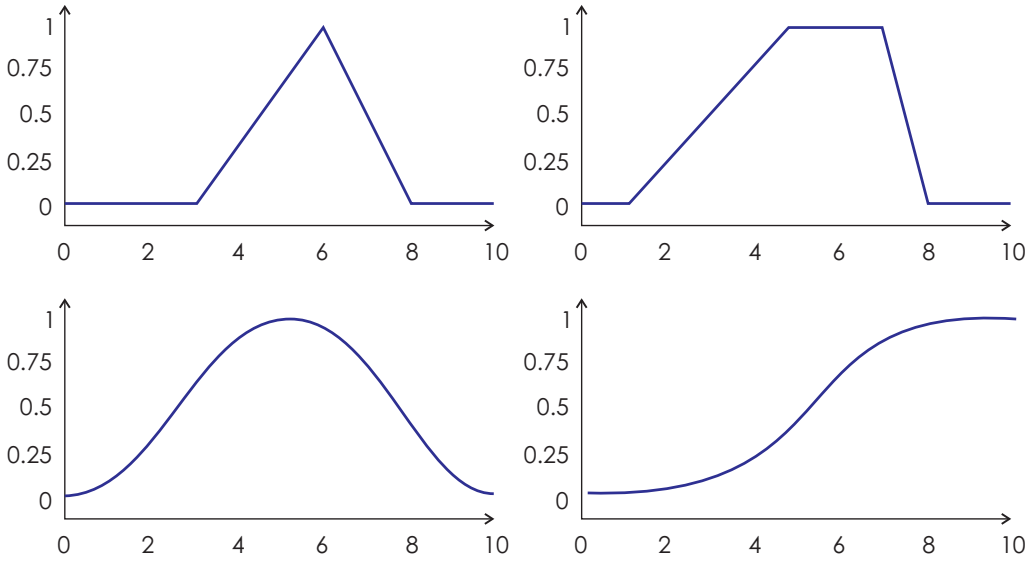


Figura 6.2: Exemple de funcţii fuzzy de apartenenţă: triangulară, trapezoidală, Gausiană şi crescătoare (ordine de la stânga la dreapta, şi de sus în jos, axa oY corespunde valorii de adevăr, în timp ce axa oX reprezintă universul de discurs).

- o funcţie de apartenenţă trebuie să conţină cel puţin o tranziţie între valoarea maximală şi cea minimală (între 1 şi 0).

- în ceea ce priveşte simetria funcţiei, întâlnim două situaţii: fie funcţia are o anumită simetrie faţă de valoarea maximală, fie aceasta este monoton crescătoare sau descrescătoare.

Proprietăţile enumerate mai sus nu sunt întâmplătoare, acestea având rolul de a conserva veridicitatea conceptului reprezentat. Astfel, nu orice funcţie matematică poate fi folosită ca funcţie de apartenenţă fuzzy.

După cum am menţionat anterior, forma şi tipul de variaţie al funcţiei de apartenenţă, şi astfel a valorilor de adevăr asociate valorilor parametrului analizat, sunt dependente de aplicaţie. Modul de variaţie al parametrului vizat influenţează forma funcţiei de apartenenţă. De exemplu, dacă măsurăm o anumită temperatură, funcţia de apartenenţă la conceptul de "temperatură medie" va fi cel mai probabil o funcţie simetrică faţă de valoarea maximală (vezi funcţia Gausiană din Figura 6.2). Acest lucru se datorează faptului că

acest concept implică ca parametrul măsurat, şi anume temperatura T, să ia valori la mijlocul gamei, când temperatura va fi considerată ca fiind medie cu o valoare de adevăr de 1. Dacă în acest caz s-ar folosi, de exemplu o funcţie monotonă crescătoare, atunci toate temperaturile ridicate vor fi catalogate ca fiind medii cu o valoare de adevăr de 1, afirmaţie ce este complet eronată. De asemenea, folosirea unei funcţii nesimetrice nu îşi are sensul în acest caz deoarece felul în care valoarea de adevăr descreşte pentru valorile de temperatură scăzute şi respectiv ridicate nu are nici un motiv să fie diferit, cele două variaţii fiind, putem spune, echiprobabile.

În cele mai multe situaţii, funcţia de apartenenţă este definită *experimental* (empiric) pe baza expertizei manuale a modului de variaţie al parametrilor vizaţi. De exemplu, pentru a defini funcţia de apartenenţă la conceptul de "temperatură ridicată" în contextul unui reactor nuclear, avem nevoie de expertiza specialiştilor din domeniu. În funcţie de experienţa acestora în domeniul fizicii nucleare, funcţia de apartenenţă va fi determinată în conformitate cu modul în care temperatura variază în funcţie de procesele fizice din reactor. Generarea automată a funcţiilor de apartenenţă, cu excepţia unor situaţii simplificate, va duce în mod sigur la obţinerea de rezultate eronate. Astfel, formalizarea unui concept folosind logica fuzzy este dependentă de intervenţia expertizei umane.

6.3 Formalizarea pe baza regulilor fuzzy

Procesul de formalizare pe baza regulilor de decizie fuzzy implică crearea de noi concepte prin combinarea a o serie de concepte fuzzy iniţiale. Acesta implică de regulă două etape, şi anume:

- în prima etapă, pornind de la parametrii numerici de nivel scăzut ai sistemului, se definesc o serie de *variabile fuzzy* (*mulţimi fuzzy*) ce vor parametriza sistemul folosind concepte semantice,
- etapa a doua constă în folosirea de operatori logici specifici, ca de exemplu, operatorii de conjuncţie şi disjuncţie logică, ce sunt aplicaţi variabilelor fuzzy definite anterior pentru a lua decizii cu privire la semnificaţia şi la relaţiile existente între acestea. Deciziile sunt luate folosind o bază de reguli fuzzy. Regulile sunt formulate sub forma unor propoziţii de tip "dacă ... atunci" ("if ... then") şi vor fi definite pentru totalitatea variabilelor fuzzy. Această a doua etapă poartă numele de *inferenţă fuzzy*.

În cele ce urmează, vom detalia fiecare dintre aceste două etape specifice formalizării fuzzy.

6.3.1 Variabilele fuzzy

Înaintea caracterizării semantice fuzzy propriu-zise a unui sistem, sunt definite o serie de variabile fuzzy ce vor caracteriza anumite proprietăţi importante ale acestuia. Astfel, fiecărui parametru numeric de nivel scăzut al sistemului îi va fi asociată o anumită variabilă fuzzy. Aceasta este de regulă exprimată sub forma unui concept lingvistic ce poate lua anumite valori textuale.

Logica fuzzy este bazată pe aceste variabile fuzzy, numite şi *variabile lingvistice*, ce au o valoare lingvistică în universul de discurs U. Fiecare valoare lingvistică constituie o submulţime fuzzy a universului de discurs. De exemplu, dacă U este dat de gama de temperaturi de la 0 la 200 de grade Celsius, atunci o posibilă variabilă lingvistică este conceptul de "temperatură", concept ce poate lua valorile lingvistice: "*foarte rece*", "*rece*", "*temperat*", "*cald*" şi respectiv "*foarte cald*".

Corespondenţa între valorile numerice ale parametrilor de nivel scăzut măsuraţi şi conceptele lingvistice asociate este realizată pe baza *funcţiilor de apartenenţă fuzzy*, ce au fost prezentate în secţiunea anterioară (vezi Figura 6.2). Funcţiile de apartenenţă sunt alese în aşa fel încât să redea cât mai bine relaţia dintre variaţia parametrului modelat şi simbolurile textuale asociate.

Pentru a înţelege mai bine mecanismul de definire a variabilelor fuzzy, în cele ce urmează vom ilustra un exemplu concret (sursă [Lescieux 06]). Să presupunem că vrem să evaluăm cât de înaltă este o anumită persoană în funcţie de înălţimea acesteia. Parametrul de nivel scăzut măsurat va fi în acest caz înălţimea H exprimată în metri. Conceptul lingvistic "*înălţimea unei persoane*" va fi astfel asociat parametrului H pe baza unei reprezentări fuzzy. Acest concept va avea, de exemplu, trei valori lingvistice posibile, şi anume: "*înălţime mică*", "*înălţime medie*" şi "*înălţime mare*", valori ce constituie submulţimile fuzzy ale parametrului H. O modalitate de asociere a valorilor parametrului H celor trei simboluri poate fi realizată folosind funcţiile de apartenenţă ilustrate în Figura 6.3.

Pe baza acestei definiţii, putem spune că înălţimea unei persoane este considerată ca fiind: "mică" cu valoarea de adevăr 1 dacă $H < 1.6$ metri, "medie" cu valoarea de adevăr 1 dacă $H = 1.7$ metri şi respectiv "mare" cu o valoare de adevăr 1 dacă $H > 1.8$ metri. Partiţia fuzzy a universului de discurs este astfel determinată de ansamblul funcţiilor de apartenenţă la simboluri, în cazul nostru de cele trei funcţii ilustrate în Figura 6.3 (vezi ultimul grafic).

Pe baza acestei reprezentări putem transforma valorile numerice ale parametrului H din intervalul de interes $[1.5; 1.9]$, într-o reprezentare simbolică, ce este similară modului de percepţie uman. De exemplu, dacă o anumită

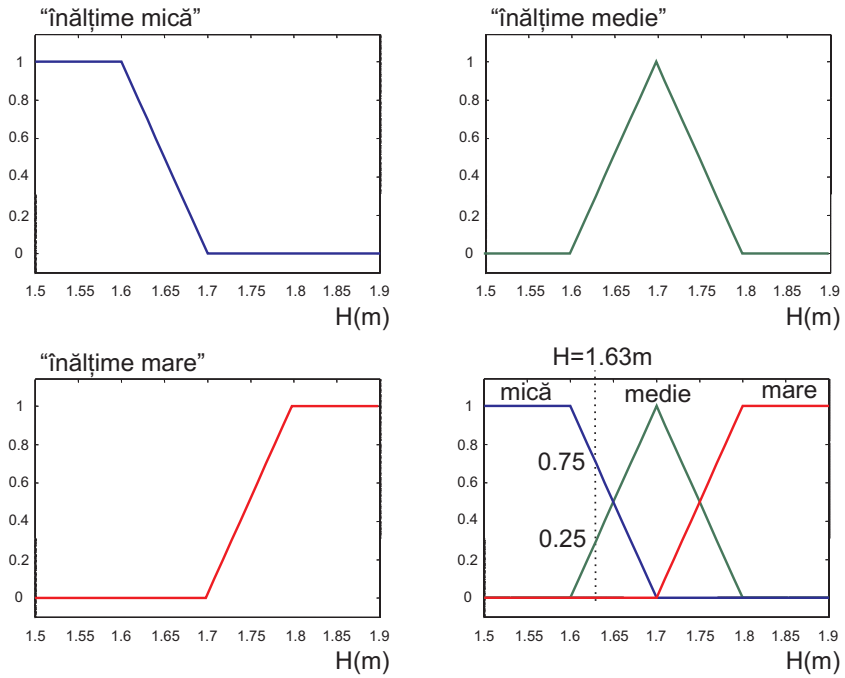


Figura 6.3: Funcţiile de apartenenţă la valorile lingvistice ale conceptului fuzzy de ”înălţimea unei persoane” şi partiţia fuzzy a universului de discurs (ultimul grafic). Axa oY corespunde valorii de adevăr. Exemplu de apartenenţă la fiecare simbol pentru $H = 1.63$ metri.

persoană are înălţimea de 1.63 metri, folosind partiţia fuzzy definită anterior, putem conclude că:

- propoziţia ”persoana are o înălţime mică” are valoarea de adevăr 0.75 (foarte probabil),
- propoziţia ”persoana are o înălţime medie” are valoarea de adevăr 0.25 (puţin probabil),
- propoziţia ”persoana are o înălţime mare” are valoarea de adevăr 0 (imposibil).

Procedând în acelaşi fel pentru toate variabilele numerice din sistem, vom defini totalitatea mulţimilor fuzzy ale acestuia. Pe baza acestora, vom putea analiza mai departe relaţiile vizibile, precum şi ascunse, care există între variabilele fuzzy, ceea ce ne va permite să luăm decizii semantice cu privire la comportamentul sistemului.

6.3.2 Principiul inferenţei fuzzy

Inferenţa fuzzy este definită ca fiind procesul de *luare de decizii* pornind de la o anumită *bază de reguli* de tip "dacă ... atunci". Această bază de reguli este construită, fie pe baza expertizei manuale a relaţiilor existente între diversele mulţimi fuzzy folosite pentru a reprezenta universul de discurs U, sau folosind informaţii "a priori" despre domeniul de aplicaţie. Dacă variabilele fuzzy caracterizează proprietăţile sistemului, inferenţa fuzzy caracterizează din punct de vedere semantic relaţiile care există între diversele variabile ale sistemului.

Proprietăţile unei baze de reguli fuzzy

O bază de reguli este caracterizată de o serie pe proprietăţi. În general, sistemele de inferenţă fuzzy pot fi considerate ca fiind *aproximatori universali* [Wang 92], ceea ce a condus la tendinţa ca metodele de inducere a bazei de reguli să fie în principal ghidate şi evaluate în raport cu acest unic criteriu de performanţă numerică [Guillaume 01].

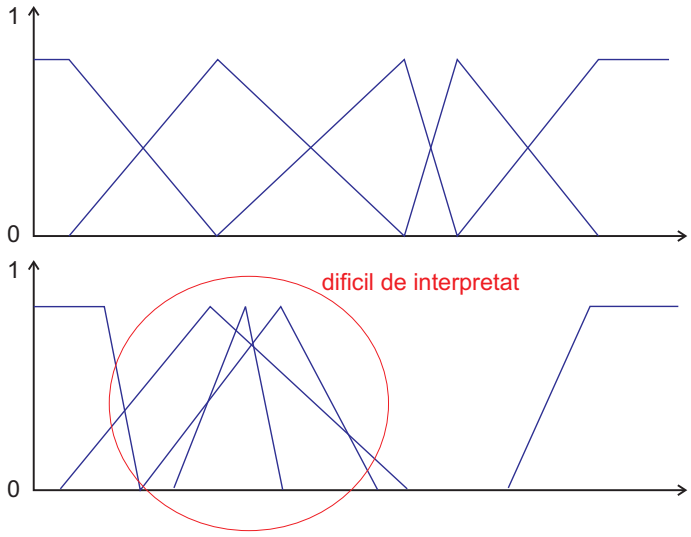


Figura 6.4: Exemple de partiţii fuzzy (axa oX corespunde universului de discurs iar axa oY valorii de adevăr).

Pe de altă parte, obiectivul de "extragere a cunoaşterii" din universul de discurs al parametrilor măsuraţi ai sistemului, impune bazei de reguli o proprietate suplimentară, şi anume ca aceasta *să respecte sensul semantic*. Submulţimile fuzzy trebuie să poată fi interpretate în termeni lingvistici,

lucru ce nu este garantat doar de utilizarea unui formalism fuzzy. În Figura 6.4 am ilustrat tocmai acest lucru. Modul de suprapunere a submulţimilor fuzzy din prima imagine a Figurii 6.4 permite ordonarea acestora şi astfel interpretarea lor cu termeni lingvistici, ca de exemplu, de la "foarte mic", la, "foarte mare". În a doua imagine a Figurii 6.4, partiţia fuzzy este imposibil de a fi interpretată deoarce nu este posibilă etichetarea distinctă a celor trei submulţimi fuzzy centrale (vezi regiunea marcată), acestea fiind mult prea suprapuse.

Pe langă cele două proprietăţi enumerate anterior, o bază de reguli fuzzy mai este caracterizată de trei proprietăţi fundamentale, şi anume: *coerenţă*, *continuitate* şi respectiv *completitudine*, astfel:

- *coerenţa bazei de reguli* este una dintre proprietăţile esenţiale ale acesteia, indiferent de domeniul de aplicaţie al sistemului. O bază de reguli este coerentă, dacă concluziile regulilor ce sunt simultan activate[1] de un anumit exemplu nu sunt contradictorii. O anumită regulă prezintă incoerenţe, dacă distribuţia valorilor de ieşire a exemplelor activate de aceasta este foarte heterogenă. Aceste incoerenţe traduc faptul că regula nu este suficient de specifică pentru a ţine cont de datele de antrenare[2]. Aproximatorii universali divizează spaţiul de intrare prin adăugarea de reguli tocmai pentru a elimina aceste incoerenţe.

- proprietatea de *continuitate a unei baze de reguli* presupune faptul că micile variaţii ale intrării sistemului nu produc variaţii importante ale ieşirii acestuia.

- proprietatea de *completitudine a bazei de reguli* asigură faptul că fiecare dintre valorile posibile ale intrării sistemului activează cel puţin o regulă, astfel încât, în orice situaţie va fi inferată o valoare de ieşire. În cazul aplicaţiilor ce presupun definirea bazei de reguli pe baza expertizei domeniului respectiv, definiţia completitudinii este mai puţin restrictivă. În acest caz, completitudinea nu va fi măsurată raportat la totalitatea valorilor posibile ale intrării sistemului, ci doar în raport cu cele conţinute în setul de exemple de antrenare furnizate de expert.

În funcţie de tipul aplicaţiei vizate, o bază de reguli poate prezenta, fie toate proprietăţile enumerate anterior, fie doar o parte dintre acestea. În Tabelul 6.1 am sintetizat cerinţele pentru o bază de reguli în cazul câtorva dintre cele mai frecvente aplicaţii ale acestora.

[1] prin definiţie, un exemplu activează o regulă, sau o regulă activează un exemplu, dacă valoarea de adevăr a regulii pentru exemplul în cauză, nu este nulă.

[2] antrenarea supervizată constă în inducerea de relaţii între intrarea şi ieşirea unui sistem pe baza unui set de exemple cunoscute "a priori" (vezi şi Secţiunea 7.2).

Aplicaţie	**Semantică**	**Coerenţă**	**Continuitate**	**Completit.**
Comandă	+	+++	+++	+++
Clasificare	+	++	+	+++
Decizie	+++	+++	+++	++
Caracterizare	+++	+++	++	+
Diagnostic	+++	+++	++	+
Simulare	+++	+++	+++	+

Tabelul 6.1: Proprietăţile necesare unei baze de reguli fuzzy pentru diverse tipuri de aplicaţii (fiecărei proprietăţi i-am asociat un scor empiric ce este proporţional cu importanţa acesteia pentru aplicaţia respectivă, sursă [Guillaume 01]).

Astfel, în contextul indexării semantice după conţinut, pentru a ajunge la o caracterizare semantică a conţinutului datelor, baza de reguli trebuie în primul rând să respecte sensul semantic şi să fie coerentă. Aceasta, în mod obligatoriu, trebuie să poată fi capabilă de a furniza rezultate necontradictorii ce pot fi interpretate lingvistic (textual). În ceea ce priveşte proprietatea de completitudine, aceasta nu este strict necesară deoarece regulile vor fi determinate în acest caz folosind expertiza domeniului respectiv.

Relaţiile dintre submulţimile fuzzy

După cum am menţionat anterior, construcţia unei baze de reguli fuzzy implică analiza relaţiilor existente între diversele mulţimi fuzzy (variabile fuzzy) ce modelează sistemul. Relaţiile dintre acestea sunt exprimate folosind *operatori specifici logicii fuzzy*. Dintre aceştia, operatorii cei mai frecvent folosiţi sunt:

- **operatorul de intersecţie**: acesta este un operator de conjuncţie logică, reprezentat cu $\cap$ sau $\wedge$, şi care este denotat de cuvântul cheie *ŞI*,
- **operatorul de reuniune**: acesta este un operator de disjuncţie logică, reprezentat cu $\cup$ sau $\vee$, şi care este denotat de cuvântul cheie SAU,
- **operatorul de complementaritate**: acesta este un operator de negaţie logică şi este denotat de cuvântul cheie NOT.

Folosirea acestor operatori pentru a caracteriza relaţiile dintre variabilele fuzzy ale unui anumit sistem implică "traducerea" acestora cu o serie de ope-

ratori matematici specifici. Aceşti operatori sunt definiţi pe baza noţiunilor de *t-conormă* şi respectiv *t-normă*.

Astfel, prin definiţie, o *t-normă* este o aplicaţie $T(x, y)$ între variabilele x şi y ce satisface următoarele proprietăţi:

- valoarea 1 este element neutru:

$$\forall x \in [0; 1] \quad T(x, 1) = T(1, x) = x \tag{6.3}$$

- este comutativă:

$$T(x, y) = T(y, x) \tag{6.4}$$

- este asociativă:

$$T(x, T(y, z)) = T(T(x, y), z) \tag{6.5}$$

- este monotonă:

$$daca\ x \leq z,\ y \leq w\ atunci\ T(x, y) \leq T(z, w) \tag{6.6}$$

În mod similar, o *t-conormă* este o aplicaţie $S(x, y)$ ce satisface aceleaşi proprietăţi ca *t-norma*, cu excepţia faptului că în acest caz elementul neutru este valoarea 0, astfel:

$$\forall x \in [0; 1] \quad S(x, 0) = x \tag{6.7}$$

Orice t-normă poate fi folosită pentru a defini *operaţia de intersecţie* fuzzy, iar orice t-conormă poate fi folosită pentru a defini *operaţia de reuniune* fuzzy.

În ecuaţiile următoare am prezentat câteva dintre t-normele şi respectiv t-conormele cel mai frecvent folosite, astfel:

- operatorii lui Zadeh:

$$T(x, y) = Min(x, y) \tag{6.8}$$
$$S(x, y) = Max(x, y) \tag{6.9}$$

- operatorii probabilişti:

$$T(x, y) = x \cdot y \tag{6.10}$$
$$S(x, y) = x + y + x \cdot y \tag{6.11}$$

- operatorii Lukasiewicz:

$$T(x, y) = Max(0, x + y - 1) \tag{6.12}$$
$$S(x, y) = Min(1, x + y) \tag{6.13}$$

- operatorii Zadeh condiţionali:

$$T(x,y) = \begin{cases} Min(x,y) & \text{dacă } x = 1, y = 1 \\ 0 & \text{altfel} \end{cases} \tag{6.14}$$

$$S(x,y) = \begin{cases} Max(x,y) & \text{dacă } x \cdot y = 0 \\ 1 & \text{altfel} \end{cases} \tag{6.15}$$

- operatorii Yager ($p > 0$):

$$T(x,y) = 1 - Min([(1-x)^p + (1-y)^p]^{1/p}, 1) \tag{6.16}$$

$$S(x,y) = Min((x^p + y^p)^{1/p}, 1) \tag{6.17}$$

- operatorii Weber ($\lambda > -1$):

$$T(x,y) = Max\left(0, \frac{x + y - 1 + \lambda \cdot x \cdot y}{1 + \lambda}\right) \tag{6.18}$$

$$S(x,y) = Min(x + y + \lambda \cdot x \cdot y, 1) \tag{6.19}$$

- operatorii Hamacher ($\gamma > 0$):

$$T(x,y) = \frac{x \cdot y}{\gamma + (1-\gamma) \cdot (x + y - x \cdot y)} \tag{6.20}$$

$$S(x,y) = \frac{x + y - x \cdot y - (1-\gamma) \cdot x \cdot y}{1 - (1-\gamma) \cdot x \cdot y} \tag{6.21}$$

În cele ce urmează, folosind exemplul prezentat la pagina 185, unde conceptul fuzzy de "*înălţimea unei persoane*" era asociat parametrului măsurat H, vom ilustra modul în care operatorii fuzzy sunt folosiţi pentru a crea noi variabile fuzzy în cadrul bazei de reguli (sursă [Lescieux 06]).

Dacă A reprezintă submulţimea fuzzy "*înălţime mică*", dată de funcţia de apartenenţă fuzzy $\alpha_A()$, iar B este submulţimea fuzzy "*înălţime medie*", dată de funcţia de apartenenţă $\alpha_B()$ (vezi Figura 6.5.a), atunci noua mulţime fuzzy "*înălţime mică*" *SAU* "*înălţime medie*" va avea funcţia de apartenenţă, $\alpha_{A \cup B}()$, dată de ecuaţia următoare:

$$\alpha_{A \cup B}(x) = \alpha_A(x) \cup \alpha_B(x), \quad \forall x \in U \tag{6.22}$$

unde x ia valori în universul de discurs U, cu $A, B \subset U$.

Dacă operatorul de reuniune este exprimat, de exemplu, cu t-conorma Zadeh, atunci funcţia de apartenenţă $\alpha_{A \cup B}()$ poate fi exprimată în felul următor:

$$\alpha_{A \cup B}(x) = Max(\alpha_A(x), \alpha_B(x)), \quad \forall x \in U \tag{6.23}$$

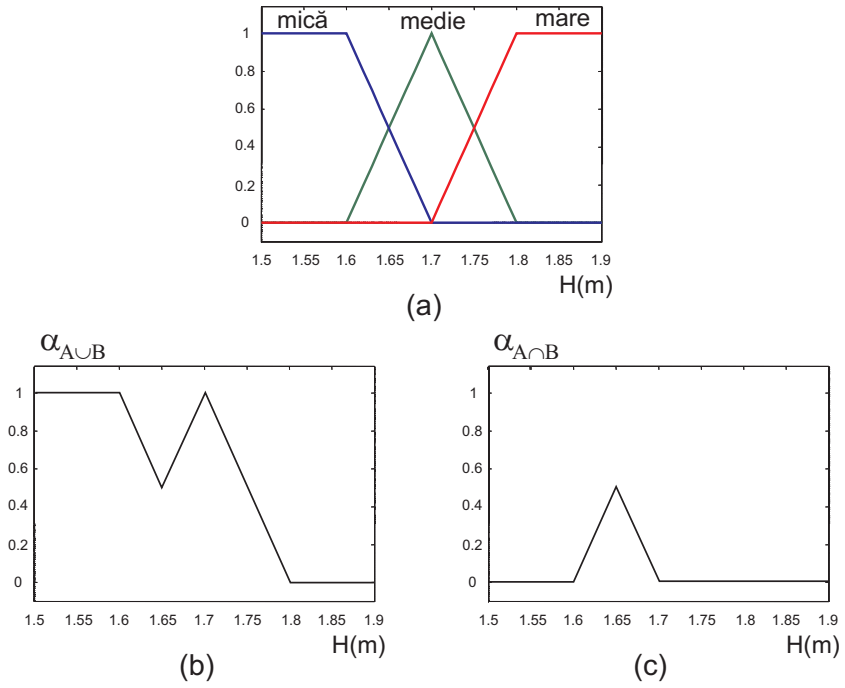


Figura 6.5: Operatori fuzzy: (a) partiţia universului de discurs, (b) funcţia de apartenenţă $\alpha_{A\cup B}$, (c) funcţia de apartenenţă $\alpha_{A\cap B}$ (A, B reprezintă submulţimile fuzzy "*înălţime mică*" şi respectiv "*înălţime medie*").

Similar, mulţimea fuzzy "*înălţime mică*" *ŞI* "*înălţime medie*" va avea funcţia de apartenenţă dată de relaţia:

$$\alpha_{A\cap B}(x) = \alpha_A(x) \cap \alpha_B(x) = Min(\alpha_A(x), \alpha_B(x)), \quad \forall x \in U \tag{6.24}$$

unde operatorul de intersecţie a fost definit folosind t-norma Zadeh. Noile funcţii de apartenenţă astfel obţinute pe baza operatorilor Zadeh sunt ilustrate în Figura 6.5.

Baza de reguli fuzzy ce va modela comportamentul dinamic al sistemului este definită în funcţie de context pe baza diferitelor combinaţii relevante dintre mulţimile fuzzy ce descriu caracteristicile sistemului.

Generarea regulilor fuzzy

În general, o regulă fuzzy este exprimată folosind următoarea structură:

$$\begin{aligned} &\textbf{DACĂ } (X_1 \textit{ este } A_1) \textit{ op } ... \textit{ op } (X_n \textit{ este } A_n) \textbf{ ATUNCI} \\ &\qquad (Y_1 \textit{ este } B_1) \textit{ op } ... \textit{ op } (Y_m \textit{ este } B_m) \end{aligned} \tag{6.25}$$

unde X_i şi Y_j reprezintă variabilele lingvistice de intrare şi respectiv de ieşire ale sistemului, A_i şi B_j reprezintă submulţimile fuzzy ce definesc partiţionarea spaţiului de intrare şi respectiv de ieşire al sistemului, $i = 1, ..., n$, $j = 1, ..., m$, o propoziţie de tip "X *este* A" este cuantificată de gradul de apartenenţă al variabilei lingvistice X la submulţimea fuzzy A iar *op* reprezintă un operator fuzzy (de exemplu, intersecţie sau reuniune). Această modalitate de formulare este cunoscută şi sub numele de Mamdani, având particularitatea că valoarea returnată este tot o mulţime fuzzy, similară intrării.

Astfel, în formularea unei reguli fuzzy întâlnim trei părţi distincte, şi anume: *premisele* regulii, *implicaţia* acestora precum şi *concluzia*:

- **premisele** reprezintă ipotezele de plecare ce sunt exprimate sub forma unuei succesiuni de afirmaţii, interconectate de operatori fuzzy. Acestea sunt încadrate în regula fuzzy de cuvintele cheie "DACĂ" şi "ATUNCI",
- **implicaţia** premiselor este anunţată de cuvântul cheie "ATUNCI",
- **concluzia** este formulată de regulă în mod similar cu premisele, şi anume folosind o succesiune de afirmaţii. Aceasta se găseşte după cuvântul cheie "ATUNCI" şi reprezintă rezultatul premiselor iniţiale.

Dacă luăm ca exemplu următoarea regulă simplă:

DACĂ (*Vremea* **este** "frumoasă") **şi** (*Momentul* **este** "dimineaţa")
ATUNCI (*Moralul* **este** "ridicat")

atunci premisele acesteia vor fi date de variabilele *Vremea* şi *Momentul* ce iau valorile "frumoasă" şi respectiv "dimineaţa", premise ce duc la implicaţia concluziei că variabila *Moralul* va lua valoarea "ridicat".

Similar variabilelor fuzzy, regulile fuzzy au şi ele o valoare de adevăr. Fiind dată o regulă, r, valoarea de adevăr a acesteia pentru un anumit exemplu, numită şi pondere a regulii, notată cu ω_r, rezultă dintr-o operaţie logică între elementele premisei, astfel:

$$\omega_r = \alpha_{A_1}(x_1)\ op\ \alpha_{A_2}(x_2)\ op\ ...\ op\ \alpha_{A_n}(x_n) \tag{6.26}$$

unde $\alpha_{A_i}(x_i)$ reprezintă gradul de apartenenţă a valorii x_i din universul de discurs la submulţimea fuzzy A_i, cu $i = 1, ..., n$ iar *op* este un operator fuzzy.

O altă modalitate de definire a unei reguli este folosind modelul Takagi-Sugeno. Diferenţa faţă de modelul Mamdani constă în faptul că în acest caz concluzia regulii va fi una netă. Astfel, valoarea de ieşire j a sistemului pentru regula r este calculată ca fiind o combinaţie liniară a valorilor de intrare, astfel:

$$y_j^r = b_{j,0} + b_{j,1} \cdot x_1 + b_{j,2} \cdot x_2 + ... + b_{j,n} \cdot x_n \tag{6.27}$$

unde $b_{j,i}$, cu $i = 1, ..., n$ şi $j = 1, ..., m$, reprezintă coeficienţii de pondere ce sunt determinaţi în funcţie de importanţa fiecărei valori de intrare, x_i, pentru valoarea ieşirii sistemului.

Procesul de inducere a bazei de reguli fuzzy a unui sistem constă în formularea tuturor mulţimilor de reguli ce vor face legătura între ieşirile sistemului şi intrările acestuia. Pentru un studiu detaliat a literaturii de specialitate din teoria mulţimilor fuzzy, cititorul se poate raporta la lucrările [Guillaume 01] şi [Klir 95].

6.4 Avantajele reprezentării fuzzy

Avantajele furnizate de logica fuzzy la formalizarea conceptuală a informaţiilor numerice, cât şi simbolice, de nivel scăzut, nu sunt deloc neglijabile. Acestea pot fi sintetizate cu următoarele:

- sistemele de inferenţă fuzzy sunt **aproximatori universali**. Universul de discurs, care de regulă este foarte vast sau chiar infinit, este convertit cu ajutorul formalizării fuzzy într-un număr limitat de concepte. Cercetările din domeniu au demostrat de-a lungul timpului că sistemele de inferenţă fuzzy sunt cel puţin la fel de performante ca alte tehnici consacrate de aproximare a datelor [Wang 92].

- formalizarea fuzzy **reduce complexitatea** sistemului modelat. În cazul în care cantitatea de informaţie ce trebuie prelucrată este mult prea mare pentru a putea fi controlată în totalitate, iar înţelegerea proceselor existente în sistem este limitată, formalizarea fuzzy permite reducerea complexităţii prin introducerea conceptului de incertitudine. În general, s-a dovedit faptul că, cu cât este tolerată mai multă incertitudine în modelarea sistemului, cu atât complexitatea acestuia tinde să scadă, iar credibilitatea modelului obţinut tinde să crească [Klir 95].

- datorită conceptului de incertitudine, variabilele fuzzy **reprezintă realitatea**, care este în general una incertă, mult mai fidel decât o fac variabilele clasice nete. Această proprietate a fost enunţată chiar de Albert Einstein în 1921: *pe cât legile matematice se raportează la realitate, pe atât acestea nu sunt veridice. Şi pe cât sunt veridice, acestea nu se raportează la realitate.*

- în logica fuzzy, conceptele vagi sunt reprezentate într-un **limbaj natural**. Această proprietate este una dintre cele mai impotante. Formalizarea fuzzy transformă mărimi numerice în concepte lingvistice ce

sunt exprimate într-un mod similar modului de percepţie uman. Dacă universul de discurs este unul numeric, formalizarea fuzzy este realizată cu variabile lingvistice ce iau valori textuale [Lescieux 06].

- modul de funcţionare al logicii fuzzy este foarte similar cu modul de percepţie uman. Însuşi creierul uman funcţionează după principiile logicii fuzzy. Creierul uman apreciază valorile stimulilor de intrare (valori continue) într-un mod aproximativ. De exemplu, apreciem un obiect ca fiind apropiat sau depărtat şi nu ca fiind la distanţa punctuală d. Astfel, formalizarea fuzzy tinde să fie **în concordanţă cu percepţia semantică**, facilitând astfel "extragerea de cunoaştere" ("knowledge") din date numerice de nivel scăzut.

- faptul că formalizarea fuzzy este construită în general pe baza expertizei şi a cunoaşterii "a priori" a domeniul de aplicaţie, conferă modelului obţinut mult mai multă **coerenţă** decât unui model clasic.

- formalizarea fuzzy **normalizează valorile datelor** de intrare. Conceptelor lingvistice le sunt asociate valori de adevăr ce sunt valori reale, de regulă normalizate între 0 şi 1, unde 0 reprezintă negaţia sigură iar 1 afirmaţia sigură. În acest fel, comparaţia parametrilor cu plaje de valori diferite este simplificată.

- reprezentarea clasică a datelor pe baza logicii booleene este un caz particular al logicii fuzzy. Astfel, **formalizarea fuzzy include formalizarea netă**.

În cele ce urmează, vom prezenta câteva exemple de aplicaţii practice ale formalizării fuzzy, atât în domeniul indexării după conţinut a secvenţelor de imagini, cât şi în domenii conexe.

6.5 Aplicabilitatea sistemelor fuzzy

Fiind foarte apropiată de modul în care percepem realitatea înconjurătoare, logica fuzzy a câştigat rapid teren în aproape toate domeniile de activitate existente, înlocuind parţial sau chiar în totalitate logica clasică booleană. Acest lucru se datorează în mare parte faptului că aceasta modelează însuşi modul de funcţionare al creierului uman, care este unul fuzzy şi nu net.

Creierul uman apreciază diversele variabile din mediul înconjurător într-un mod aproximativ, ca de exemplu folosind atribute precum redus/ridicat, aproape/departe, etc. Acelaşi lucru este valabil şi pentru răspunsul la anumiţi stimuli exteriori, de exemplu când conducem o maşină frânăm mai puţin sau

mai tare în funcţie de indicaţia semaforului şi de distanţa până la acesta. Astfel, pe baza acestor date aproximative, creierul uman îşi dezvoltă o serie de reguli, concepute într-un mod similar modului de inferare a regulilor fuzzy, reguli ce vor determina răspunsul nostru în funcţie de diversele valori ale variabilelor de intrare.

Sistemele fuzzy îşi găsesc aplicaţie practic în toate domeniile existente ce se bazează pe luarea de decizii, ca de exemplu: asistare la diagnostic şi luarea de decizii (domeniul medical, orientare profesională, etc.), baze de date (obiecte fuzzy şi formularea cererilor de căutare cu mulţimi fuzzy), recunoaştere automată a formelor, vedere asistată de calculator, prelucrare de imagini, sisteme de agregare multicriteriu, sisteme de optimizare, comandă fuzzy a sistemelor, etc. Standardizarea anumitor metode de analiză fuzzy a dus chiar la implementarea hardware sau software a acestora în diverse produse finite destinate publicului larg sau specialiştilor din domeniu, ca de exemplu: aparate electrocasnice, sisteme audio-vizuale, procesoare dedicate, interfeţe de dezvoltare specifice precum procesorul Motorola 68HC12 sau Thomson WARP [Lescieux 06].

În domeniul prelucrării de imagini şi al vederii asistate de calculator, aportul logicii fuzzy este incontestabil. Aceasta constituie puntea de legătură între datele numerice de nivel scăzut obţinute în urma diverselor măsurători ale proprietăţilor fizice ale fenomenelor analizate, sau în urma anumitor etape de prelucrare, şi modul de percepţie al acestora. Conceptele fuzzy sunt exprimate într-un limbaj natural ceea ce face ca "cifrele să capete sens".

Datorită diversităţii foarte mare de aplicaţii posibile ale logicii fuzzy, este aproape imposibil să realizăm o clasificare a metodelor existente. Astfel că, în cele ce urmează ne vom limita la prezentarea a doar câteva exemple semnificative de aplicaţii ale logicii fuzzy în domeniul analizei şi prelucrării secvenţelor de imagini.

Un exemplu de folosire a logici fuzzy pentru a simplifica şi a da sens comparaţiei datelor numerice cu plaje de valori diferite, este metoda propusă în [Doulamis 00c]. În aceasta, variabilele fuzzy sunt folosite pentru a caracteriza proprietăţile de culoare şi de adâncime a secvenţelor de imagini stereoscopice[3] în vederea rezumării automate a conţinutului acestora. Pentru aceasta, conţinutul vizual al fiecărei imagini este mai întâi reprezentat cu o serie de parametri extraşi la nivel de segment (obiect), formând astfel vectorul de caracteristici al imaginii. Cum numărul de segmente variază de la o imagine la alta, dimensiunea vectorului de caracteristici va fi de asemenea

[3]imaginile stereoscopice sau imaginile 3D sunt imagini ce sunt capabile să redea informaţia spaţială 3D prin crearea iluziei de adâncime. În cazul imaginilor sau a secvenţelor de imagini, efectul de adâncime este obţinut prin înregistrarea a două imagini din perspective un pic diferite, ce vor fi proiectate independent fiecărui ochi.

variabilă, fapt ce duce la imposibilitatea de a evalua similaritatea imaginilor pe baza comparaţiei vectorilor de caracteristici.

Pentru a soluţiona această problemă sunt folosite variabile fuzzy. Segmentele sunt mai întâi clasificate într-un anumit număr de clase predefinite, formând astfel o histogramă multidimensională. Fiecare element al vectorului de caracteristici va corespunde în acest fel unei anumite clase sau cu alte cuvinte, unui anumit bin al histogramei. Pentru a evita clasificarea a două segmente similare în clase diferite, clasificare ce poate duce la erori de comparaţie semnificative, fiecare clasă va fi reprezentată de un grad de apartenenţă fuzzy. Astfel, fiecărui segment i se va permite să aparţină la mai multe clase, dar cu grade diferite de apartenenţă.

Din punct de vedere matematic, fiecare segment S_i, cu $i = 1, ..., K$, unde K reprezintă numărul de segmente extrase din imagine, va fi reprezentat de vectorul de caracteristici s_i următor: $s_i = [c_{S_i}^T d_{S_i} l_{S_i}^T a_{S_i}]$, unde matricea c_{S_i} conţine componentele medii de culoare ale segmentului S_i, d_{S_i} şi a_{S_i} reprezintă adâncimea şi respectiv dimensiunea acestuia iar l_{S_i} este o matrice ce conţine coordonatele orizontale şi verticale ale centrului segmentului (^{T} denotă operaţia de transpunere matriceală).

Domeniul de valori al elementului j din vectorul de caracteristici s_i va fi partiţionat în Q regiuni folosind funcţiile de apartenenţă fuzzy $\mu_{n_j}(s_{i,j})$ (funcţii triunghiulare cu suprapunere de 50%), unde $n_j = 1, ..., Q$. Astfel, $\mu_{n_j}(s_{i,j})$ va reprezenta gradul de apartenenţă al lui $s_{i,j}$ la clasa de indice n_j. Ansamblul claselor n_j pentru toate valorile lui j, cu $j = 1, ..., L$, va fi dat de clasa L-dimensională $n = [n_1 n_2 ... n_L]^T$. În acest fel, gradul de apartenenţă al vectorului de caracteristici s_i la clasa n, $\mu_n(s_i)$, poate fi definit ca un produs de funcţii de apartenenţă fuzzy, astfel:

$$\mu_n(s_i) = \prod_{j=1}^{L} \mu_{n_j}(s_{i,j}) \tag{6.28}$$

Mai departe, histograma fuzzy de caracteristici, $H_f()$, este construită ca fiind:

$$H_f(n) = \frac{1}{K} \sum_{i=1}^{K} \mu_n(s_i) \tag{6.29}$$

unde valoarea $H_f(n)$ va reprezenta gradul de apartenenţă al întregii imagini la clasa n. Vectorul de caracteristici al imagini va fi dat astfel de valorile lui $H_f()$ pentru toate clasele posibile n. Acesta va fi folosit mai departe pentru a evalua similaritatea imaginilor în vederea selectării imaginilor cheie ale rezumatului.

În [Dorado 04] variabilele fuzzy sunt folosite pentru a "copia" modul de percepţie uman în încercarea de adnotare automată a conţinutului secvenţelor de imagini. Astfel, datele de nivel scăzut extrase din secvenţă vor fi convertite într-o serie de concepte semantice. Aceste concepte sunt definite de un anumit lexicon format, atât din simboluri lingvistice, cât şi din simboluri grafice considerate ca fiind semnificative pentru descrierea conţinutului secvenţei. Etichetarea secvenţei cu simbolurile din lexicon va fi realizată folosind o bază de reguli. Aceasta este obţinută într-o etapă de învăţare, pe baza expertizei domeniului de aplicaţie.

Principiul de extragere a "cunoaşterii" ("knowledge") din parametrii numerici de nivel scăzut este următorul: dacă W^f reprezintă setul de cuvinte asociate parametrilor de nivel scăzut consideraţi, $W^f \subset L$, unde L reprezintă lexiconul semantic, atunci procesul de asociere a conceptelor semantice poate fi definit ca fiind o aplicaţie $\tilde{M}(w, \tilde{A})$ ce face trecerea de la un set de cuvinte din L la un set de interpretări ale acestora definite de universul Y. Fiecare cuvânt, $w \in L$, corespunde unei mulţimi fuzzy, $\tilde{A} \in Y$, mulţime ce reprezintă interpretarea semantică a lui w. Gradul de adevăr al acestei apartenenţe va fi furnizat de funcţiile de apartenenţă fuzzy: $\mu_{\tilde{M}}(w, y) = \mu_{\tilde{A}}(y)$.

De exemplu, în cazul secvenţelor de ştiri, folosind ca parametri de nivel scăzut decupajul în plane video şi culorile predominante ale anumitor imagini cheie ale secvenţei, se poate defini un lexicon care să conţină două concepte, şi anume: conceptul de "anchorperson"[4] ce survine în cazul în care culorile predominante prezintă variaţii slabe pe parcursul secvenţei, şi respectiv conceptul de "reportaj" în cazul în care culorile predominante au o variaţie semnificativă. În acest caz, reprezentarea semantică este dată de două mulţimi fuzzy ce sunt legate de valoarea medie a schimbării de culoare.

Un alt exemplu este sistemul propus în [Ionescu 08] unde mulţimile fuzzy sunt folosite pentru a descrie percepţia conţinutului de culoare din secvenţele artistice de animaţie. Metoda folosită este una clasică de inferenţă fuzzy. Conţinutul de culoare al secvenţei este mai întâi caracterizat cu o serie de parametri statistici de nivel scăzut, precum distribuţia globală de culoare a secvenţei, distribuţia de culori elementare, gradul de variabilitate al culorilor, procentul de culori închise sau deschise, etc.

Trecerea la un nivelul semantic al descrierii se realizează în două etape. Mai întâi, un nivel simbolic de descriere este obţinut prin asocierea de mulţimi fuzzy fiecărui parametru de nivel scăzut. Funcţiile de apartenenţă sunt definite pe baza expertizei domeniului particular al filmului de animaţie. De

[4] "anchorperson" este un termen folosit în domeniul televiziunii pentru a desemna reporterul ce coordonează o anumită transmisiune TV în care intervin mai mulţi corespondenţi.

exemplu, variabila lingvistică "light color content" este asociată parametrului P_{light} ce reprezintă procentul de culori deschise prezente în secvenţă. Conceptul este descris mai departe cu trei simboluri, şi anume: "low-light color content", "mean-light color content" şi respectiv "high-light color content". Astfel, secvenţa are o distribuţie de culoare săracă în culori deschise (valoare de adevăr 1) dacă $100 \cdot P_{light} < 33\%$, o distribuţie cu un conţinut mediu de culori deschise (valoare de adevăr 1) dacă $100 \cdot P_{light} > 50\%$ şi $100 \cdot P_{light} < 60\%$, şi respectiv o distribuţie de culoare bogată în culori deschise (valoare de adevăr 1) dacă $100 \cdot P_{light} > 66\%$.

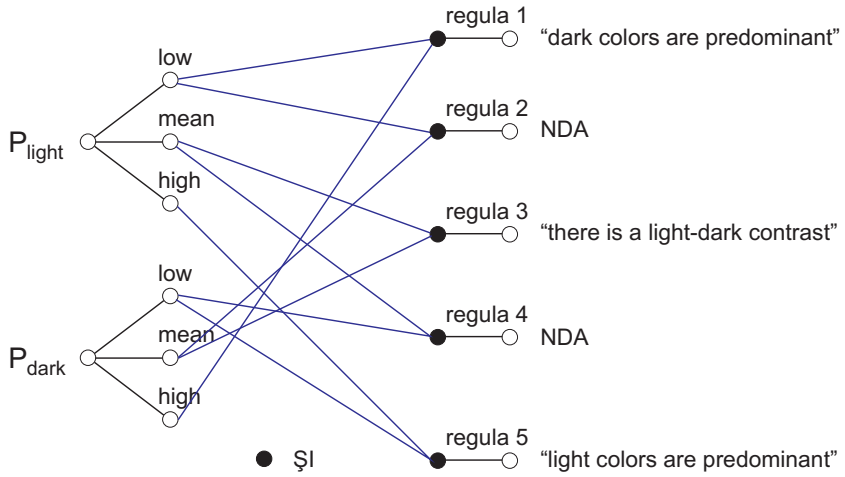


Figura 6.6: Exemplu de bază de reguli fuzzy folosită la descrierea intensităţii de culoare în filmele artistice de animaţie (NDA = "fără descriere", sursă [Ionescu 08]).

Nivelul de descriere semantic propriu-zis este obţinut prin inducerea de reguli fuzzy. Descrierile vizate sunt tehnicile artistice de culoare prezente în filmele de animaţie, precum cele şapte contraste ale lui Itten sau schemele de armonie a culorilor (vezi Secţiunea 4.2.3). Un exemplu este prezentat în Figura 6.6. Descrierile obţinute, atât simbolice cât şi semantice sunt folosite ulterior pentru indexarea automată după conţinut a unei baze de secvenţe de imagini, precum şi ca informaţii de conţinut pentru facilitarea navigării în baza de date.

6.6 Concluzii

În acest capitol am prezentat modul în care formalizarea fuzzy poate fi folosită la descrierea semantică, pe bază de concepte lingvistice, a conţinutului

datelor. Logica fuzzy, spre deosebire de logica clasică booleană, prin introducerea conceptului de incertitudine are avantajul de a fi apropiată modului de percepţie uman, care este unul aproximativ. În logica fuzzy, percepţia fenomenelor înconjurătoare nu este o problemă de validare sau invalidare, ci mai degrabă o problemă de grad de adevăr.

Formalizarea fuzzy implică de regulă două etape distincte. Prima etapă este etapa de "fuzzyficare" şi constă în asocierea de concepte lingvistice parametrilor numerici ce modelează sistemul. Acestea sunt definite de o serie de valori lingvistice ce constituie submulţimile fuzzy ale universului de discurs. Etapa a doua este etapa de inferenţă fuzzy şi constă în inducerea unei baze de reguli fuzzy pentru a modeliza, atât relaţiile vizibile, cât şi cele ascunse ce pot exista între diverşii parametri ai sistemului.

Unul dintre principalele avantaje ale formalizării fuzzy, relativ la utilizarea acesteia la descrierea anumitor procese de interes, constă în trecerea de la o reprezentare numerică de nivel scăzut a proprietăţilor sistemului modelat la o reprezentare lingvistică de nivel semantic superior. Astfel, formalizarea fuzzy *dă sens* valorilor numerice obţinute în diverse etape de prelucrare. Dacă universul de discurs este unul numeric, universul conceptelor fuzzy este unul textual.

Unul dintre principalele dezavantaje ale formalizării fuzzy este dat de faptul că în cele mai multe cazuri, exceptând sistemele cu un grad de complexitate foarte redusă, aceasta nu este un proces automat. Formalizarea fuzzy necesită intervenţia unui operator sau a unui expert. Acesta, pe baza expertizei domeniului de aplicaţie va determina mecanismul de "fuzzyficare", şi eventual pe cel de inferenţă fuzzy, astfel încât sistemul modelat să răspundă cerinţelor de prelucrare. Pe de altă parte, tocmai implicarea raţionamentului uman în procesul de proiectare face ca formalizarea fuzzy să fie în conformitate cu realitatea.

Datorită caracterului semantic, evaluarea unui model fuzzy este un proces subiectiv. Totuşi, implicarea umană în procesul de proiectare al mecanismului fuzzy implică într-o anumită măsură şi veridicitatea rezultatelor obţinute.

CAPITOLUL 7

Clasificarea după conţinut a datelor

Rezumat: *Căutarea informaţiei utile într-o colecţie mare de date este un lucru relativ dificil de realizat. Aceasta se datorează imposibilităţii noastre de a procesa şi tria volume mari de date. Gruparea automată a datelor după anumite criterii de similaritate, transferă problema căutării globale într-o problemă de localizare a datelor într-o colecţie restrânsă de date de acelaşi tip. Metodele de clasificare automată, vizează tocmai partiţionarea automată a datelor în clase omogene din punct de vedere al conţinutului. În acest capitol vom prezenta tehnicile existente de clasificare, atât supervizată cât şi nesupervizată, punând accentul pe utilitatea acestora în procesul de indexare după conţinut.*

Tehnicile de clasificare a datelor permit, pe baza definirii unor criterii de similaritate, regruparea automată în funcţie de conţinut a volumelor vaste de date. Aceste date, accesate în mod direct, sunt de regulă greu accesibile utilizatorului.

Din punct de vedere matematic, problematica abordată de metodele de clasificare existente este următoarea: având la dispoziţie o mulţime de N obiecte ale căror proprietăţi fizice sunt sintetizate prin intermediul unor vectori multidimensionali de caracteristici, problema care se pune este definirea unei anumite partiţii, cât mai relevante, a acestor obiecte într-un anumit număr de clase [Hinneburg 00]. În acest scop se defineşte conceptul de simi-

laritate între vectorii de caracteristici asociaţi obiectelor, concept pe baza căruia obiectele aşa zise similare vor fi atribuite aceleiaşi clase. Principiul clasificării este sintetizat în Figura 7.1.

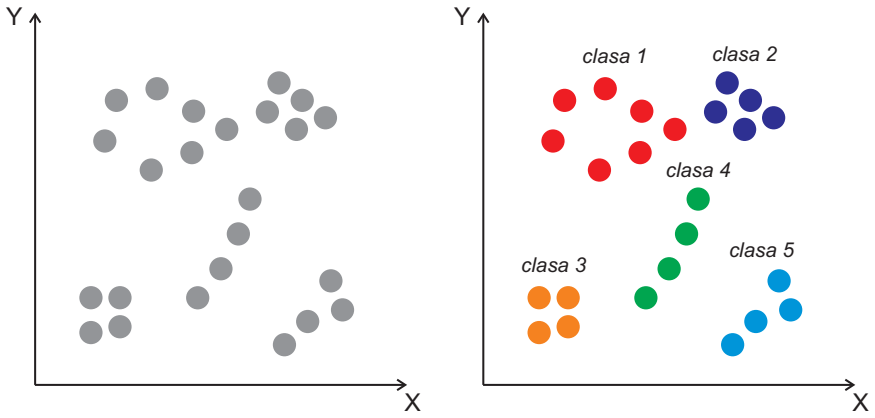


Figura 7.1: Principiul clasificării datelor: (a) datele de intrare reprezentate în spaţiul de caracteristici, (b) repartiţia în clase obţinută în urma clasificării (obiectele din aceeaşi clasă sunt reprezentate cu aceeaşi culoare).

Tehnicile de clasificare existente sunt utilizate într-o gamă foarte largă de aplicaţii. Acestea deservesc diverse obiective, dintre care putem menţiona drept cele mai importante următoarele:

- **reducerea volumului informaţional**: tehnicile de clasificare a datelor permit regruparea unui ansamblu de date în grupuri omogene, lucru ce facilitează reducerea volumului informaţional disponibil. De exemplu, fiecare grup (clasă) de date poate fi reprezentat, în etapele de prelucrare ulterioare, doar de informaţia cea mai reprezentativă a grupului. La un alt nivel, clasificarea datelor permite eliminarea redundanţei informaţionale prin reducerea spaţiului de caracteristici.

- **punerea în evidenţă a "cunoaşterii"**: tehnicile de clasificare permit localizarea într-un volum mare de date a unor grupuri de informaţii ce prezintă anumite caracteristici de interes. Localizarea acestora furnizează utilizatorului o cunoaştere nouă a relaţiilor existente între date, cunoaştere ce nu era disponibilă anterior căutării. Acest proces este cunoscut în literatura de specialitate şi sub numele de "data mining".

- **punerea în evidenţă a relevanţei claselor**: tehnicile de clasificare permit localizarea anumitor grupuri de date ce sunt reprezentative pentru ansamblul datelor analizate,
- **punerea în evidenţă a datelor atipice**: tehnicile de clasificare permit de asemenea localizarea datelor ce nu corespund niciunui criteriu de similaritate, date ce sunt considerate ca fiind atipice pentru criteriile considerate. Acestea sunt importante, deoarece sunt un caz particular şi trebuiesc analizate separat. Un astfel de exemplu sunt datele ce se găsesc pe frontiera dintre două clase diferite, date ce pot fi considerate ca aparţinând ambelor clase cât şi ca o clasă independentă.

Din punct de vedere al problematicii indexării datelor, subiect ce face obiectul acestei cărţi, tehnicile de clasificare sunt indispensabile unui sistem de indexare după conţinut. Clasificarea datelor intervine în general în însuşi procesul de căutare al informaţiei. Utilizatorul, prin formularea cererii de căutare va defini spaţiul de caracteristici ce va fi folosit pentru localizarea datelor dorite. Pe baza acestuia, datele din baza de date pot fi grupate în funcţie de similaritate sau cu alte cuvinte în funcţie de asemănarea dintre vectorii de caracteristici asociaţi. Astfel, grupul sau grupurile de date ce sunt suficient de similare vectorului de caracteristici asociat cererii de căutare vor fi furnizate utilizatorului drept rezultat.

Metodele de clasificare existente se împart în două mari categorii. Prima categorie de metode o constituie *metodele probabilistice* sau de *clasificare supervizată*. Clasificarea supervizată implică clasarea datelor pe baza unor modele predefinite de clase sau "date de antrenament". Acestea reprezintă de regulă o clasificare de referinţă ce corespunde realităţii (similară unei "realităţi de teren" sau "groundtruth"[1]), folosită iniţial pentru antrenarea sistemului înaintea clasificării propriu-zise a datelor. În literatura de specialitate, termenul asociat metodelor din această categorie este de "metode de clasificare" sau "classification methods".

O a doua categorie de metode de clasificare sunt *metodele de clasificare nesupervizată sau automată*, desemnate în literatura de specialitate prin termenul de "clustering"[2]. Clasificarea nesupervizată, spre deosebire de clasificarea supervizată, propune o partiţie optimală a spaţiului de caracteristici din punct de vedere al unui anumit criteriu matematic, fără a folosi informaţii "a priori" (de exemplu, o partiţie de referinţă). Avantajul acestor metode este dat de faptul că sunt complet automate (nu necesită intervenţia utilizatorului) şi pot fi folosite pentru clasarea datelor despre care nu dispunem de

[1] vezi explicaţia de la pagina 170.

[2] de notat este faptul că în limba română, termenul de clasificare este folosit generic pentru a desemna, în funcţie de context, atât o clasificare supervizată cât şi nesupervizată.

informaţii relative la conţinutul acestora (număr de clase, prototipul clasei, etc.). Pe de altă parte, fiind un proces automat, relevanţa claselor tinde să fie mai redusă decât în cazul clasificării supervizate, aceasta fiind dependentă de metoda folosită cât şi de puterea discriminatorie a spaţiului de caracteristici folosit.

În cele ce urmează, vom face o trecere în revistă a tehnicilor de clasificare supervizată şi nesupervizată existente punând în evidenţă avantajele cât şi dezavantajele fiecărei abordări.

7.1 Clasificarea nesupervizată a datelor

O catalogare interesantă a metodelor de clasificare nesupervizată existente în funcţie de proprietăţile contrastante ale acestora este propusă în [Jain 99], astfel, metodele de clasificare nesupervizată sunt:

- **acumulative sau partiţionale**: această proprietate este legată de modul de structurare al algoritmului folosit. Metodele acumulative pornesc clasificarea de la o anumită partiţie în clase, clase care pe parcursul algoritmului sunt fuzionate iterativ până când este satisfăcut un anumit criteriu de convergenţă. Pe de altă parte, metodele partiţionale pornesc de la o singură clasă ce este divizată iterativ până când criteriul de convergenţă considerat este satisfăcut.

- **monotetice sau politetice**[3]: aceste proprietăţi sunt legate de modul de utilizare a vectorilor de caracteristici în procesul de clasificare, care poate fi secvenţial sau simultan. Marea parte a metodelor existente sunt politetice, astfel că pentru estimarea distanţei dintre obiecte sunt folosiţi toţi parametrii disponibili ("features"). De asemenea, decizile de clasare sunt luate pe baza acestei măsuri de distanţă. Pe de altă parte, metodele monotetice folosesc parametrii în mod secvenţial pentru a constitui progresiv clasele, de exemplu, parametrul x_1 este folosit pentru a diviza datele în două clase, mai departe, parametrul x_2 este folosit pentru divizarea claselor anterioare, şi aşa mai departe.

- **nete sau fuzzy**. O clasificare netă presupune alocarea fiecărui obiect unei singure clase, astfel apartenenţa fiind sigură. Pe de altă parte, o clasificare fuzzy asociază fiecărui obiect un grad de apartenenţă la una sau mai multe clase, apartenenţa la clase fiind de această dată incertă.

[3]termenul de monotetic desemnază o anumită clasă ai cărei membrii sunt identici din punct de vedere al tuturor caracteristicilor acestora, în timp ce termenul de politetic desemnează o anumită clasă ai cărei membrii sunt similari dar nu identici.

Clasificarea fuzzy permite trecerea la o clasificare netă prin asocierea obiectelor claselor cu gradul de apartenenţă cel mai semnificativ.

- **deterministe sau stohastice**: aceste proprietăţi sunt relevante mai ales pentru abordările partiţionale ce se bazează pe optimizarea unei erori pătratice. Această optimizare poate fi realizată, fie în mod clasic (determinist), fie pe baza unei căutări aleatoare în spaţiul format de toate clasificările posibile (stohastic).

- **incrementale sau non-incrementale**: aceste proprietăţi intervin atunci când volumul de date este foarte mare şi constrângerile impuse (timp de execuţie, putere de calcul, etc.) tind să influenţeze arhitectura algoritmului. Astfel, în general, metodele incrementale încearcă să minimizeze numărul de citiri al datelor, să reducă numărul de repartiţii în clase analizate pe parcursul algoritmului sau să reducă dimensiunea structurilor de date folosite de algoritm.

În general, diversitatea metodelor de clasificare existente lasă suficientă flexibilitate clasificării astfel încât aceasta să poată fi implementată optimal pentru aplicaţia dorită.

7.1.1 Etapele clasificării nesupervizate

În ciuda diversităţii abordărilor existente, metodele de clasificare nesupervizată implică o serie de etape comune [Jain 99], astfel:

- **reprezentarea datelor**: această primă etapă constă în alegerea parametrilor metodei de clasificare, cât şi a datelor folosite, de exemplu: numărul de clase anticipate, numărul de modele disponibile, tipul şi dimensiunea vectorilor de caracteristici ce vor defini spaţiul în care se va efectua clasificarea. De asemenea, tot în această etapă se poate realiza *selecţia de caracteristici* sau/şi *extragerea de caracteristici*. Selecţia caracteristicilor reprezintă procesul de identificare a unei submulţimi optimale a setului iniţial de caracteristici. Această etapă permite astfel reducerea dimensiunii datelor de intrare (reducerea dimensiunii vectorilor de caracteristici) şi astfel a complexităţii de calcul. Pe de altă parte, extragerea de caracteristici foloseşte o anumită transformare pentru a proiecta vectorii de caracteristici iniţiali într-un alt spaţiu de reprezentare ce oferă o topologie mai eficientă pentru clasificare.

- **alegerea unei măsuri de distanţă**: similaritatea vectorilor de caracteristici este cuantificată pe baza unei anumite măsuri de distanţă sau funcţii de similaritate. Distanţa Euclidiană este una dintre măsurile

cele mai des întâlnite, datorită simplităţii şi eficienţei acesteia. Alte măsuri uzuale sunt distanţa Mahalanobis[4] sau distanţa Hausdorff[5]. Această etapă permite definirea noţiuni de "proximitate" între date, astfel că alegerea măsurii de distanţă este definitorie pentru performanţele algoritmului de clasificare. Pentru un studiu bibliografic al măsurilor de distanţă folosite la clasificarea nesupervizată a datelor, cititorul se poate raporta la lucrarea [Jain 99].

- **clasificarea propriu-zisă**: această etapă permite regruparea datelor în clase omogene în funcţie de vectorii de caracteristici asociaţi acestora. Abordările folosite au fost discutate în paragraful anterior (vezi Secţiunea 7.1).

- **reducerea vectorilor de caracteristici**: această etapă este realizată doar dacă este necesară şi constă în reducerea dimensiunii vectorilor de caracteristici, de regulă n-dimensionali cu $n > 3$, la o dimensiune ce poate fi reprezentată grafic, de regulă $n = 3$. Procesul de reducere a informaţiei trebuie să conserve pe cât posibil informaţia utilă conţinută în clase. Un exemplu clasic de astfel de metodă este analiza în componente principale sau PCA ("Principal Components Analysis"[6]).

- **evaluarea rezultatelor**: evaluarea pertinenţei repartiţiei în clase este de asemenea o etapă opţională. Metodele folosite sunt, fie *matematice* precum evaluarea tendinţei de fragmentare a claselor sau analiza validităţii claselor, fie *empirice* precum analiza manuală a repartiţiei în clase pe baza vizualizării grafice a acestora.

7.1.2 Metodele existente de clasificare nesupervizată

Tehnicile de clasificare în general, au fost studiate intensiv de mai bine de 30 de ani. Diversitatea metodelor existente face dificilă găsirea metodei care să corespundă cel mai bine necesităţilor de prelucrare ale aplicaţiei vizate.

În cele ce urmează vom realiza o trecere în revistă a metodelor de clasificare nesupervizată cel mai frecvent întâlnite în literatura de specialitate, evidenţiind particularităţile şi avantajele fiecărei dintre acestea. Astfel, vom prezenta:

[4]distanţa Mahalanobis a fost introdusă pentru prima oară de statisticianul P.C. Mahalanobis în anul 1936. Aceasta se bazează pe analiza corelaţiei dintre date şi este folosită pentru a măsura gradul de separare a două grupuri de date. De exemplu, distanţa Mahalanobis dintre doi vectori aleatori de aceeaşi distribuţie, x şi y, este dată de relaţia următoare: $d_M^2(x, y) = (x - y)^T \Sigma^{-1}(x - y)$, unde Σ reprezintă matricea de covarianţă.

[5]vezi explicaţia de la pagina 172.

[6]vezi explicaţia de la pagina 123.

- principiul *clasificării ierarhice*,
- algoritmul "*k-means*",
- clasificarea pe bază de *grafuri*,
- clasificarea *pe bază de modele*,
- *clasificarea fuzzy*,
- abordările bazate pe *teoria evoluţiei*.

Clasificarea ierarhică

Având la dispoziţie un set de N obiecte şi o matrice de similaritate (distanţă) între acestea, de dimensiune $N \times N$, principiul clasificării ierarhice al acestora constă în următorul algoritm [Johnson 67]:

1. în primă fază, fiecare obiect este atribuit unei clase diferite, astfel iniţial vom dispune de N clase. Similaritatea dintre clase va fi dată în acest punct de similaritatea dintre obiectele claselor,
2. se caută cele mai similare două clase ce vor fi fuzionate într-una singură, astfel numărul de clase disponibile devenind $N - 1$,
3. se recalculează similaritatea dintre noua clasă şi fiecare dintre celelalte clase existente,
4. se repetă paşii 2 şi 3 până când toate obiectele se vor regăsi într-o clasă unică de dimensiune N.

În funcţie de metoda folosită, pasul 3 poate fi realizat în mai multe moduri. Astfel întâlnim: *înlănţuire simplă* ("simple-linkage"), *înlănţuire completă* ("complete-linkage") şi respectiv *înlănţuire medie* ("average-linkage").

În înlănţuirea simplă se consideră ca distanţă între două clase distanţa minimă dintre membrii unei clase la membrii celeilalte clase. Acest lucru se traduce prin aprecierea similarităţii între două clase ca fiind dată de cele mai similare două obiecte din aceste două clase.

Opus înlănţuirii simple, înlănţuirea completă, consideră ca distanţă între două clase, distanţa cea mai semnificativă dintre membrii unei clase la membrii celeilalte clase.

În cele din urmă, înlănţuirea medie consideră distanţa dintre două clase ca fiind media distanţelor dintre toţi membrii unei clase la toţi membrii celeilate clase. O altă versiune a înlănţuirii medii constă în alegerea distanţei ca fiind

valoarea mediană a valorilor de distanţă obţinute. Aceasta se dovedeşte în cele mai multe situaţii a fi mai eficientă decât valoarea medie, reducând numărul de clase atipice.

Clasificarea ierarhică este o metodă de tip acumulativ, deoarece clasele sunt fuzionate progresiv. Totuşi, clasificarea ierarhică poate fi implementată şi ca metodă partiţională prin inversarea algoritmului de clasificare. Astfel, se porneşte cu o singură clasă ce conţine toate obiectele, clasă ce este divizată progresiv în clase mai mici în funcţie de disimilaritatea obiectelor.

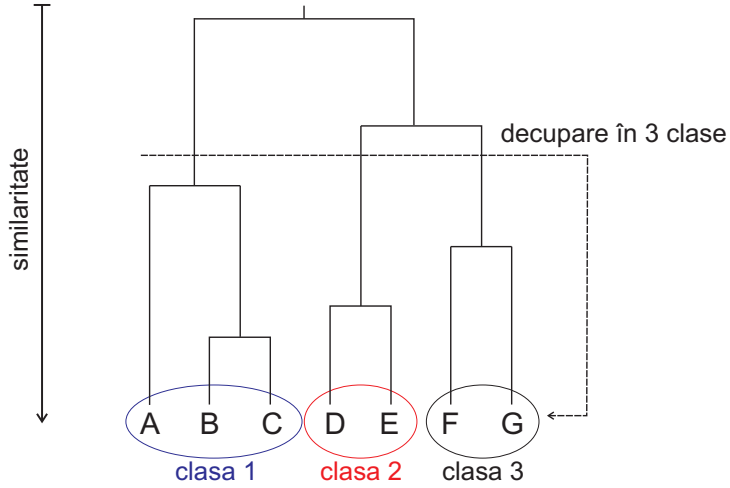


Figura 7.2: Exemplu de clasificare ierarhică, obiectele clasificate sunt notate cu litere de la A la G.

În raport cu celelalte metode existente, principalul avantaj al metodelor de clasificare ierarhică este dat de adaptabilitatea numărului de clase. În urma clasificării ierarhice, clasele obţinute pot fi reprezentate sub forma unei dendograme[7]. Un exemplu este prezentat în Figura 7.2. Astfel, în funcţie de cerinţele aplicaţiei, prin alegerea unui anumit prag de similaritate între clase, utilizatorul poate opta pentru o partiţie în clase ce variază de la o singură clasă, până la un număr de clase egal cu numărul de obiecte N. De notat este faptul că această repartiţie variabilă este disponibilă imediat fără a mai fi necesară relansarea algoritmului de clasificare.

Clasificarea K-means

Clasificarea k-means este una dintre cele mai populare metode de clasificare nesupervizată a datelor [MacQueen 67]. Acest lucru se datorează în mare

[7]vezi explicaţia de la pagina 70.

parte simplităţii şi eficienţei acesteia. Clasificarea k-means este folosită pentru a partaja un anumit număr de obiecte într-un număr de clase, k, ce este fixat "a priori".

Pentru aceasta, în primă fază se vor defini centrele celor k clase, numite şi centroizi. Modul în care sunt aleşi centroizii va influenţa rezultatul clasificării. Cea mai eficientă soluţie constă în alegerea acestora cât mai depărtaţi unii de alţii în spaţiul de caracteristici.

Următorul pas al clasificării constă în asocierea fiecărui obiect clasei cu centroidul cel mai apropiat de acesta. După ce fiecare obiect a fost asociat unei clase, poziţiile centroizilor vor fi recalculate în funcţie de membrii clasei, iar obiectele vor fi realocate în funcţie de distanţa faţă de noii centroizi. Pe măsură ce procesul este repetat, centroizii claselor îşi vor schimba succesiv poziţiile, până în mometul în care nu se va mai produce nici o schimbare. Repartiţia în clase astfel obţinută corespunde rezultatului final al clasificării.

Având în vedere că rezultatul clasificării depinde de modul în care au fost aleşi centroizii iniţiali, astfel că seturi diferite de centrozi produc clase diferite, în cele mai multe cazuri la finalul clasificării se calculează o anumită funcţie de cost, J. Clasificarea poate fi repetată de mai multe ori folosind seturi diferite de centroizi, optându-se în final pentru soluţia care minimizează valoarea lui J.

În cazul k-means, funcţia J este aleasă în sensul erorii pătratice, şi este definită astfel:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \tag{7.1}$$

unde k reprezintă numărul de clase, n reprezintă numărul de obiecte ce trebuiesc clasificate şi $||x_i^{(j)} - c_j||$ reprezintă o măsură de distanţă între obiectul $x_i^{(j)}$ al clasei j şi centroidul c_j al acesteia. Definită în acest fel, funcţia J este o măsură a distanţei globale a obiectelor faţă de centrul claselor. Cu cât valoarea lui J este mai importantă, cu atât clasele sunt mai neomogene.

Printre diversele variante ale clasificării k-means putem menţiona algoritmul ISODATA [Jain 99]. Algoritmul ISODATA încearcă să rezolve problema numărului fix de clase ce este impus în algoritmul k-means, prin reducerea claselor aşa zise redundante. Astfel că, atunci când unui centroid nu i se asociază un număr suficient de obiecte, acesta este vizat pentru a fi eliminat. În acest fel, numărul de clase obţinute va fi mai mult sau mai puţin optimal. Pentru ca selecţia să fie eficientă, numărul iniţial de clase k trebuie ales suficient de mare.

Raportat la celelalte metode de clasificare nesupervizată existente, clasificarea k-means este una dintre cele mai eficiente din punct de vedere al complexităţii de calcul şi al calităţii rezultatelor obţinute. De exemplu, pen-

tru o clasificare a 60 de obiecte în 5 clase, marea parte a metodelor enunţate anterior sunt de 500 de ori, până la 2500 de ori, mai lente decât algoritmul k-means [Jain 99]. Din acest motiv, clasificarea k-means şi corespondentul acesteia din categoria de metode ce folosesc reţele neuronale (reţelele neuronale Kohonen) au fost folosite cu succes la clasificarea volumelor foarte vaste de date. Celelalte tehnici de clasificare mai complexe sunt de regulă folosite pentru volume de date mai restrânse din cauza timpului de execuţie ridicat.

Clasificarea pe bază de grafuri

Din categoria metodelor de clasificare ce folosesc teoria grafurilor[8], algoritmul cel mai cunoscut este bazat pe construcţia arborelui minimal, numit şi "*Minimum Spanning Tree*" sau MST.

Având la dispoziţie un graf nedirecţional, definim ca arbore al acestuia, subgraful ce conectează toate muchiile[9] grafului. Un anumit graf poate avea astfel mai mulţi arbori. De asemenea, fiecărei muchii a grafului i se asociază o anumită pondere ce este aleasă ca fiind proporţională cu costul parcurgerii acesteia în graf. În aceste condiţii, un arbore minimal este definit ca fiind arborele cu ponderea totală cea mai mică sau cel mult egală cu ponderea unui alt arbore posibil, unde ponderea unui arbore este dată de suma ponderilor muchiilor acestuia.

Reprezentând obiectele ce trebuie clasificate sub forma unui graf în spaţiul de caracteristici, unde ponderea muchiilor poate fi dată de o anumită măsură de distanţă între noduri, atunci putem construi arborele minimal MST al acestuia. O strategie de clasificare pe baza MST constă mai departe în eliminarea progresivă a segmentelor arborelui cu lungimea cea mai semnificativă. Un exemplu este prezentat în Figura 7.3 [Jain 99], în care am ilustrat arborele minimal obţinut pentru 9 obiecte (notate cu litere de la A la I), ce sunt reprezentate într-un spaţiu de caracteristici bidimensional. Astfel, prin ruperea legăturii cu ponderea cea mai importantă, şi anume segmentul CD (pondere 6) vom obţine două clase, clasa c_1 ce conţine obiectele $\{A, B, C\}$ şi respectiv clasa c_2 ce conţine obiectele $\{D, E, F, G, H, I\}$. Mai departe, prin ruperea legăturii EF (pondere 4.5), clasa c_2 poate fi divizată în două subclase, şi anume: $c_{21} = \{D, E\}$ şi $c_{22} = \{F, G, H, I\}$. Procesul poate continua astfel până la obţinerea granularităţii (numărului de clase) dorite.

Principalul avantaj al metodelor de clasificare pe bază de grafuri este dat de faptul că o clasificare complexă multidimensională este redusă la o

[8] vezi explicaţia de la pagina 159.

[9] într-un graf, o muchie este dată de o pereche de noduri. Dacă graful este orientat, atunci aceasta se numeşte arc.

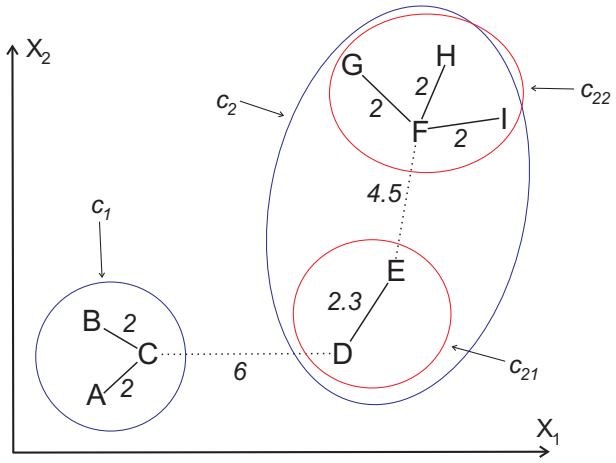


Figura 7.3: Exemplu de clasificare pe baza construcţiei arborelui minimal MST ($X_1 o X_2$ reprezintă spaţiul de caracteristici iar numerele asociate muchiilor reprezintă ponderea acestora).

problemă de partiţionare a unui arbore bidimensional, fară a pierde însă din informaţiile esenţiale pentru clasificare. Un alt avantaj important este dat de faptul că clasificarea pe bază de grafuri nu depinde de forma geometrică a claselor. În ultimă instanţă, aceasta poate fi folosită ca etapă intermediară la implementarea altor tehnici de clasificare mai riguroase, dar care necesită un timp de calcul important.

Clasificarea pe bază de modele statistice

O altă abordare a problematicii clasificării datelor constă în modelarea statistică a claselor şi încercarea de optimizare a modelului obţinut, relativ la datele reale. Astfel, clasificarea pe bază de modele foloseşte reprezentarea matematică a fiecărei clase cu o anumită distribuţie parametrică, ca de exemplu, distribuţia Gaussiană[10] (continuă) sau Poisson[11] (discretă). În acest fel, întregul set de date va fi modelat de un amestec de astfel de distribuţii ce

[10]distribuţia Gaussiană sau normală a unei variabile aleatoare x este dată de funcţia de densitate de probabilitate continuă: $\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-(x-\mu)^2/2 \cdot \sigma^2}$, unde μ reprezintă valoarea medie a lui x iar σ^2 reprezintă varianţa.

[11]distribuţia Poisson este o distribuţie de probabilitate discretă ce exprimă probabilitatea ca o serie de evenimente să aibă loc într-o anumită perioadă de timp, dacă aceste evenimente au o rată de apariţie medie cunoscută şi apar independent de timpul scurs de la ultimul eveniment. Probabilitatea ca un eveniment să apară de k ori într-un interval de timp pentru care valoarea medie de apariţii este λ, este dată de: $f(k, \lambda) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$.

constituie şi modelul datelor.

Principiul matematic este următorul [Fraley 96]: având la dispoziţie o populaţie de date ce conţine un număr G de sub-populaţii diferite (potenţiale clase), atunci densitatea de probabilitate a observaţiei p-dimensionale x, din sub-populaţia de indice k, este dată de funcţia $f_k(x;\theta)$, unde θ reprezintă un anumit vector de parametri necunoscut. Având la dispoziţie observaţiile $x = \{x_1, ..., x_n\}$ ce reprezintă datele ce trebuie clasificate, se defineşte vectorul de etichete al claselor ca fiind:

$$\gamma = (\gamma_1, ..., \gamma_n)^T \tag{7.2}$$

unde $\gamma_i = k$ dacă x_i aparţine sub-populaţiei de indice k.

Pentru clasificare, parametrii θ şi etichetele γ sunt astfel aleşi încât să maximizeze o anumită funcţie de plauzibilitate ("likelihood function"), şi anume:

$$L(x;\theta,\gamma) = \prod_{i=1}^{G} f_{\gamma_i}(x;\theta) \tag{7.3}$$

De cele mai multe ori, ca distribuţie este aleasă distribuţia normală sau Gaussiană multidimensională, parametrizată de vectorul de medii μ_k şi matricea de varianţă Σ_k. Astfel, vectorul de parametri θ va fi dat de parametrii distribuţiilor pentru fiecare sub-populaţie, $\theta = (\mu_1, ..., \mu_G, \Sigma_1, ..., \Sigma_G)$.

Înlocuind în relaţia anterioară, obţinem funcţia de probabilitate următoare:

$$L(x;\mu_1, ..., \mu_G, \Sigma_1, ..., \Sigma_G, \gamma) =$$
$$\prod_{k=1}^{G} \prod_{i \in I_k} (2\pi)^{-\frac{p}{2}} |\Sigma_k|^{-\frac{1}{2}} exp\left[-\frac{1}{2}(x_i - \mu_i)^T \Sigma_k^{-1}(x_i - \mu_k)\right] \tag{7.4}$$

unde I_k reprezintă setul de indici ce corespund observaţiilor din grupul de indice k, $I_k = \{i/\gamma_i = k\}$.

Simplificări ale funcţiei de plauzibilitate $L()$ pot fi obţinute prin înlocuirea parametrilor distribuţiei cu anumiţi estimatori, ca de exemplu:

$$\mu_k = \hat{\mu}_k = \frac{\sum_{i \in I_k} x_i}{n_k} \tag{7.5}$$

unde n_k reprezintă numărul de elemente al lui I_k, substituţie ce conduce la o expresie logaritmică, astfel:

$$L(x;\hat{\mu}_1, ..., \hat{\mu}_G, \Sigma_1, ..., \Sigma_G, \gamma) =$$
$$-\frac{pnlog(2\pi)}{2} - \frac{1}{2}\sum_{k=1}^{G}\left[tr(W_k\Sigma_k^{-1}) + n_k log|\Sigma_k|\right] \tag{7.6}$$

unde $W_k = \sum_{i \in I_k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$ iar $tr(A)$ ("trace") reprezintă operatorul ce returnează suma elementelor de pe diagonala principală a matricei A.

Totuşi, chiar şi în această situaţie, maximizarea funcţiei $L()$ poate fi dificilă, având o complexitate de calcul ridicată. Din acest motiv, în practică, se preferă o metodă derivată şi anume algoritmul EM - "Expectation-Maximization" [Dempster 77].

Datorită faptului că în momentul clasificării datelor, nu se cunosc exact, nici distribuţia acestora şi nici parametrii modelului mixt, similar cu algoritmul k-means (vezi pagina 208), algoritmul EM se foloseşte de iterare. Astfel, clasificarea porneşte cu un set de parametri aleşi arbitrar. Valorile acestora sunt folosite pentru estimarea probabilităţilor claselor, iar mai departe, aceste probabilităţi sunt folosite pentru re-estimarea parametrilor, procesul repetându-se în mod iterativ.

Prima etapă de calcul a probabilităţilor se numeşte "expectation", deoarece acestea reprezintă valorile *aşteptate* ale claselor. A doua etapă ce constă în estimarea parametrilor modelului poartă numele de "maximization" şi constă în *maximizarea* funcţiei de plauzibilitate.

Spre deosebire de algoritmul k-means, în care iterarea se opreşte atunci când clasele nu se mai modifică, în cazul algoritmului EM convergenţa este mai dificilă. Teoretic, algoritmul converge către un punct fix dar practic nu ajunge niciodată acolo. Totuşi, se poate estima cât de aproape suntem de convergenţă prin calcularea unei funcţii globale de plauzibilitate pentru ansamblul datelor folosite, prin înmulţirea probabilităţilor pentru fiecare iteraţie.

În practică, se preferă calculul logaritmului acestei funcţii. Decizia de oprire a iteraţiei se ia atunci când pe parcursul a n iteraţii diferenţa dintre valorile funcţiei de plauzibilitate este inferioară valorii 10^{-d}, un exemplu de astfel de valori fiind $n = 10$ şi $d = 10$ [Witten 05].

Global, metodele de clasificare din această categorie prezintă o serie de avantaje. În primul rând, folosind drept criteriu de convergenţă maximizarea funcţiei de plauzibilitate, acesta duce la obţinerea de clase compacte ce vor îngloba tipurile de date dominante (furnizate de distribuţia componentelor modelului mixt). Un alt avantaj al folosirii modelării statistice a claselor constă în baza teoretică foarte avansată şi bine instrumentată disponibilă în acest domeniu, şi anume al inferenţei statistice. Acest lucru oferă de asemenea şi o flexibilitate în alegerea distribuţiei statistice adaptate datelor ce trebuie clasificate.

Pe de altă parte, convergenţa unei astfel de abordări este complexă, atât din punct de vedere al modelării matematice, cât şi al complexităţii de calcul.

Classificarea fuzzy

Dacă metodele de clasificare tradiţionale furnizează o partiţionare netă a datelor în clase, astfel, fiecare obiect aparţinând unei singure clase, iar clasele fiind disjuncte, *clasificarea fuzzy* se foloseşte de teoria mulţimilor fuzzy pentru a introduce conceptul de incertitudine (vezi Capitolul 6). În acest fel, problema apartenenţei la clase nu mai este o problemă de afirmare sau negare, ci o problemă de grad de apartenenţă. Fiecare obiect va aparţine astfel unei clase într-un anumit grad, ce este furnizat de coeficientul fuzzy.

Principiul este ilustrat în Figura 7.4, în care datele ce trebuie clasificate au fost reprezentate cu cifre de la 1 la 9. Astfel, în cazul unei clasificări nete, clasele obţinute sunt de tipul claselor disjuncte N_1 şi N_2, în timp ce clasele obţinute cu o clasificare fuzzy, F_1 şi F_2, se întrepătrund.

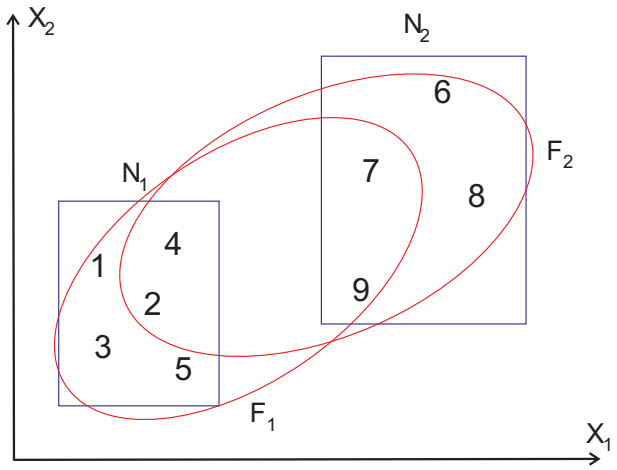


Figura 7.4: Principiul clasificării fuzzy (X_1oX_2 reprezintă spaţiul de caracteristici, dreptunghiurile reprezintă clasele nete, în timp ce elipsele marchează clasele fuzzy).

Dintre algoritmii de clasificare fuzzy cei mai populari, putem menţiona algoritmul *Fuzzy C-Means* sau FCM [Dunn 73]. Acesta poate fi considerat ca fiind echivalentul în logică fuzzy al algoritmului k-means (vezi pagina 208). FCM se bazează pe minimizarea unei funcţii obiective, J_m, ce este definită în felul următor:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m ||x_i - c_j||^2 \tag{7.7}$$

unde m este un număr real, $m \geq 1$, u_{ij} reprezintă gradul de apartenenţă al lui x_i la clasa de indice j, $u_{ij} \in [0; 1]$ (valoarea 1 indică apartenenţa sigură, iar

valoarea 0 negaţia sigură), x_i reprezintă data de indice i din setul de N date ce trebuie clasificate, c_j reprezintă centrul clasei j, C reprezintă numărul de clase iar $||.||$ este o anumită normă ce exprimă similaritatea dintre x_i şi c_j.

Partiţionarea fuzzy este realizată prin optimizarea iterativă a funcţiei J_m, timp în care funcţiile de apartenenţă, u_{ij}, şi centrele claselor, c_j, sunt recalculate progresiv, astfel:

$$u_{ij} = \left[\sum_{l=1}^{C} \left(\frac{||x_i - c_j||}{||x_i - c_l||} \right)^{\frac{2}{m-1}} \right]^{-1} \tag{7.8}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_j}{\sum_{i=1}^{N} u_{ij}^m} \tag{7.9}$$

Iterarea se va opri atunci când condiţia următoare este satisfăcută:

$$max_{i,j}\{|u_{ij}^{(k+1)} - u_{ij}^{(k)}|\} < \epsilon \tag{7.10}$$

unde k reprezintă indicele iteraţiei iar ϵ este un anumit prag ales între 0 şi 1. În acest fel, algoritmul va converge către un minim sau o valoare fixă a funcţiei J_m.

Astfel, algoritmul de clasificare FCM poate fi sintetizat în felul următor:

1. în primă fază este iniţializată matricea $U = U^{(0)}$, matrice ce conţine gradele de apartenenţă fuzzy la clase, $U = [u_{ij}]$, iar indicele iteraţiei devine $k = 0$,

2. pe baza matricei $U^{(k)}$, se calculează centrele claselor, $C^{(k)} = [c_j]$, folosind relaţia 7.9,

3. mai departe, matricea de la iteraţia $k + 1$, $U^{(k+1)}$, este calculată prin reactualizarea valorilor matricei $U^{(k)}$ pe baza relaţiei 7.8,

4. se testează condiţia de oprire şi anume dacă $||U^{(k+1)} - U^{(k)}|| < \epsilon$, caz în care clasificarea se opreşte. În caz contrar, algoritmul trece la iteraţia următoare, $k \leftarrow k + 1$, şi acesta se reia de la punctul 2.

Raportat la algoritmul echivalent în logică booleană, şi anume algoritmul k-means, algoritmul FCM converge rapid către un minim local. Principalul dezavantaj al clasificării fuzzy constă totuşi în dificultatea de alegere adecvată a funcţiilor de apartenenţă fuzzy, alegere care de cele mai multe ori este determinantă pentru calitatea clasificării.

Clasificarea bazată pe teoria evoluţiei

Abordările bazate pe teoria evoluţiei sunt inspirate de *principiul evoluţiei naturale a lumii vii*. Principiul care stă la baza acestora, constă în aplicarea, asupra unei anumite populaţii de soluţii, a o serie de operatori genetici specifici. În acest fel, populaţia tinde să evolueze spre o soluţie global optimală de partiţionare a datelor [Jain 99].

Potenţialele soluţii ale problemei de clasificare sunt codate folosind cromozomi[12]. Dintre operatorii genetici cei mai frecvent utilizaţi, putem menţiona: *selecţia*, *recombinarea* şi *mutaţia*. Fiecare dintre aceştia transformă unul sau mai mulţi cromozomi de intrare în unul sau mai mulţi cromozomi de ieşire. De asemenea, pentru fiecare cromozom se evaluează o anumită funcţie obiectivă ce determină probabilitatea de supravieţuire a acestuia în următoarea generaţie.

Astfel, un algoritm generic de clasificare pe baza teoriei evoluţiei poate fi enunţat în felul următor [Jain 99]:

1. în primă fază se alege aleator o populaţie de soluţii, fiecare dintre soluţii corespunzând unei anumite partiţii valide a datelor. Fiecărei soluţii i se asociază o funcţie obiectivă, de regulă aleasă în aşa fel încât să fie invers proporţională cu eroarea pătratică. Astfel, o soluţie cu o eroare pătratică mică, va corespunde unei valori ridicate a funcţiei obiective,

2. folosind operatorii de selecţie, recombinare şi mutaţie, se generează o nouă populaţie de soluţii. Se re-evaluează funcţia obiectivă a acestei noi populaţii,

3. se repetă pasul 2 până când este îndeplinit un anumit criteriu de convergenţă.

Dintre tehnicile bazate pe teoria evoluţiei cele mai cunoscute, putem enumera *algoritmii genetici* (GAs), *strategiile evoluţioniste* (ESs) şi *programarea evoluţionistă* (EP) [Jain 99]. Dintre acestea, algoritmii genetici sunt folosiţi cel mai frecvent în probleme de clasificare. În GAs, soluţiile sunt exprimate binar iar propagarea de la o generaţie la alta este realizată pe baza funcţiei obiective prin intermediul operatorului de selecţie. Selecţia se foloseşte de o strategie probabilistică pentru a desemna soluţiile cu o valoare semnificativă a funcţiei obiective ca soluţii cu o probabilitate importantă de reproducere. Operatorul de mutaţie este folosit în GAs doar pentru a se asigura că întreg

[12] cromozomul, în limba greacă chromo-nuanţă şi soma-obiect, reprezintă macromolecule de ADN care conţin mai multe gene şi secvenţe nucleotide cu rol în păstrarea informaţiei ereditare a celulei.

spaţiul soluţiilor este explorat. Spaţiul soluţiilor este explorat folosind operatorul de recombinare ("crossover"). Pe de altă parte, ESs şi EP, diferă de GAs, prin modul de reprezentare al soluţiilor şi prin operatorii de mutaţie folosiţi. De exemplu, în EP nu se foloseşte operatorul de recombinare, ci doar cei de mutaţie şi selecţie.

Global, abordările bazate pe teoria evoluţiei sunt tehnici de căutare a unei soluţii globale a problemei clasificării, ceea ce le diferenţiază de marea parte a celorlalte tehnici existente ce propun de regulă o căutare locală a soluţiei. Mai mult, abordările bazate pe teoria evoluţiei sunt capabile să găsească soluţia optimală chiar şi în cazul în care funcţia asociată criterilor de selecţie este discontinuă.

7.2 Clasificarea supervizată

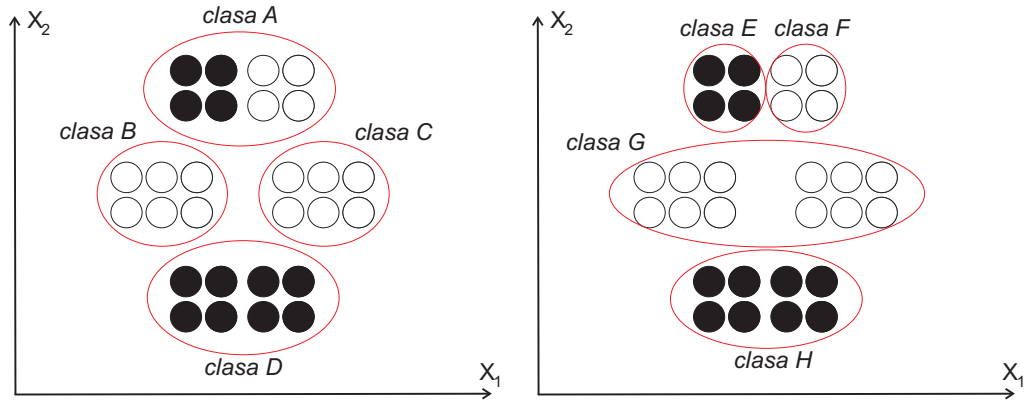
După cum am menţionat şi în partea introductivă a acestui capitol, diferenţa dintre clasificarea nesupervizată şi cea supervizată constă în faptul că aceasta din urmă se foloseşte de o serie de exemple de clasificări (date de antrenare), de regulă obţinute în urma expertizei domeniului de aplicaţie[13], pentru a identifica clasele ce prezintă o densitate de probabilitate importantă relativ la o singură clasă de referinţă. Mai mult, clasificarea supervizată, tinde să păstreze un număr de clase relativ scăzut [Eick 04].

Din punct de vedere al modului în care obiectele sunt asociate claselor, în clasificarea supervizată se foloseşte noţiunea de "apropiere" în sensul unei anumite măsuri de distanţă, spre deosebire de clasificarea nesupervizată în care de regulă se calculează o anumită eroare de clasificare ce minimizează distanţa dintre obiectele din interiorul unei clase.

Pentru a înţelege mai bine diferenţa dintre cele două tipuri de abordări, în Figura 7.5 am ilustrat exemplul următor [Eick 04]: să presupunem că obiectele ce trebuie clasificate reprezintă subspecii ale plantei din familia Iris, şi anume Setosa (culoare neagră) şi Virginica (culoare albă). În urma unei clasificări nesupervizate, este foarte probabil ca obiectele să fie împărţite în patru clase, aşa cum am ilustrat în Figura 7.5.a. Acest lucru se datorează faptului că, clasificarea nesupervizată ţine cont în primul rând de minimizarea distanţei obiectelor în spaţiul de caracteristici. Astfel, clasificarea obţinută nu este foarte interesantă din punct de vedere obiectiv deoarece clasele obţinute, fie combină cele două subspecii, fie separă aceeaşi specie în clase diferite. De

[13] datele de antrenare pot fi văzute în sensul unei realităţi de teren ("groundtruth") folosite de regulă pentru validarea rezultatelor anumitor algoritmi de detecţie (vezi şi explicaţia de la pagina 170).

exemplu, clasa A conţine atât subspecia Virginica cât şi Setosa, iar clasele B şi C conţin aceeaşi subspecie Virginica.



(a) clasificare nesupervizată (b) clasificare supervizată

Figura 7.5: Diferenţa dintre clasificarea nesupervizată şi clasificarea supervizată ($X_2 o X_1$ reprezintă spaţiul de caracteristici, obiectele de acelaşi tip sunt figurate cu aceeaşi culoare) [Eick 04].

Pe de altă parte, o metodă de clasificare supervizată ce are ca scop maximizarea "purităţii" claselor, va diviza clasa A obţinută anterior în două clase, şi anume: clasa E şi respectiv F (vezi Figura 7.5.b). De asemenea, în scopul reducerii numărului de clase fără însă a compromite omogenitatea acestora, clasele B şi C obţinute anterior vor fuziona într-o singură clasă G.

Global, clasele obţinute vor fi pe cât posibil omogene, conţinând obiecte de acelaşi tip, lucru ce se datorează în principal antrenării prealabile a clasificatorului cu exemple din fiecare clasă.

7.2.1 Etapele clasificării supervizate

În general, un clasificator supervizat implică introducerea noţiunii de învăţare sau "machine learning". Aceasta presupune adaptarea algoritmului la domeniul de aplicaţie şi implică mai multe etape de prelucrare. În Figura 7.6 am sintetizat diagrama de funcţionare generală a unui clasificator supervizat [Kotsiantis 07].

Astfel, prima etapă constă în identificarea setului de date ce trebuie clasificat. Tot în această etapă sunt alese şi atributele ce vor fi folosite pentru definirea spaţiului de caracteristici în care se va efectua clasificarea propriu-zisă. Acestea trebuie alese în aşa fel încât să fie cât mai reprezentative pentru criterile de selecţie vizate.

A doua etapă constă în pre-prelucrarea acestor date. De regulă operaţiile efectuate constau în reducerea zgomotului existent în date cât şi simplificarea procesului de manevrare a seturilor voluminoase de date de antrenare ("instance selection"), sau reducerea dimensiunii setului de date prin eliminarea datelor ce sunt irelevante sau redundante pentru clasificare ("feature subset selection").

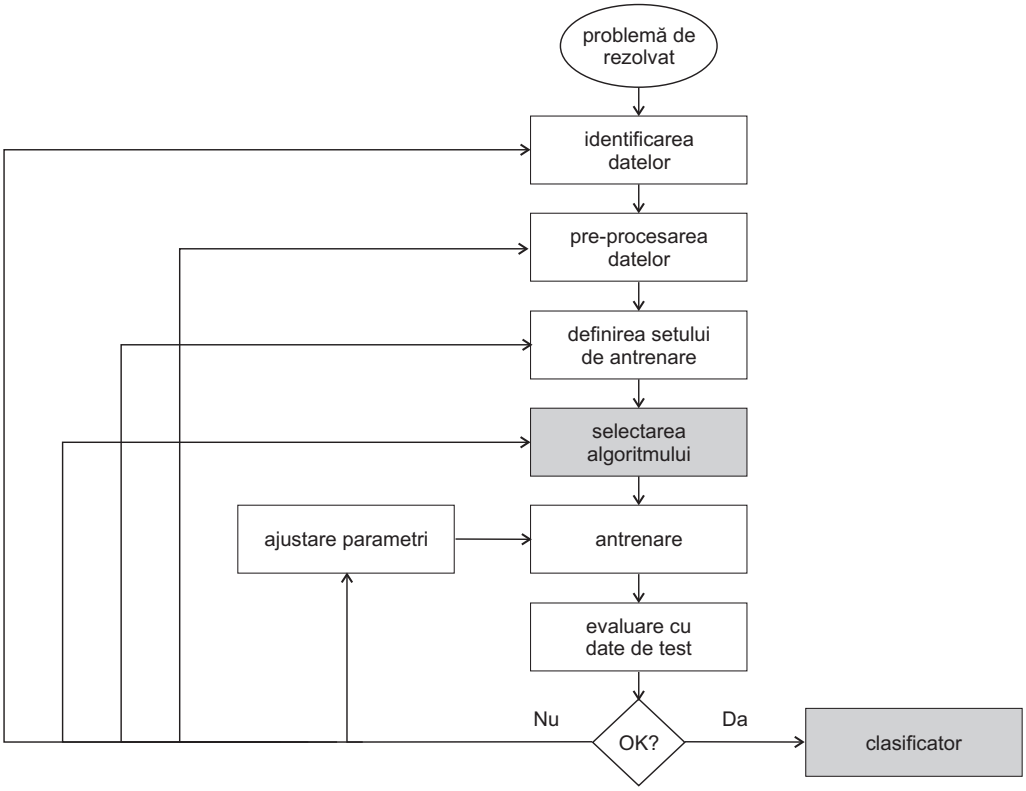


Figura 7.6: Procesul de clasificare supervizată [Kotsiantis 07].

După ce datele de antrenare au fost definite, acestea fiind de regulă disponibile "a priori" clasificării, se trece la alegerea algoritmului de clasificare supervizată ce corespune cel mai bine aplicaţiei vizate. Acest pas este foarte important deoarece influenţează decisiv rezultatele clasificării, algoritmi diferiţii furnizând rezultate uneori foarte diferite. Din această cauză se preferă testarea rezultatelor clasificării folosind setul de antrenare. În momentul în care clasificatorul testat este considerat ca fiind satisfăcător pentru aplicaţie, de abia atunci se poate trece la folosirea acestuia cu datele de intrare ne-etichetate (datele propriu-zise).

Evaluarea clasificatorului constă de regulă în estimarea preciziei predicţiei,

aceasta fiind definită ca raportul dintre numărul de predicţii corecte şi numărul total de predicţii[14].

Dintre metodele existente, menţionăm trei dintre cele mai frecvent folosite. O primă abordare constă în divizarea setului de date de antrenare în trei părţi, dintre care două sunt folosite pentru antrenare iar una dintre acestea pentru estimarea preciziei predicţiei clasificatorului. O altă abordare, cunoscută şi sub numele de "validare încrucişată" ("cross-validation"), constă în împărţirea setului de antrenare în subseturi disjuncte şi egale. Pentru fiecare subset, clasificatorul va fi antrenat folosind setul format prin reuniunea tuturor celorlalte subseturi existente, iar evaluarea se va realiza pe acesta. Eroarea clasificatorului va fi dată de media tuturor erorilor obţinute pentru fiecare subset în parte. Un caz particular al validării încrucişate este algoritmul "leave-one-out". În acest caz, fiecare subset de test va fi constituit dintr-o singură instanţă. Această metodă este cea mai complexă din punct de vedere al calculelor, dar pe de altă parte şi cea mai precisă din punct de vedere al estimării erorii de predicţie a clasificatorului.

În cazul în care eroarea de predicţie nu este suficient de mică, atunci etapele anterioare trebuie revizate. Mai mulţi factori pot fi responsabili pentru aceasta, ca de exemplu: este posibil să nu fi ales atributele cele mai relevante, setul de antrenare să fie prea redus, complexitatea problemei să fie prea ridicată, clasificatorul ales să nu fie cel mai adaptat problemei vizate, parametrii acestuia să fie reglaţi incorect, şi aşa mai departe.

7.2.2 Metodele existente de clasificare supervizată

Clasificarea supervizată este realizată cu predilecţie de ceea ce numim Sisteme Inteligente. Dintre tehnicile existente, ne vom focaliza în continuare pe prezentarea celor mai semnificative dintre acestea, şi anume:

- *algoritmi logici* (arbori de decizie, seturi de reguli),
- *algoritmi bazaţi pe perceptroni* (uni-strat, reţele neuronale),
- *algoritmi statistici* (clasificatori Bayes, reţele Bayes),
- *algoritmi bazaţi pe instanţe* (kNN),
- algoritmi de tip *"Support Vector Machine"*.

[14]termenul de predicţie este folosit în acest caz pentru a desemna modul în care clasificatorul etichetează datele de antrenare. Dacă etichetele asociate de clasificator corespund etichetelor deja existente, atunci predicţia este corectă. În caz contrar, aceasta este eronată.

Algoritmi logici

În această categorie de metode de clasificare supervizată putem menţiona construcţia *arborilor de decizie* şi a *regulilor de decizie*.

Un **arbore de decizie** ("decision tree") [Yuan 95] este un arbore ce clasifică anumite instanţe ale datelor de intrare pe baza trierii lor în funcţie de valorile atributelor acestora. Fiecare nod al arborelui reprezintă un atribut al unei anumite instanţe ce trebuie clasificată. Fiecare ramură a arborelui reprezintă o anumită valoare pe care nodul şi-o poate atribui. Instanţele datelor de intrare sunt clasificate pornind de la rădăcina arborelui, fiind sortate progresiv în funcţie de valorile atributelor. În Figura 7.7 am exemplificat arborele de decizie asociat datelor de antrenare din Tabelul 7.1 [Kotsiantis 07].

Atributul 1	Atributul 2	Atributul 3	Atributul 4	Clasa nr.
$A1$	$A2$	$A3$	$A4$	1
$A1$	$A2$	$A3$	$B4$	1
$A1$	$B2$	$A3$	$A4$	2
$A1$	$B2$	$B3$	$B4$	3
$A1$	$C2$	$A3$	$A4$	4
$A1$	$C2$	$A3$	$B4$	5
$B1$	$B2$	$B3$	$B4$	2
$C1$	$B2$	$B3$	$B4$	6

Tabelul 7.1: Exemplu de date de antrenare în 6 clase (valorile atributelor sunt reprezentate cu litere majuscule de la A la C).

Astfel, instanţa $\{$*Atributul 1*$= A1$, *Atributul 2*$= B2$, *Atributul 3*$= A3$, *Atributul 4*$= B4\}$ va parcurge arborele prin nodurile *Atributul 1*, *Atributul 2* şi respectiv *Atributul 3*, ceea ce duce la clasificarea instanţei ca aparţinând clasei 2.

Una dintre problemele unei astfel de abordări de clasificare este dată de adaptarea arborelui strict la datele de antrenare, ceea ce va duce la o posibilă clasificare eronată a datelor reale. Acest fenomen poartă numele de "overfitting". Fiind dat un spaţiu de ipoteze H, o ipoteză $h \in H$ (de exemplu, un arbore de decizie) este considerată că se "supra-adaptează" datelor de antrenare dacă există o altă ipoteză alternativă $h' \in H$, iar h obţine o eroare inferioară lui h' pe datele de antrenare, dar h' duce la o eroare mai mică decât h pentru întreaga distribuţie de instanţe de intrare [Venkataraman 99].

De regulă metodele folosite pentru combaterea acestui fenomen constau, fie în *reducerea algoritmului de antrenare* prin folosirea doar a unui subset de date de antrenare, sau folosirea în totalitate a acestora şi *"rafinarea"*

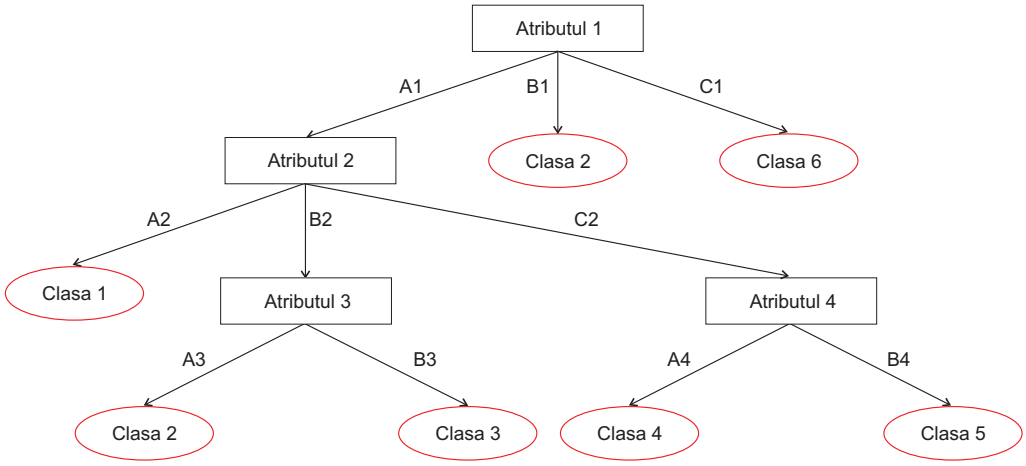


Figura 7.7: Arborele de decizie asociat datelor de antrenare din Tabelul 7.1.

ulterioară a arborelui de decizie obţinut prin eliminarea anumitor ramuri sau frunze[15] ("pruning"). De regulă, dacă doi arbori de decizie diferiţi, testaţi pe aceleaşi date, obţin aceeaşi precizie a predicţiei, atunci se preferă arborele ce conţine cele mai puţine frunze.

Spre deosebire de alte metode, din punct de vedere al complexităţii calculelor, arborii de decizie au avantajul de a fi în general uni-variabilă, deoarece divizarea ramurilor este realizată prin analiza unui singur atribut. Totuşi, din această cauză, arborii de decizie se limitează la a furniza o partiţionare a spaţiului de intrare ortogonală pe una dintre axele spaţiului de caracteristici şi paralelă cu toate celelalte. Astfel, clasele rezultate în urma partiţionării sunt hiperdreptunghiuri[16].

O altă metodă de clasificare supervizată din această categorie sunt **regulile de decizie**. Acestea pot fi obţinute indirect pe baza arborilor de decizie, prin asocierea unei reguli pentru fiecare cale de parcurgere a arborelui de la rădăcină până la o frunză, sau pot fi induse direct pe baza datelor de antrenare (vezi studiul prezentat în [Furnkraz 99]).

Algoritmi bazaţi pe perceptroni

Perceptronul [Rosenblatt 62] poate fi considerat ca fiind cel mai simplu exemplu de reţea neuronală. Acesta este un clasificator liniar. În versiunea simplificată, uni-strat, perceptronul nu are decât o singură ieşire la care sunt

[15] termenul de "frunză" a unui arbore este folosit pentru a desemna un nod terminal.

[16] în geometrie, un hiperdreptunghi reprezintă generalizarea noţiunii de dreptunghi pentru dimensiuni de ordin superior lui 2, fiind definit ca un produs Cartezian de intervale.

conectate toate intrările (vezi Figura 7.8).

Principiul de funcţionare este următorul: dacă x_i, cu $i = 1, ..., n$, reprezintă valorile atributelor de intrare, iar w_i sunt nişte ponderi asociate conexiunilor de intrare (de regulă alese ca fiind numere reale în intervalul $[-1; 1]$, numite şi vector de predicţie), atunci perceptronul va calcula suma ponderată a intrărilor:

$$y = \sum_{i=1}^{n} x_i \cdot w_i \tag{7.11}$$

iar ieşirea acestuia va fi dată de valoarea binară ce rezultă din filtrarea cu un anumit prag τ a valorii y obţinute. Astfel, dacă $y > \tau$ valoarea de ieşire este 1, iar în caz contrar se returnează valoarea 0.

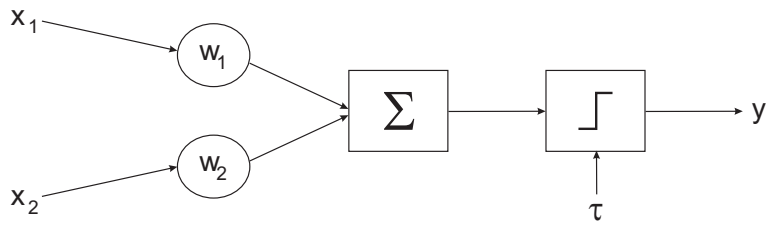


Figura 7.8: Exemplu de perceptron uni-strat cu două intrări (x reprezintă intrările, w ponderile acestora, τ pragul folosit iar y ieşirea perceptronului).

Modalitatea cea mai simplă de antrenare a perceptronului pentru un anumit set de date de antrenare, constă în rularea repetitivă a algoritmului prezentat anterior pentru acest set de date, până când se obţine un vector de predicţie care este corect pentru toate datele de antrenare.

Un alt algoritm de antrenare mai elaborat este metoda WINNOW. Aceasta modifică ponderile conexiunilor în felul următor: dacă valoarea predicţiei la un moment dat este $y' = 0$ iar valoarea actuală este $y = 1$, atunci înseamnă că ponderile sunt valori prea mici. În acest caz, ponderile vor fi mărite folosind următoarea relaţie:

$$w_i = \alpha \cdot w_i \tag{7.12}$$

unde $\alpha > 1$ reprezintă parametrul de creştere a ponderii ("promotion parameter").

Dacă valoarea predicţiei este $y' = 1$ iar valoarea actuală este $y = 0$, atunci înseamnă că ponderile au valori prea ridicate. În acest caz, se recurge la diminuarea acestora, astfel:

$$w_i = \beta \cdot w_i \tag{7.13}$$

unde $0 < \beta < 1$ reprezintă parametrul de diminuare a ponderii ("demotion parameter").

Perceptronii sunt clasificatori binari ce pot fi folosiţi pentru a clasifica doar seturi liniar separabile de date de intrare. Astfel, aceştia oferă doar o partajare liniară a spaţiului de intrare. În cazul în care datele de intrare nu sunt liniar separabile, se poate opta pentru perceptroni multi-strat sau ceea ce numim reţele neuronale artificiale (ANN - Artificial Neural Networks) [Rumelhart 86].

O reţea neuronală este constituită dintr-un număr mare de unităţi de prelucrare, numite şi neuroni, ce sunt interconectate pe straturi pentru a forma o anumită structură complexă de prelucrare. Un exemplu este ilustrat în Figura 7.9. De regulă, neuronii unei reţele neuronale sunt de trei tipuri:

- *neuroni de intrare* ce primesc informaţia ce trebuie prelucrată,
- *neuroni ascunşi* ce efectuează anumite etape intermediare de prelucrare şi înregistrează datele,
- *neuroni de ieşire* ce furnizează rezultatele obţinute.

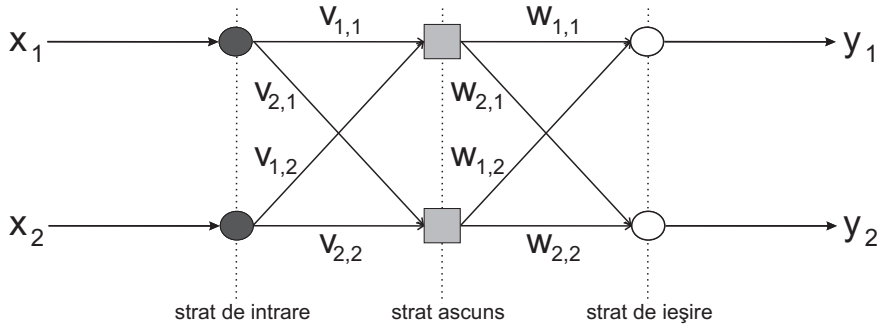


Figura 7.9: Exemplu de reţea neuronală cu două intrări şi două ieşiri (x reprezintă intrările, v şi w reprezintă ponderile conexiunilor iar y reprezintă ieşirile).

Un neuron primeşte o serie de valori de intrare, fie de la o sursă externă, fie de la un alt neuron. Ca şi în cazul perceptronului, fiecare valoare de intrare are asociată o pondere care poate fi modificată astfel încât acesta să se adapteze modelului de învăţare. Valorile de ieşire ale neuronului pot să devină valori de intrare pentru alţi neuroni.

Dacă consideram că x_j, cu $j = 1, ..., m$, reprezintă valorile de intrare ale neuronului i, iar w_{ij} reprezintă ponderea legăturii de la neuronul j la neuronul

i, atunci ieşirea neuronului i este dată de ecuaţia următoare:

$$y_i = f(\sum_{j=1}^{m} w_{ij} \cdot x_j) \tag{7.14}$$

unde $f()$ reprezintă funcţia de activare a neuronului iar suma $\sum_{j=1}^{m} w_{ij} \cdot x_j$ poartă numele de intrare a reţelei şi este de regulă notată cu net_i, astfel:

$$y_i = f(net_i) \tag{7.15}$$

Antrenarea unei reţele neuronale se realizează simplu prin folosirea datelor de antrenare pentru a modifica ponderile reţelei. Datele de antrenare sunt de regulă perechi de intrări şi ieşiri ale reţelei. Ponderile obţinute sunt stocate şi folosite apoi pentru clasificarea datelor reale.

De menţionat este faptul că reţelele neuronale pot exista şi în varianta nesupervizată, cum ar fi de exemplu reţelele cu auto-organizare de tip Kohonen. Dintre reţelele neuronale existente, menţionăm pe cele mai importante, şi anume: reţelele RBF (Radial Basis Function) în care funcţiile de activare sunt funcţii radiale, reţelele recurente RNN (Recurrent Neural Networks) în care conexiunile dintre neuroni sunt circulare, reţelele neurale stohastice ce introduc variaţii aleatoare în reţea sau reţelele neuronale modulare.

Global, ca metode de clasificare, reţelele neuronale folosesc doar date de intrare numerice, ceea ce restricţionează aplicabilitatea acestora doar la clasificarea obiectelor ce pot fi reprezentate numeric. Mai mult, varietatea importantă de reţele existente, precum şi numărul important de parametri ce trebuie reglaţi (de exemplu: numărul de straturi, numărul de neuroni, funcţiile de activare, etc.), fac dificilă alegerea metodei care să fie cea mai adaptată problemei de clasificare vizată. Totuşi, un avantaj important al acestora îl constituie arhitectura pur paralelă ce permite procesarea simultană a unui volum important de date (de exemplu, pentru clasificarea conţinutului unei imagini, se poate asocia un neuron fiecărui pixel al acesteia, calculele realizându-se astfel simultan).

Algoritmi statistici

Spre deosebire de reţelele neuronale, abordările statistice se evidenţiază prin modelarea repartiţiei în clase pe baza unui model statistic. Astfel, datelor de intrare li se asociază o anumită probabilitate de apartenenţă la fiecare dintre clase, şi nu doar unei singure clase ca în cazul metodelor prezentate anterior.

Dintre metodele statistice cele mai cunoscute putem menţiona *reţelele Bayes*. În varianta cea mai simplă sunt clasificatorii NB sau "Naive Bayes Classifiers". Un clasificator NB este un clasificator probabilistic simplu ce se

bazează în mod "naiv" pe ipoteza de independenţă a datelor. Cu alte cuvinte, un clasificator NB consideră că prezenţa (sau absenţa) unui anumit atribut al unei clase este total independentă de prezenţa (sau absenţa) oricărui alt atribut.

Modelul probablistic folosit de clasificatorii Bayes este un model condiţional. Acesta poate fi exprimat pe baza probabilităţii "a posteriori", $P(C|F_1, ..., F_n)$, unde C reprezintă variabila clasă iar F_i, cu $i = 1, ..., n$, reprezintă atributele claselor.

Pe baza teoremei Bayes, această probabilitate poate fi rescrisă în felul următor:

$$P(C|F_1, ..., F_n) = \frac{P(C)P(F_1, ..., F_n|C)}{P(F_1, ..., F_n)} \tag{7.16}$$

unde $P(C)$ reprezintă probabilitatea "a priori", $P(F_1, ..., F_n|C)$ reprezintă verosimilitatea ("likelihood") iar $P(F_1, ..., F_n)$ reprezintă evidenţa.

Mai departe, expresia $P(C)P(F_1, ..., F_n|C)$ poate fi rescrisă în felul următor:

$$P(C)P(F_1, ..., F_n|C) = P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2) \cdot \\ ... \cdot P(F_n|C, F_1, ..., F_{n-1}) \tag{7.17}$$

Folosind ipoteza de independenţă a atributelor F_i obţinem:

$$P(F_i|C, F_j) = P(F_i|C) \tag{7.18}$$

şi mai departe:

$$P(C|F_1, ..., F_n) = \frac{1}{Z}P(C)\prod_{i=1}^{n} P(F_i|C) \tag{7.19}$$

unde $Z = P(F_1, ..., F_n)$ reprezintă un factor de scală ce depinde doar de atributele F_i, fiind constant dacă valorile acestora sunt cunoscute.

Pentru a transforma modelul probabilistic prezentat în ecuaţia 7.19 într-un clasificator, se vor adăuga anumite reguli de decizie. Una dintre regulile cel mai frecvent folosite este dată de alegerea ipotezei cea mai probabilă, abordare cunoscută şi sub numele de "maximum a posteriori" sau MAP.

Clasificatorul obţinut în acest fel este descris de funcţia următoare:

$$Clasif(f_1, ..., f_n) = argmax_c P(C = c)\prod_{i=1}^{n} P(F_i = f_i|C = c) \tag{7.20}$$

unde $Clasif()$ reprezintă funcţia de clasificare a datelor, f_i, cu $i = 1, ..., n$, reprezintă valorile atributelor F_i iar c valorile variabilei clasă C. Astfel, datele

de intrare reprezentate prin valorile f_i ale atributelor, vor fi clasate în clasa c, pentru care probabilitatea dată de expresia 7.19 este maximă.

O abordare mai complexă o constituie *reţelele Bayes* sau BN ("Bayesian Networks"). O reţea Bayes este un model grafic de reprezentare a relaţiilor probabilistice ce pot exista între valorile unui set de atribute. O astfel de reţea este structurată sub forma unui graf orientat aciclic sau DAG ("Directed Acyclic Graph"[17]) în care nodurile corespund atributelor X, pe baza cărora se realizează clasificarea. Un exemplu este ilustrat în Figura 7.10.

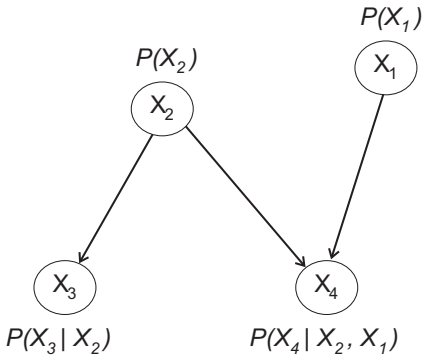


Figura 7.10: Exemplu de reţea Bayes cu patru noduri (X reprezintă atributele iar $P(X_i|X_j)$ probabilităţile condiţionale).

Muchiile grafului corespund relaţiilor cauzale ce pot exista între atribute, în timp ce lipsa muchiilor indică independenţa statistică dintre atribute. Dacă există o conexiune de la nodul A la nodul B, atunci nodul A este considerat părintele nodului B iar acesta din urmă copilul nodului A. Astfel, un graf orientat aciclic este o reţea Bayes dacă probabilitatea nodurilor poate fi exprimată ca un produs de probabilităţi locale ale fiecărui nod şi a părinţilor acestuia, astfel:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|Părinţi(X_i)) \tag{7.21}$$

unde funcţia $Părinţi(X_i)$ returnează părinţii nodului X_i [Heckerman 96].

Pentru a răspunde unei problemele de clasificare supervizată, o reţea Bayes trebuie mai întâi antrenată. În cazul cel mai simplu, un expert în domeniu va specifica parametrii reţelei astfel încât aceasta să fie adaptată

[17] un graf orientat aciclic este un graf orientat (sensul de parcurgere este luat în calcul) în care nu există bucle. Cu alte cuvinte, pentru orice nod v nu există o cale nevidă prin care se porneşte din v şi se ajunge tot în v.

cerinţelor de clasificare dorite. În cazul în care acest lucru nu este posibil sau dacă reţeau este mult prea complexă pentru a fi concepută manual, parametrii acesteia precum şi distribuţiile locale trebuiesc extrase automat din datele de antrenare. Problema care se pune este de a găsi reţeaua Bayes care modelează cel mai bine datele de antrenare, şi anume corelaţiile de referinţă dintre atributele de intrare şi respectiv de ieşire. Pentru o descriere detaliată a metodelor de învăţare folosite pentru reţelele Bayes, cititorul se poate raporta la lucrarea [Niculescu 06].

Clasificarea unui set de date de intrare reprezentate prin mulţimea de atribute $F_1, ..., F_n$ se traduce într-o reţea Bayes prin clasarea acestora în clasa de etichetă c care maximizează probabilitatea "a posteriori", şi anume $P(c|F_1, ..., F_n)$ (vezi ecuaţia 7.16).

Principalul avantaj al reţelelor Bayes, relativ la celelalte metode de clasificare (de exemplu: reţelele neuronale sau arborii de decizie), constă în posibilitatea acestora de a se modela sau adapta problemei de clasificare prin luarea în calcul a tuturor informaţiilor disponibile "a priori" despre aceasta. Astfel, conexiunile dintre noduri pot fi modificate pentru a lua în calcul eventualele ipoteze adiţionale. Pe de altă parte, în cazul în care setul de atribute este complex şi conţine un număr considerabil de atribute, folosirea unui clasificator pe bază de reţele Bayes nu este eficientă datorită volumului reţelei, cât şi a timpului de calcul. De asemenea, trebuie luat în calcul faptul că într-o reţea Bayes de cele mai multe ori valorile atributelor de intrare trebuie eşantionate [Kotsiantis 07].

Algoritmi bazaţi pe instanţe

Această categorie de metode de clasificare supervizată intră tot sub incidenţa metodelor de clasificare statistică. Totuşi, metodele bazate pe instanţe sunt algoritmi de învăţare "leneşi" ("lazy-learning algorithms") întrucât aceştia întârzie procesul de inducţie şi de generalizare până în momentul în care se realizează clasificarea. Din această cauză, etapa de învăţare a unui algoritm "leneş" are de regulă o complexitate de calcul mai redusă decât în cazul algoritmilor "intensivi" ("eager-learning algorithms", precum arborii de decizie sau reţelele Bayes). Pe de altă parte, procesul de clasificare se dovedeşte a fi mai complex.

Dintre metodele cele mai populare, putem menţiona metoda de clasificare de tip "cel mai apropiat vecin" sau k-NN ("k-Nearest Neighbor") [Fix 51]. Aceasta se bazează pe principiul că instanţele dintr-un anumit set de date (valorile atributelor) se vor găsi în general în proximitatea altor instanţe cu proprietăţi similare. Astfel, dacă o instanţă este clasată ca aparţinând unei anumite clase, atunci o altă instanţă, ce nu a fost încă clasificată, poate fi

atribuită unei clase doar prin observarea clasei din care face parte cel mai apropiat vecin al acesteia[18]. Mai general, metoda k-NN, analizează cele mai apropiate k instanţe, iar noua instanţă va fi în principiu alocată clasei predominante. Clasele vor fi diferenţiate între ele pe baza analizei similarităţii dintre datele conţinute de acestea.

Etapa de antrenare a metodei k-NN constă doar în stocarea datelor de antrenare, şi anume a perechilor: valori atribute - etichetă clasă. Clasificarea propriu-zisă a unei noi instanţe se realizează prin calcularea în primă fază a unei măsuri de distanţă între aceasta şi instanţele deja stocate. Dintre măsurile de distanţă folosite, putem enumera următoarele (unde x şi y reprezintă doi vectori m-dimensionali de caracteristici):

- distanţa Minkowski:

$$D(x,y) = \left(\sum_{i=1}^{m} |x_i - y_i|^r \right)^{1/r} \tag{7.22}$$

  unde r este o valoare întreagă.

- distanţa Manhattan:

$$D(x,y) = \sum_{i=1}^{m} |x_i - y_i| \tag{7.23}$$

- distanţa Chebyshev:

$$D(x,y) = \max_{i=1}^{m} |x_i - y_i| \tag{7.24}$$

- distanţa Euclidiană:

$$D(x,y) = \left(\sum_{i=1}^{m} |x_i - y_i|^2 \right)^{1/2} \tag{7.25}$$

- distanţa Canberra:

$$D(x,y) = \sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i + y_i|} \tag{7.26}$$

[18] noţiunea de apropiere între instanţe se referă la proximitatea acestora în spaţiul de caracteristici considerat.

- corelaţia Kendall:

$$D(x,y) = 1 - \frac{2}{m(m-1)} \sum_{i=j}^{m} \sum_{j=1}^{i-1} sign(x_i - x_j) sign(y_i - y_j) \quad (7.27)$$

unde funcţia $sign(E)$ returnează semnul expresiei E.

Calculând D pentru toate perechile de instanţe, se va obţine o matrice de distanţe ce este folosită mai departe pentru a analiza gradul de vecinătate dintre acestea. Fiecare instanţă i, va fi însoţită de eticheta clasei din care face parte, $c_i \in \{c_1, ..., c_n\}$, unde n reprezintă numărul total al instanţelor.

Mai departe, pentru instanţa curentă analizată, folosind matricea de distanţe, sunt localizate cele mai apropiate k instanţe. Eticheta clasei predominante va fi atribuită instanţei analizate. În cazul în care una sau mai multe etichete de clase sunt prezente în număr egal, atunci testul este relansat de această dată pentru $k-1$ instanţe. Dacă egalitatea se păstrează, numărul de obiecte este diminuat recursiv până când se ajunge la $k = 1$, moment în care eticheta clasei rămase va fi atribuită implicit instanţei analizate.

În ciuda simplităţii sale, metoda k-NN se dovedeşte a fi o metodă de clasificare eficientă pentru multe domenii de aplicaţie. Totuşi, aceasta prezintă şi o serie de dezavantaje, printre care cele mai importante sunt: dependenţa acesteia de noţiunea de proximitate (concept vag, dependent de măsura de distanţă folosită) precum şi dificultatea de alegere adecvată a valorii parametrului k (metodele existente propuse având o complexitate de calcul ridicată).

Support Vector Machines

"Support Vector Machines" sau SVM realizează clasificarea datelor prin construcţia hiperplanului[19] ce separă în mod optimal datele de intrare în două categorii [Welling 05].

Ca şi în cazul celorlalte metode de clasificare, datele sunt reprezentate ca fiind vectori $n-$dimensionali în spaţiul de caracteristici, iar SVM încearcă să determine dacă aceste puncte pot fi separate de un hiperplan $(n-1)$ dimensional. Aceasta este o problemă de clasificare liniară. Având în vedere că există o multitudine de hiperplane ce pot separa datele, SVM restricţionează căutarea la acele hiperplane ce permit o separare maximă între cele două clase (maximizarea "marginii" dintre date). Cu alte cuvinte, se caută hiperplanul cu proprietatea ca acesta să maximizeaze distanţa faţă de cel mai apropiat

[19] un hiperplan este un concept folosit în domeniul algebrei liniare pentru a generaliza noţiunea de linie, folosită în geometria Euclidiană a planului, sau de plan, folosită în geometria Euclidiană tridimensională, pentru cazul n-dimensional, cu $n > 3$.

punct din spaţiul de caracteristici. Acesta este cunoscut şi sub numele de "hiperplanul marginii maximale" ("maximum-margin hyperplane").

Formalizarea problemei de clasificare abordată de SVM este următoarea: având la dispoziţie un set de date de antrenare, D, constituit ca fiind:

$$D = \{(X_i, c_i) | X_i \in R^n, c_i \in \{-1, 1\}\} \tag{7.28}$$

unde X_i este un vector $n-$dimensional (vector de caracteristici), c_i indică clasa din care face parte vectorul X_i (valori etichete -1 şi 1), i reprezintă indicele vectorului curent, cu $i = 1, ..., p$, iar p reprezintă numărul de vectori consideraţi; se caută hiperplanul marginii maximale ce permite separarea punctelor din clasa $c_i = 1$ de cele din clasa $c_i = -1$ (vezi Figura 7.11).

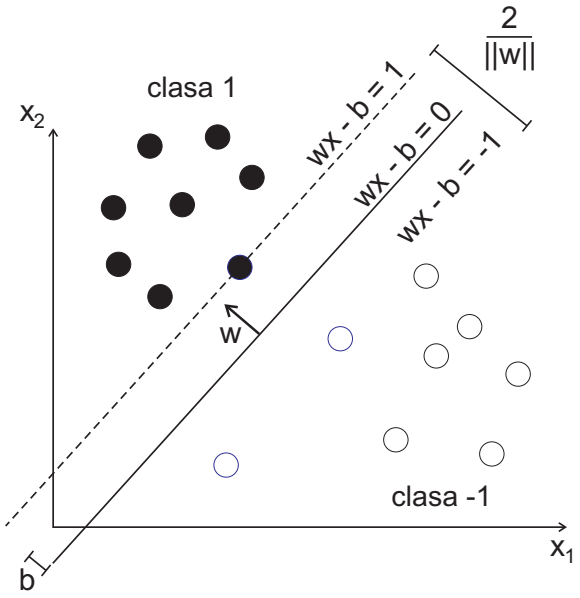


Figura 7.11: Hiperplanul marginii maximale în cazul a două clase (cercurile reprezintă vectorii de caracteristici, X_1 şi X_2 formează spaţiul de caracteristici, sursă Wikipedia "http://en.wikipedia.org/wiki/Support_vector_machine").

Un hiperplan oarecare poate fi definit ca un set de puncte X ce satisfac următoarea relaţie:

$$W \cdot X - b = 0 \tag{7.29}$$

unde W reprezintă un vector normal (perpendicular pe hiperplan) iar parametrul $\frac{b}{||W||}$ va defini decalajul hiperplanului faţă de originea axei de coordonate, de-a lungul vectorului W (vezi Figura 7.11).

În scopul definirii marginii maximale, căutăm valorile lui W şi b astfel încât acestea să maximizeze distanţa dintre hiperplanele paralele, cele mai depărtate, dar care încă separă datele. Acestea sunt date de ecuaţiile:

$$W \cdot X - b = 1 \tag{7.30}$$
$$W \cdot X - b = -1 \tag{7.31}$$

Distanţa dintre acestea este $\frac{2}{||W||}$, astfel că problema maximizării se transformă într-o problemă de minimizare a valorii $||W||$.

De asemenea, pentru a preveni ca punctele să se găsească pe margini, se folosesc o serie de constrângeri suplimentare, astfel marginea maximală este determinată de condiţiile următoare:

$$W \cdot X_i - b \geq 1, \quad X_i \in c_1 \tag{7.32}$$
$$W \cdot X_i - b \leq -1, \quad X_i \in c_{-1} \tag{7.33}$$

sau

$$c_i \cdot (W \cdot X_i - b) \geq 1 \tag{7.34}$$

pentru oricare $i \in [1; p]$.

Transformată într-o problemă de optimizare, clasificarea SVM poate fi enunţată astfel: alege parametrii W şi b astfel încât să minimizeze valoarea $||W||$ cu constrângerea că $c_i \cdot (W \cdot X_i - b) \geq 1$, pentru oricare i. Această clasificare este valabilă totuşi doar în cazul în care datele sunt liniar separabile.

Pentru a crea un clasificator SVM neliniar, la maximizarea marginii dintre clase se folosesc ceea ce numim funcţii nucleu sau "kernel functions". Algoritmul rezultat este unul similar din punct de vedere al principiului, doar că operaţiile de înmulţire sunt înlocuite acum de nuclee de funcţii neliniare. În acest fel, hiperplanul marginii maximale va fi potrivit datelor într-un spaţiu de caracteristici transformat neliniar. Dintre nucleele, $k()$, cel mai frecvent folosite putem menţiona următoarele:

- nucleu polinomial omogen:

$$k(X, X') = (X \cdot X')^d \tag{7.35}$$

  unde d este un număr întreg,

- nucleu polinomial neomogen:

$$k(X, X') = (X \cdot X' + 1)^d \tag{7.36}$$

- funcţie radială:

$$k(X, X') = \exp(-\gamma||X - X'||^2) \tag{7.37}$$

unde $\gamma > 0$,

- funcţie radială Gaussiană:

$$k(X, X') = \exp\left(-\frac{||X - X'||^2}{2\sigma^2}\right) \tag{7.38}$$

unde σ^2 reprezintă varianţa statistică,

- funcţie sigmoidă:

$$k(X, X') = \tanh(\kappa \cdot X \cdot X' + c) \tag{7.39}$$

unde *tanh* reprezintă tangenta hiperbolică, $\kappa > 0$ iar $c < 0$.

Din punct de vedere al performanţelor de clasificare, SVM sunt capabile să clasifice eficient date multidimensionale precum şi vectori de caracteristici continui. Totuşi, pentru a obţine o precizie a predicţiei maximală, este de regulă necesar un set de date de antrenare suficient de vast. În ceea ce priveşte limitările SVM, una dintre cele mai importante este faptul că SVM este un clasificator binar, datele fiind clasificate în doar două clase. Pentru a realiza o clasificare în mai mult de două clase, aceasta trebuie descompusă într-un set de clasificări binare [Kotsiantis 07].

7.3 Concluzii

În acest capitol am prezentat problematica clasificării automate după conţinut a datelor. Metodele de clasificare existente au ca obiectiv general regruparea datelor unei colecţii mari de date în categorii cât mai *omogene*. Gradul de omogenitate este controlat de utilizator prin introducerea noţiunii de similaritate între date. Aceasta este definită de regulă într-un spaţiu n-dimensional, dat de o serie de atribute numerice reprezentative a datelor, spaţiu numit şi *spaţiul de caracteristici*.

În ciuda diversităţii ridicate de metode existente, acestea se împart în două mari categorii, şi anume:

- vorbim de *clasificare nesupervizată* în cazul în care partiţia optimală a spaţiului de caracteristici, din punct de vedere al unui anumit criteriu matematic, este realizată direct, fără a folosi o analiză prealabilă a

conţinutului datelor sau alte informaţii referitoare la acestea. Avantajul acestei abordări este dat de automatizarea procesului de clasificare, lucru ce tinde să diminueze relevanţa claselor obţinute,

- pe de altă parte, vorbim de *clasificare supervizată* în cazul în care clasarea datelor se face pornind de la o serie de modele predefinite de clase sau clasificări de referinţă. Clasificarea supervizată implică astfel o etapă prealabilă clasificării în care sistemul învaţă din exemple. Avantajul acestui tip de abordare este dat în principal de relevanţa în general ridicată a conţinutului claselor obţinute. Totuşi, o limitare a acestor metode este dată de faptul că datele de antrenare nu sunt întotdeauna disponibile.

Din punct de vedere al problematicii indexării datelor, tehnicile de clasificare sunt indispensabile unui sistem de indexare după conţinut. Clasificarea datelor intervine în general în însuşi procesul de căutare al informaţiei. Utilizatorul, prin formularea cererii de căutare va defini spaţiul de caracteristici ce va fi folosit pentru localizarea datelor dorite. Pe baza acestuia, datele din baza de date pot fi grupate în funcţie de similaritate sau cu alte cuvinte în funcţie de asemănarea dintre vectorii de caracteristici asociaţi. Astfel, grupul sau grupurile de date ce sunt suficient de similare vectorului de caracteristici asociat cererii de căutare vor fi furnizate utilizatorului drept rezultat.

La indexarea bazelor de date generice se caută folosirea unei metode de clasificare nesupervizată, deoarece în acest caz, nu dispunem de informaţii referitoare la conţinutul acesteia sau la o posibilă repartiţie în clase. Astfel, se caută un clasificator care să poată pune în evidenţă noi relaţii dintre date, sau ceea ce numim "cunoaştere". Clasificarea supervizată este folosită atunci când domeniul de aplicaţie este cunoscut "a priori" şi o expertiză a acestuia este disponibilă. Astfel, algoritmul de clasificare poate fi antrenat pentru a răspunde la anumite cerinţe de selecţie.

Diversitatea metodelor de clasificare disponibile face dificilă alegerea metodei adecvate aplicaţiei dorite. Şi pentru ca lucrurile să fie şi mai complicate, găsirea metodei optimale nu garantează clasificarea corectă. Aceasta mai depinde, în principal, şi de alegerea eficientă a atributelor (cât mai reprezentative), de măsura de similaritate folosită cât şi de validarea corectă a rezultatelor (de regulă dificil de reprezentat).

Bibliografie

[4i2i 06] 4i2i. *H.263 Video Coding Tutorial*. `http://www.4i2i.com/h263_video_codec.htm`, 2006.

[Accame 98] M. Accame, F.G.B. De Natale & D.D.Giusto. *High Performance Hierarchical Block-based Motion Estimation for Real-Time Video Coding*. Real-Time Imaging, vol. 4, pag. 67–79, 1998.

[Acosta 02] E. Acosta, L. Torres, A. Albiol & E. Delp. *An Automatic Face Detection and Recognition System for Video Indexing Applications*. IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pag. 3644–3647, 2002.

[Adames 02] B. Adames, C. Dorai & S. Venkatesh. *Towards Automatic Extraction of Expressive Elements of Motion Pictures: Tempo*. IEEE Transactions on Multimedia, vol. 4, nr. 4, pag. 472–481, decembrie 2002.

[Adjeroh 01] D.A. Adjeroh & M.C. Lee. *On Ratio-Based Color Indexing*. IEEE Transactions on Image Processing, vol. 10, nr. 1, pag. 36–48, ianuarie 2001.

[Agoston 87] G.A. Agoston. *Color Theory and Its Application in Art and Design*. Optical Sciences, Springer-Verlag, 1987.

[Aigrain 95] P. Aigrain, P. Jolly & V. Longueville. *Medium Knowledge-Based Macro-Segmentation into Sequences*.

Working notes of IJCAI Workshop on Intelligent Multimedia Information Retrieval, pag. 5–14, Montreal, Canada 1995.

[Akutsu 92] A. Akutsu, Y. Tonomura, H. Hashimoto & Y. Ohba. *Video Indexing Using Motion Vectors*. SPIE Visual Communications Image Processing, vol. 1818, 1992.

[Alatan 01] A.A. Alatan, A.N. Akasu & W. Wolf. *Multimodal Dialog Scene Detection Using Hidden Markov Models for Content-Based Multimedia Indexing*. Multimedia Tools and Applications, vol. 14, nr. 2, pag. 137–151, 2001.

[Alattar 93] A.M. Alattar. *Detecting and Compressing Dissolve Regions in Video Sequences with a DVI Multimedia Image Compression Algorithm*. IEEE International Symposium on Circuits and Systems, vol. 1, pag. 13–16, mai 1993.

[Alattar 97] A.M. Alattar. *Detecting Fade Regions in Uncompressed Video Sequences*. IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pag. 3025–3028, 1997.

[Allen 83] J.F. Allen. *Maintaining Knowledge About Temporal Intervals*. Communications of the ACM, vol. 26, nr. 11, noiembrie 1983.

[Aner 01] A. Aner & J.R. Kender. *Mosaic-Based Clustering of Scene Locations in Videos*. IEEE Workshop on Content-based Access of Image and Video Libraries, decembrie, Hawaii, USA 2001.

[ARGOS 06] ARGOS. *Campagne d'Evaluation d'Outils de Surveillance de Contenus Vidéo*. `http://www.irit.fr/argos`, 2006.

[Ariki 03] Y. Ariki, M. Kumano & K. Tsukada. *Highlight Scene Extraction in Real Time From Baseball Live Video*. ACM 5th International Workshop on Multimedia Information Retrieval, pag. 209–214, 2003.

[Arman 93a] F. Arman, A. Hsu & M.Y. Chiu. *Feature Management for Large Video Databases*. SPIE Storage and Retrieval

for Image and Video Databases, vol. 1908, pag. 2–12, februarie 1993.

[Arman 93b] F. Arman, A. Hsu & M.Y. Chiu. *Image Processing on Compressed Data for Large Video Database*. ACM International Conference on Multimedia, pag. 267–272, august, Anaheim, USA 1993.

[A.R.Weeks 95] A.R.Weeks, G.E.Hague & H.R. Myler. *Histogram Equalization of 24-bits Color Images in the Color Difference (c-y) Color Space*. Journal of Electronic Imaging, vol. 4, nr. 1, pag. 15–22, ianuarie 1995.

[Babaguchi 02] N. Babaguchi, Y. Kawai & T. Kitahashi. *Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration*. IEEE Transactions on Multimedia, vol. 4, nr. 1, 2002.

[Barnard 03] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D.M. Blei & M.I. Jordan. *Matching Words and Pictures*. Journal of Machine Learning Research, vol. 3, pag. 1107–1135, 2003.

[Barron 94] J.L. Barron, D.J. Fleet & S.S. Beauchemin. *Performance of Optical Flow Techniques*. International Journal of Computer Vision, vol. 12, nr. 1, pag. 43–77, februarie 1994.

[Beaver 94] F. Beaver. *Dictionary of Film Terms*. New York: Twayne, 1994.

[Ben-Yacoub 99] S. Ben-Yacoub, B. Fasel & J. Luettin. *Fast Face Detection Using MLP and FFT*. Second International Conference on Audio and Video-Based Biometric Person Authentication, pag. 31–36, 1999.

[Benavente 04] R. Benavente & M. Vanrell. *Fuzzy Colour Naming Based on Sigmoid Membership Functions*. The Second European Conference on Colour Graphics, Imaging and Vision, pag. 135–139, aprilie 2004.

[Benitez 01] A. Benitez, S.-F. Chang & J.R. Smith. *IMKA: A Multimedia Organization System Combining Perceptual and*

Semantic Knowledge. ACM Multimedia, vol. 18, pag. 121–129, 2001.

[Benoit 07] H. Benoit & M. Bernard. *Interactive Video.* Signals and Communication Technology, Springer Berlin Heidelberg, vol. 1, pag. 27–42, 2007.

[Berlin 91] B. Berlin & P. Kay. *Basic Color Terms: Their Universality and Evolution.* University of California Press, Berkeley 1991.

[Bimbo 99] A. Del Bimbo. *Visual Information Retrieval.* Morgan Kaufmann Publishers, San Francisco, USA 1999.

[Birren 69] F. Birren. *Principles of Color - A Review of Past Traditions and Modern Theories of Color Harmony.* New York: Reinhold, 1969.

[Boccignone 00] G. Boccignone, M. De Santo & G. Percannella. *Automated Threshold Selection for the Detection of Dissolves in Mpeg Video.* IEEE International Conference on Multimedia and Expo, vol. 3, pag. 1535–1538, 2000.

[Boreczky 98] J.S. Boreczky & L.D. Wilcox. *A Hidden Markov Model Framework for Video Segmentation Using Audion and Image Features.* IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle, USA 1998.

[Bouthemy 98] P. Bouthemy & R. Fablet. *Motion Characterization from Temporal Cooccurrences of Local Motion-Based Measures for Video Indexing.* Pattern Recognition, vol. 1, pag. 905–908, august 1998.

[Bretl 99] W. Bretl & M. Fimoff. *MPEG Tutorial.* `http://www.zenith.com/sub_hdtv/mpeg_tutorial/`, 1999.

[Calic 02a] J. Calic & E. Izquierdo. *Efficient Key-Frame Extraction and Video Analysis.* International Conference on Information Technology: Coding and Computing, pag. 28–33, 2002.

[Calic 02b] J. Calic & E. Izquierdo. *A Multiresolution Technique for Video Indexing and Retrieval.* IEEE International

Conference on Image Processing, vol. 1, pag. 952–955, 2002.

[Calic 04] J. Calic & E. B.T. Thomas. *Spatial Analysis in Key-Frame Extraction Using Video Segmentation*. Workshop on Image Analysis for Multimedia Interactive Services, Lisboa, Portugal 2004.

[Campisi 99] P. Campisi, A. Longari & A. Neri. *Automatic Key Frame Selection Using a Wavelet Based Approche*. SPIE, vol. 3813, pag. 861–872, 1999.

[Chang 98] S.-F. Chang, W. Chen, H. Meng, H. Sundara & D. Zhong. *A Fully Automatic Content-Based Video Search Engine Supporting Multi-Object Spatio-Temporal Queries*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 8, pag. 602–615, 1998.

[Chang 99] S.-F. Chang, W. Chen, H.J. Meng, H. Sundaram & D. Zhong. *Evaluation of Texture Segmentation Algorithms*. International Conference on Computer Vision and Pattern Recognition, vol. 1, pag. 294–299, 1999.

[Chanussot 98] J. Chanussot. *Approches Vectorielles ou Marginales pour le Traitement d'Images Multi-composantes*. Teză de doctorat, Université de Savoie, Annecy-France, 1998.

[Chen 99] Y. Chen, E.K. Wong, M.M. Yeunh, Y. Boon-Lock & A.C. Charles. *Augmented Image Histogram for Image and Video Similarity Search*. SPIE Conference on Storage and Retrieval for Image and Video Database VII, vol. 3656, pag. 523–532, 1999.

[Chen 05] S.-C. Chen, M.-L. Shyu & N. Zhao. *An Enhanced Query Model for Soccer Video Retrieval Using Temporal Relationships*. IEEE International Conference on Data Engineering, aprilie 2005.

[Chen 06] J.-Y. Chen, C. Taskiran, E.J. Delp & C.A. Bouman. *ViBE: A New Paradigme for Video Database Browsing and Search*. `http://stargate.ecn.purdue.edu/~ips/ViBE/`, 2006.

[Choi 98] H. Choi & R. Baraniuk. *Multiscale Texture Segmentation using Wavelet-Domain Hidden Markov Models*. 32st Asilomar Conference on Signals, Systems and Computers, vol. 2, pag. 1692–1697, 1998.

[CICA 06] CICA. *Centre International du Cinéma d'Animation*. `http://www.annecy.org`, 2006.

[Coldefy 04] F. Coldefy & P. Bouthemy. *Unsupervised Soccer Video Abstraction Based on Pitch, Dominant Color and Camera Motion Analysis*. ACM Multimedia, pag. 268–271, Ney York, USA 2004.

[Colombo 99] C. Colombo, A. Del Bimbo & P. Pala. *Semantics in Visual Information Retrieval*. IEEE Multimedia, vol. 6, nr. 3, pag. 38–53, 1999.

[Cooper 02] M. Cooper & J. Foote. *Summarizing Video Using Non-Negative Similarity Matrix Factorization*. IEEE Workshop on Multimedia Signal Processing, pag. 25–28, St. Thomas, US Virgin Islands 2002.

[Corridoni 95] J.M. Corridoni & A. Del Bimbo. *Film Semantic Analysis*. Proceedings of Computer Architectures for Machine Perception, pag. 202–209, septembrie, Como-Italy 1995.

[Corridoni 99] J.M. Corridoni, A. Del Bimbo & P. Pala. *Retrieval in Paintings Using Effects Induced by Color Features*. IEEE Multimedia, vol. 6, nr. 3, pag. 38–53, 1999.

[Cox 00] I.J. Cox, M. Miller, T.P. Minka, T.V. Papathomas & P.N. Yianilos. *The Bayesian Image Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments*. IEEE Transactions on Image Processing, vol. 9, pag. 20–37, 2000.

[CSAIL 06] CSAIL. *Color Name Dictionaries*. `http://swiss.csail.mit.edu/~jaffer/Color/Dictionaries.html`, 2006.

[Dagtas 00] S. Dagtas, W. Al-Khatib, A. Ghafoor & R.L. Kashyap. *Models for Motion-Based Video Indexing and Retrieval*. IEEE Transactions on Image Processing, vol. 9, nr. 1, pag. 88–101, ianuarie 2000.

[Dempster 77] A.P. Dempster, N.M. Laird & D.B. Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, vol. 39, nr. 1, pag. 1–38, 1977.

[Detyniecki 03] M. Detyniecki & C. Marsala. *Discovering Knowledge for Better Video Indexing Based on Colors*. IEEE International Conference on Fuzzy Systems, vol. 2, pag. 1177–1181, Paris, France 2003.

[Délibéré 89] M. Délibéré. *La Couleur*. Presses Universitaires de France, 6éme édition, collection: Que sais-je?, vol. 220, 1989.

[Donderler 04] M.E. Donderler, O. Ulusoy & U. Gudukbay. *Rule-Based Spatio-Temporal Query Processing for Video Databases*. International Journal on Very Large Data Bases, vol. 13, nr. 1, ianuarie 2004.

[Dorado 04] A. Dorado, J. Calic & E. Izquierdo. *A Rule-Based Video Annotation System*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, nr. 5, pag. 622–633, mai 2004.

[Doulamis 98] N.D. Doulamis, A.D. Doulamis, Y.S. Avrithis & S.D. Kollias. *Video Content Representation Using Optimal Extraction of Frames and Scenes*. IEEE International Conference on Image Processing, vol. 1, pag. 875–879, Chicago, USA 1998.

[Doulamis 00a] A.D. Doulamis, N. Doulamis & S. Kollias. *Non-Sequential Video Content Representation Using Temporal Variation of Feature Vectors*. IEEE Transactions on Consumer Electronics, vol. 46, nr. 3, pag. 758–768, 2000.

[Doulamis 00b] A.D. Doulamis, N.D. Doulamis & S.D. Kollias. *A Fuzzy Video Content Representation for Video Summarization and Content-Based Retrieval*. Signal Processing, vol. 80, nr. 6, pag. 1049–1067, iunie 2000.

[Doulamis 00c] N.D. Doulamis, A.D. Doulamis, Y.S. Avrithis, K.S. Ntalianis & S.D. Kollias. *Efficient Summarization of Stereoscopic Video Sequences*. IEEE Transactions on

Circuits and Systems for Video Technology, vol. 10, nr. 4, pag. 501–517, iunie 2000.

[Drew 00] M.S. Drew, Z.N. Li & X. Zhong. *Video Dissolve and Wipe Detection Via Spatio-Temporal Images of Chromatic Histograms Differences*. IEEE International Conference on Image Processing, vol. 3, pag. 929–932, 2000.

[Duan 06] L.-Y. Duan, J.S. Jin & C.-S. Xu Q. Tian. *Nonparametric Motion Characterization for Robust Classification of Camera Motion Patterns*. IEEE Transactions on Multimedia, vol. 8, nr. 2, pag. 323–340, aprilie 2006.

[Dufaux 00] F. Dufaux. *Key Frame Selection to Represent a Video*. IEEE International Conference on Multimedia and Expo, vol. 2, pag. 275–278, 2000.

[Dunn 73] J.C. Dunn. *A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters*. Journal of Cybernetics, vol. 3, pag. 32–57, 1973.

[Eick 04] C.F. Eick, N. Zeidat & Z. Zhao. *Supervised Clustering - Algorithms and Benefits*. 16th IEEE International Conference on Tools with Artificial Intelligence, pag. 774 – 776, 2004.

[Eidenberger 04] H. Eidenberger. *A Video Browsing Application Based on Visual MPEG-7 Descriptors and Self-Organising Maps*. TFSA International Journal of Fuzzy Systems, vol. 6, nr. 3, pag. 122–135, septembrie 2004.

[Elad 99] M. Elad, P. Teo & Y. Hel-Or. *Optimal Filters for Gradient-Based Motion Estimation*. IEEE International Conference on Computer Vision, pag. 559–565, Corfu, Greece 1999.

[Electric 05] Mitsubishi Electric. *MERL - Timetunnel Interface for Video Browsing*. `http://www.merl.com/projects/timetunnel2/`, 2005.

[Erol 00] B. Erol & F. Kossentini. *Automatic Key Video Object Plane Selection Using the Shape Information in the MPEG-4 Compressed Domain*. IEEE Transactions on Multimedia, vol. 2, pag. 129–138, 2000.

[Erol 03] B. Erol & D.-S.H.J. Lee. *Multimodal Summarization of Meeting Recordings*. IEEE International Conference on Multimedia and Expo, vol. 3, pag. 25–28, 2003.

[Estrela 04] V. Estrela, L.A. Rivera & M.H.S. Bassani. *Pel-Recursive Motion Estimation Using the Expectation-Maximization Technique and Spatial Adaptation*. International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, pag. 47–54, februarie, Plzen, Czech Republic. 2004.

[Fablet 02] R. Fablet, P. Bouthemy & P. Pérez. *Nonparametric Motion Characterization Using Causal Probabilistic Models for Video Indexing and Retrieval*. IEEE Transactions on Image Processing, vol. 11, nr. 4, pag. 393–407, 2002.

[Fan 01] J. Fan, W.G. Aref, A.K. Elmagamid, M.-S. Hacid, M.S. Marzouk & X. Zhu. *Multi-Level Video Content Representation and Retrieval*. Journal of Electronic Imaging, Special Issue on Multimedia Database, vol. 10, nr. 4, pag. 895–908, 2001.

[Fan 04] J. Fan, H. Luo & A.K. Elmagarmid. *Concept-Oriented Indexing of Video Databases: Toward Semantic Sensitive Retrieval and Browsing*. IEEE Transactions on Image Processing, vol. 13, nr. 7, pag. 974–991, 2004.

[Farnebäck 00] G. Farnebäck. *Fast and Accurate Motion Estimation using Orientation Tensors and Parametric Motion Models*. 15th International Conference on Pattern Recognition, vol. 1, pag. 135–139, septembrie, Barcelona, Spain 2000.

[Faugeras 79] O.D. Faugeras. *Digital Color Image Processing within the Framework of a Human Visual Model*. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, nr. 4, pag. 380–393, 1979.

[Fauvet 04] B. Fauvet, P. Bouthemy, P. Gros & F. Spindler. *A Geometrical Key-Frame Selection Method Exploiting Dominant Motion Estimation in Video*. International Conference on Image and Video Retrieval, vol. 3115, pag. 419–427, Dublin, Ireland 2004.

[Ferecatu 01] M. Ferecatu, N. Boujemaa & S. Boughorbel. *Local Color Structure Signatures for Image Retrieval*. International Workshop on Multimedia Content-Based Indexing and Retrieval, Rocquencourt, France 2001.

[Ferman 99] A.M. Ferman & A.M. Tekalp. *Probabilistic Analysis and Extraction of Video Content*. IEEE International Conference on Image Processing, vol. 2, pag. 91–95, octombrie, Kobe, Japan 1999.

[Fernando 99] W.A.C. Fernando, C.N. Canagarajah & D.R.Bull. *Fade and Dissolve Detection in Uncompressed and Compressed Video Sequence*. IEEE International Conference on Image Processing, vol. 3, pag. 299–303, octombrie, Kobe, Japan 1999.

[Fernando 01] W.A.C. Fernando, C.N. Canagarajah & D.R.Bull. *Scene Change Detection Algorithms for Content-Based Video Indexing and Retrieval*. IEE Electronics and Communication Engineering Journal, pag. 117–126, iunie 2001.

[Fix 51] E. Fix & J. Hodges. *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*. Technical Report 11, USAF School of Aviation Medicine, Randolph Field, Texas, USA 1951.

[Flickner 95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Patkovic, D. Steele & P. Yanker. *Query by Image and Video Content: The QBIC System*. IEEE Computer, vol. 28, nr. 9, pag. 23–32, septembrie 1995.

[Folimage 06] Studio Folimage. *Présentation du studio et de ses productions*. `http://www.folimage.com`, 2006.

[Fraley 96] C. Fraley. *Algorithms for Model-Based Gaussian Hierarchical Clustering*. Technical Report No.311, University of Washington, Department of Statistics, Seattle, USA 1996.

[Furht 95] B. Furht, S.W. Smoliar & H. Zhang. *Video and Image Processing in Multimedia Systems*. Norwell, Kluwer 1995.

[Furnkraz 99] J. Furnkraz. *Separate-and-Conquer Rule Learning.* Artificial Intelligence Review, vol. 13, pag. 3–54, 1999.

[Gargi 00] U. Gargi, R. Kasturi & S.H. Strayer. *Performance Characterization of Video-Shot-Change Detections Methods.* IEEE Transactions on Circuits, Systems for Video Technology, vol. 10, nr. 1, pag. 1–13, februarie 2000.

[Ge 02] J. Ge & G. Mirchandani. *A New Hybrid Block-Matching Motion Estimation Algorithm.* IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pag. 4190, 2002.

[Gibson 02] N.C.D. Gibson & B. Thomas. *Visual Abstraction of Wildlife Footage Using Gaussian Mixture Models.* International Conference on Vision Interface, pag. 814–817, 2002.

[Gilvarry 99] J. Gilvarry. *Extraction of Motion Vectors from an MPEG Stream.* Technical report Dublin City University, `http://www.cdvp.dcu.ie/Papers/MVector.pdf`, 1999.

[Gimel'farb 96] G.L. Gimel'farb & A.K. Jain. *On Retrieving Textured Images from an Image Database.* Pattern Recognition, vol. 29, nr. 9, pag. 1461–1483, 1996.

[Gong 00] Y. Gong & X. Liu. *Generating Optimal Video Summaries.* IEEE International Conference on Multimedia and Expo, vol. 3, pag. 1559–1562, New York, USA 2000.

[Gong 03] Y. Gong & X. Liu. *Video Summarization and Retrieval Using Singular Value Decomposition.* ACM Multimedia Systems Journal, vol. 9, pag. 157–168, 2003.

[Graffigne 95] C. Graffigne. *Approche Région: Méthodes Markoviennes.* Analyse d'Images: Filtrage et Segmentation, Masson, Paris, pag. 281–304, octombrie 1995.

[Gu 97] L. Gu, K. Tsui & D. Keightley. *Dissolve Detection in MPEG Compressed Video.* IEEE International Conference on Intelligent Processing Systems, vol. 2, pag. 1692–1696, 1997.

[Guillaume 01] S. Guillaume. *Induction de Règles Floues Interprétables.* Teză de doctorat, INSA Toulouse-France, noiembrie 2001.

[Guimaraes 03] S.J.F. Guimaraes, M. Couprie, A. de A. Araujo & N.J. Leite. *Video Segmentation Based on 2D Image Analysis.* Pattern Recognition Letters, nr. 24, pag. 947–957, 2003.

[H. Sundaram 00] S. Chang H. Sundaram. *Determining Computable Scenes in Films and Their Structures using Audio-Visual Memory Models.* ACM Multimedia, pag. 95 – 104, Marina del Rey, California, US 2000.

[Haering 00] N. Haering, R. Qian & I. Sezan. *A Semantic Event-Detection Approach and Its Application to Detecting Hunts in Wildlife Video.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 10, nr. 6, pag. 857–868, 2000.

[Hampapur 95] A. Hampapur, R. Jain & T.E. Weymouth. *Production Model Based Digital Video Segmentation.* Multimedia Tools and Applications, vol. 1, pag. 9–45, 1995.

[Hampson 00] F.J. Hampson & J.-C. Pesquet. *Motion Estimation in the Presence of Illumination Variations.* Signal Processing: Image Communication, vol. 16, pag. 373–381, 2000.

[Han 02a] J. Han & K.-K. Ma. *Fuzzy Color Histogram and Its Use in Color Image Retrieval.* IEEE Transactions on Image Processing, vol. 11, nr. 8, pag. 944–952, 2002.

[Han 02b] M. Han, W. Hua, W. Xu & Y. Gong. *An Integrated Baseball Digest System Using Maximum Entropy Method.* ACM Multimedia, Juan-les-Pins, France 2002.

[Hanjalic 97] A. Hanjalic, M. Ceccarelli, R.L. Lagendijk & J. Biemond. *Automation of Systems Enabling Search on Stored Video Data.* SPIE Storage and Retrieval for Image and Video Databases V, vol. 3022, pag. 427–438, februarie 1997.

[Hanjalic 02] A. Hanjalic. *Shot-Boundary Detection: Unraveled and Resolved?* IEEE Transactions on Circuits and Systems

for Video Technology, vol. 12, nr. 2, pag. 90–105, februarie 2002.

[Hauptmann 98] A.G. Hauptmann & M.J. Witbrock. *Story Segmentation and Detection of Commercials in Broadcast News Video*. Advances in Digital Libraries, pag. 168–179, aprilie, Santa Barbara, USA 1998.

[He 99] L. He, E. Sanocki, A. Gupta & J. Grudin. *Auto-Summmarization of Audio-Video Presentations*. ACM Multimedia, pag. 489–498, Orlando, USA 1999.

[Heckerman 96] D. Heckerman. *A Tutorial on Learning With Bayesian Networks*. Microsoft Research Advanced Technology Division, Technical Report MSR-TR-95-06, 1996.

[Heng 99] W.J. Heng & K.N. Ngan. *Post Shot Boundary Detection Technique: Flashlight Scene Determination*. Signal Processing and Its Applications, vol. 1, pag. 447–450, august 1999.

[Heng 01] W.J. Heng & K.N. Ngan. *Enhanced Shot Boundary Refinement for Post-Shot Boundary Detection*. IEEE International Conference on Electrical and Electronic Technology, vol. 1, pag. 259–263, 2001.

[Hinneburg 00] A. Hinneburg & D.A. Keim. *Clustering Techniques for Large Data Sets From the Past to the Future*. IEEE International Conference on Bioinformatics and Biomedical Egineering, pag. 43–49, 2000.

[Horn 81] B. Horn & B. Schunck. *Determining optical flow*. Artificial Intelligence, vol. 17, pag. 185–203, 1981.

[Houten 03] Y.Van Houten, M.Van Setten & J.-G. Schuurman. *Patch-Based Video Browsing*. Human-Computer Interaction INTERACT, 2003.

[Hsu 02] C.-T. Hsu & S.-J. Teng. *Motion Trajectory Based Video Indexing and Retrieval*. IEEE International Conference on Image Processing, vol. 1, pag. 605–608, septembrie, New York, USA 2002.

[Huang 98] J. Huang & Y. Wang Z. Liu. *Integration of Audio and Visual Information for Content-Based Video Segmentation.* IEEE International Conference on Image Processing, vol. 3, pag. 526–530, octombrie, Chicago, USA 1998.

[Ionescu 05a] B. Ionescu, D. Coquin, P. Lambert & V. Buzuloiu. *Analysis and Characterization of Animation Movies.* ORASIS Journées Francophones des Jeunes Chercheurs en Vision par Ordinateur, mai, Fournois, Puy-de-Dôme-France 2005.

[Ionescu 05b] B. Ionescu, D. Coquin, P. Lambert & V. Buzuloiu. *An Approach to Scene Detection in Animation Movies and Its Applications.* Buletin Ştiinţific Universitatea Politehnica din Bucureşti, vol. C (67), nr. 2, pag. 45–57, 2005.

[Ionescu 05c] B. Ionescu, P. Lambert, D. Coquin & L. Dîlea. *Color-Based Semantic Characterization of Cartoons.* IEEE International Symposium on Signals, Circuits and Systems, Special Issue on Statistical Models in Image Processing, vol. 1, pag. 223–226, iulie, Iaşi, România 2005.

[Ionescu 06a] B. Ionescu, V. Buzuloiu, P. Lambert & D. Coquin. *Improved Cut Detection for the Segmentation of Animation Movies.* IEEE International Conference on Acoustic, Speech and Signal Processing, mai, Toulouse-France 2006.

[Ionescu 06b] B. Ionescu, P. Lambert, D. Coquin, L. Ott & V. Buzuloiu. *Animation Movies Trailer Computation.* ACM Multimedia, octombrie, Santa Barbara, CA, USA 2006.

[Ionescu 07a] B. Ionescu. *Caractérisation Symbolique de Séquences d'Images: Application aux Films d'Animation.* Teză de doctorat, Université de Savoie, LISTIC, `http://www.listic.univ-savoie.fr/`, Annecy-France 2007.

[Ionescu 07b] B. Ionescu, P. Lambert, D. Coquin & V. Buzuloiu. *Caractérisation du Mouvement dans les Films d'Animation.* Buletin Ştiinţific Universitatea Politehnica din Bucureşti, vol. C (69), nr. 2, 2007.

[Ionescu 07c] B. Ionescu, P. Lambert, D. Coquin & V. Buzuloiu. *The Cut Detection Issue in the Animation Movie Domain*. Academy Publisher Journal of Multimedia, vol. 2, nr. 4, pag. 10–19, 2007.

[Ionescu 08] B. Ionescu, D. Coquin, P. Lambert & V. Buzuloiu. *A Fuzzy Color-Based Approach for Understanding Animated Movies Content in the Indexing Task*. Eurasip Journal on Image and Video Processing, Special Issue on Color in Image and Video Processing, 2008.

[Irani 95] M. Irani, P. Anandan & S. Hsu. *Mosaic Based Representations of Video Sequences and Their Applications*. Computer Vision, pag. 605–611, iunie 1995.

[IRISA 05] IRISA. *Motion2D*. `http://www.irisa.fr/vista/Motion2D/`, Rennes, France 2005.

[Itten 61] J. Itten. *The Art of Color: The Subjective Experience and Objective Rationale of Color*. New York: Reinhold, 1961.

[Jain 91] J.R. Jain & A.K. Jain. *Displacement measurement and its application in interframe image coding*. IEEE Transactions on Communications, vol. 29, nr. 12, pag. 1799–1806, decembrie 1991.

[Jain 99] A.K. Jain, M.N. Murty & P.J. Flynn. *Data Clustering: A Review*. ACM Computing Surveys, vol. 31, nr. 3, pag. 264–323, septembrie 1999.

[Jeannin 01] S. Jeannin & A. Divakaran. *MPEG-7 Visual Motion Descriptors*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, nr. 6, pag. 720–724, iunie 2001.

[Jin 02] R. Jin, Y. Qi & A. Hauptmann. *A Probabilistic Model for Camera Zoom Detection*. The Sixteenth Conference of the International Association for Pattern Recognition, nr. 3, pag. 859– 862, august, Quebec-Canada 2002.

[Johnson 67] S.C. Johnson. *Hierarchical Clustering Schemes*. Psychometrika, vol. 2, pag. 241–254, 1967.

[Kang 01] H.-B. Kang. *A Hierarchical Approach to Scene Segmentation.* IEEE Workshop on Content-Based Access of Image and Video Libraries, 2001.

[Kay 03] P. Kay & T. Regier. *Resolving the Question of Color Naming Universals.* National Academy of Sciences, vol. 100, nr. 15, 2003.

[Kelly 76] K.L. Kelly & D.B. Judd. *Color: Universal Language and Dictionary of Names.* National Bureau of Standards, 1976.

[Kim 00a] C. Kim & J.N. Hwang. *An Integrated Scheme for Object-Based Video Abstraction.* ACM Multimedia, pag. 303 – 311, 2000.

[Kim 00b] E.Y. Kim, K.I. Kim, K. Jung & H.J. Kim. *A Video Indexing System Using Character Recognition.* IEEE International Conference on Consumer Electronics, pag. 358–359, Los Angles, CA, USA 2000.

[Kim 02] S.H. Kim & R.H. Park. *Robust Video Indexing for Video Sequences with Complex Brightness Variation.* IASTED International Conference on Signal and Image Processing, pag. 410–414, Kauai, Hawaii 2002.

[Kim 04] J.-G. Kim, H.S. Chang, J. Kim & H.-M. Kim. *Threshold-Based Camera Motion Characterization of MPEG Video.* Electronics and Telecommunications Research Institute Journal, vol. 26, nr. 3, pag. 269–272, iunie 2004.

[Klir 95] G. J. Klir & B. Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications.* Prentice Hall, New Jersey, 1995.

[Kobla 99] V. Kobla, D. DeMenthon & D. Doermann. *Special Effect Edit Detection Using VideoTrails: A Comparison with Existing Techniques.* SPIE Storage and Retrieval for Image and Video Databases VII, vol. 3656, pag. 302–313, 1999.

[Kobla 00] V. Kobla, D. DeMenthon & D. Doermann. *Identifying Sports Video Using Replay, Text and Camera Motion Features.* SPIE Storage and Retrieval for Media Database, vol. 3972, pag. 332–343, 2000.

[Kosch 01] H. Kosch, L. Boszormányi, A. Bachlechner, B. Dorflingera, C. Hanin, C. Hofbauer, M. Lang, C. Riedler & R. Tusch. *SMOOTH - A Distributed Multimedia Database System.* International Conference on Very Large Data Bases, Rome, Italy 2001.

[Kotsiantis 07] S.B. Kotsiantis. *Supervised Machine Learning: A Review of Classification Techniques.* Informatica, vol. 31, pag. 249–268, 2007.

[Kramer 05] P. Kramer & J.B. Pineau. *Camera Motion Detection in the Rough Index Paradigm.* TREC Video Retrieval Evaluation Online Proceedings, TRECVID, noiembrie 2005.

[Kwok 04] S.H. Kwok, A.G. Constantinides & W.-C. Siu. *An Efficient Recursive Shortest Spanning Tree Algorithm Using Linking Properties.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, nr. 6, pag. 852–863, 2004.

[Kyungpook 06] National University Kyungpook. *Artificial Intelligence Laboratory.* `http://ailab.kyungpook.ac.kr/vindex/video-view.html`, 2006.

[Laganière 08] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Pays & B. Ionescu. *Video Summarization from Spatio-Temporal Features.* ACM International Conference on Multimedia, Trecvid BBC Rushes Summarization Workshop, octombrie, Vancouver-Canada 2008.

[Latecki 01] L.J. Latecki, D.D. Widldt & J. Hu. *Extraction of Key Frames from Videos by Optimal Color Composition Matching and Polygon Simplification.* Multimedia Signal Processing Conference, pag. 245–250, Cannes, France 2001.

[Lay 04] J.A. Lay & L. Guan. *Retrieval for Color Artistry Concepts.* IEEE Transactions on Image Processing, vol. 13, nr. 3, pag. 125–129, martie 2004.

[Lecce 99] V. Di Lecce, G. Dimauro, A. Guerriero, S. Impedovo, G. Pirlo & A. Salzo. *Image Basic Features Indexing*

Techniques for Video Skimming. International Conference on Image Analysis and Processing, pag. 715–720, 1999.

[Lee 01] M.-S. Lee, Y.-M. Yang & S.-W. Lee. *Automatic Video Parsing Using Shot Boundary Detection and Camera Operation Analysis*. Pattern Recognition, vol. 34, pag. 711–719, 2001.

[Lee 02a] H. Lee & S.-D. Kim. *Rate-Driven Key Frame Selection Using Temporal Variation of Visual Content*. Electronics Letters, vol. 38, nr. 5, pag. 217–218, februarie 2002.

[Lee 02b] S. Lee & M.H. Hayes. *Real-Time Camera Motion Classification for Content-Based Indexing and Retrieval using Templates*. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pag. 3664–3667, 2002.

[Lescieux 06] M. Lescieux. *Introduction à la Logique Floue: Plan du Cours*. `http://auto.polytech.univ-tours.fr/automatique/AUA/ressources/Introduct%ion_logique_floue.ppt`, 2006.

[Li 01] Y. Li, T. Zhang & D. Tretter. *An Overview of Video Abstraction Techniques*. HP Laboratories, HPL-2001-191, 2001.

[Li 03] Y. Li, S. Narayanan & C.-C.J. Kuo. *Movie Content Analysis, Indexing and Skimming via Multimodal Information*. Video Mining, Chapter 5, Eds. Kluwer Academic Publishers, 2003.

[Li 04] Z. Li, G. Schuster, A.K. Katsaggelos & B. Gandhi. *Optimal Video Summarization With a Bit Budget Constraint*. IEEE International Conference on Image Processing, vol. 1, pag. 617–620, Singapore 2004.

[Lienhart 97] R. Lienhart, C. Kuhmunch & W. Effelsberg. *On the Detection and Recognition of Television Commercials*. IEEE Conference on Multimedia Computing and Systems, pag. 509–516, Ottawa, Canada 1997.

[Lienhart 99a] R. Lienhart. *Comparison of Automatic Shot Boundary Detection Algorithms.* SPIE Storage and Retrieval for Still Image and Video Databases VII, vol. 3656, pag. 290–301, 1999.

[Lienhart 99b] R. Lienhart, S. Pfeiffer & W. Effelsberg. *Scene Determination Based on Video and Audio Features.* IEEE International Conference on Multimedia, Computing and Systems, vol. 1, pag. 685–690, iunie, Florence-Italy 1999.

[Lienhart 00] R. Lienhart. *Dynamic Video Summarization of Home Video.* SPIE Storage and Retrieval for Media Databases, vol. 3972, pag. 378–389, ianuarie 2000.

[Lienhart 01a] R. Lienhart. *Reliable Dissolve Detection.* SPIE Storage and Retrieval for Media Databases, vol. 4315, pag. 219–230, ianuarie 2001.

[Lienhart 01b] R. Lienhart. *Reliable Transition Detection in Videos: A Survey and Practitiner's Guide.* MRL, Intel Corporation, `http://www.lienhart.de/Publications/IJIG_AUG2001.pdf`, august, Santa Clara, USA 2001.

[Lim 01] S.H. Lim & A. El Gamal. *Optical Flow Estimation Using High Frame Rate Sequences.* IEEE International Conference on Image Processing, vol. 2, pag. 925–928, octombrie 2001.

[Lin 98] C.-W. Lin, Y.-J. Chang & Y.-C. Chen. *Hierarchical Motion Estimation Algorithm Based on Pyramidal Successive Elimination.* International Computer Symposium, octombrie 1998.

[Lin 02] W.-H. Lin & A.G. Hauptmann. *News Video Classification Using SVM-Based Multimodal Classifiers and Combination Strategies.* ACM Multimedia, pag. 323–326, Juan-les-Pins, France 2002.

[Liu 02a] C.-C. Liu & A.L.P. Chen. *3D-List: A Data Structure for Efficient Video Query Processing.* IEEE Transactions on Knowledge and Data Engineering, vol. 14, nr. 1, pag. 106–122, ianuarie-februarie 2002.

[Liu 02b] T. Liu & J.R. Kender. *An Efficient Error-Minimizing Algorithm for Variable-Rate Temporal Video Sampling.* IEEE International Conference on Multimedia and Expo, vol. 1, pag. 413–416, 2002.

[Liu 02c] T. Liu & J.R. Kender. *Optimization Algorithms for the Selection of Key Frames Sequences of Variable Length.* European Conference on Computer Vision, vol. 2353, pag. 403–417, London, UK 2002.

[Liu 03] T. Liu, H.-J. Zhang & F. Qi. *A Novel Video Key-Frame Extraction Algorithm Based on Perceived Motion Energy Model.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, nr. 10, pag. 1006–1013, octombrie 2003.

[Liu 04] T. Liu, X. Zhang, J. Freg & K. Lo. *Shot Reconstruction Degree: A Novel Criterion for Keyframe Selection.* Pattern Recognition Letter, vol. 25, nr. 12, pag. 1451–1457, septembrie 2004.

[Lu 03] S. Lu, I. King & M. Lyu. *Video Summarization Using Greedy Method in a Constraint Satisfaction Framework.* 9th International Conference on Distributed Multimedia Systems, pag. 456–461, Miami, Florida, USA 2003.

[Lundmark 01] A. Lundmark. *Non-Redundant Search Patterns in Log-Search Motion Estimation.* Swedish Society for Automated Image Analysis Symposium - SSAB, 2001.

[Lupatini 98] G. Lupatini, C. Saraceno & R. Leonardi. *Scene Break Detection: A Comparison.* Research Issues in Data Engineering, Workshop on Continuous Media Databases and Applications, pag. 34–41, Orlando, FL, USA 1998.

[Ma 01] Y.F. Ma, J. Sheng, Y. Chen & H.J. Zhang. *MSR-Asia at TREC-10 Video Track: Shot Boundary Detection Task.* 10th Text Retrieval Conference, pag. 371, 2001.

[MacQueen 67] J.B. MacQueen. *Some Methods for Classification and Analysis of Multivariate Observations.* 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, vol. 1, pag. 281–297, 1967.

[Maillet 03] S.M. Maillet. *Content-Based Video Retrieval: An Overview.* `http://viper.unige.ch/~marchand/CBVR/`, 2003.

[Marichal 98] X. Marichal. *Motion Estimation and Compensation for Very Low Bitrate Video Coding.* Teză de doctorat, UCL - Université Catholique de Louvain, Laboratoire de Télécommunications et Télédétection, Louvain-la-Neuve, Belgique 1998.

[Mazière 00] M. Mazière, F. Chassaing, L. Garrido & P. Salembier. *Segmentation and Tracking of Video Objects for a Content-Based Video Indexing Context.* IEEE International Conference on Multimedia Computing and Systems, vol. 2, pag. 1191–1194, New York, USA 2000.

[Mehtre 97] B.M. Mehtre, M.S. Kankanhalli & W.F. Lee. *Shape Measures for Content Based Image Retrieval: A Comparison.* Information Processing and Management, vol. 33, nr. 3, pag. 319–337, mai, 1997.

[Meng 95] J. Meng, Y. Juan & S.F. Chang. *Scene Change Detection in a MPEG Compressed Video Sequence.* SPIE Symposium, vol. 2419, pag. 14–25, februarie 1995.

[Miene 01] A. Miene, A. Dammeyer, T. Hermes & O. Herzog. *Advanced and Adaptive Shot Boundary Detection.* ECDL WS Generalized Documents, pag. 39–43, 2001.

[Miura 03] K. Miura, R. Hamada, I. Ide, S. Sakai & H. Tanaka. *Motion Based Automatic Abstraction of Cooking Videos.* IPSJ Transactions on Computer Vision and Image Media, vol. 44, 2003.

[Mojsilovic 00] A. Mojsilovic, J. Kovacevic, R.J. Safranek J. Hu & S.K. Ganapathy. *Matching and Retrieval Based on the Vocabulary and Grammar of Color Patterns.* IEEE Transactions on Image Processing, vol. 9, nr. 1, pag. 38–54, 2000.

[Morphing 08] Morphing. *Introduction to Media Computation.* `http://coweb.cc.gatech.edu/mediaComp-plan/65`, 2008.

[Nagasaka 92] A. Nagasaka & Y. Tanaka. *Automatic Video Indexing and Full-Video Search for Object Appearances*. Visual Database Systems II, pag. 113–127, Amsterdam, Netherlands 1992.

[Nam 98] J. Nam, M. Alghoniemy & A.H. Tewfik. *Audio-Visual Content-Based Violent Scene Characterization*. IEEE International Conference on Image Processing, vol. 1, pag. 353–357, Chicago, USA 1998.

[Nam 00] J. Nam & A.H. Tewfik. *Dissolve Transition Detection Using B-Spline Interpolation*. IEEE International Conference on Multimedia and Expo, vol. 3, pag. 1349–1352, iulie 2000.

[Naphade 01a] M.R. Naphade. *A Probabilistic Framework for Mapping Audio-Visual Features to High-Level Semantics in Terms of Concepts and Context*. Teză de doctorat, Department of Electrical and Computing Engineering, University of Illinois, 2001.

[Naphade 01b] M.R. Naphade & T.S. Huang. *A Probabilistic Framework for Semantic Video Indexing, Filtering and Retrieval*. IEEE Transactions on Multimedia, vol. 3, nr. 1, pag. 141–151, 2001.

[Naphade 02] M.R. Naphade & T.S. Huang. *Extracting Semantics from Audiovisual Content: The Final Frontier in Multimedia Retrieval*. IEEE Transactions on Neural Networks, vol. 13, nr. 2, pag. 793–810, 2002.

[Netravali 79] A.N. Netravali & J.D. Robbins. *Motion-compensated television coding: Part I*. BELL System Technical Journal, vol. 58, nr. 3, pag. 631–670, martie 1979.

[Ngo 00] C.-W. Ngo, T.-C. Pong, H.-J. Zhang & R.T. Chin. *Motion Characterization by Temporal Slices Analysis*. IEEE International Conference on Computer Vision and Pattern Recognition, vol. 2, pag. 768–773, 2000.

[Ngo 03] C.-W. Ngo, Y.-F. Ma & H.-J. Zhang. *Automatic Video Summarization by Graph Modeling*. IEEE International Conference on Computer Vision, vol. 1, pag. 104, Nice, France 2003.

[Niculescu 06] R.S. Niculescu, T.M. Mitchell & R.B. Rao. *Bayesian Network Learning with Parameter Constraints*. Journal of Machine Learning Research, vol. 7, pag. 1357–1383, 2006.

[Ohta 80] Y.I. Ohta, T. Kanade & T. Sakai. *Color Information for Region Segmentation*. Computer Graphics amd Image Processing, vol. 13, pag. 222–241, 1980.

[Otsuji 91] K. Otsuji, Y. Tonomura & Y. Ohba. *Video Browsing Using Brightness Data*. SPIE Visual Communications and Image Processing, vol. 1606, pag. 980–989, 1991.

[Ott 07] L. Ott, P. Lambert, B. Ionescu & D. Coquin. *Animation Movie Abstraction: Key Frame Adaptative Selection based on Color Histogram Filtering*. International Conference on Image Analysis and Processing, Computational Color Imaging Workshop, septembrie, Modena, Italy 2007.

[Pan 00] C. Pan & S. Ma. *3D Motion Estimation of Human by Genetic Algorithm*. 15th International Conference on Pattern Recognition, vol. 1, pag. 1159, 2000.

[Pan 01] H. Pan, P. Beek & M. Sezan. *Detection of Slow-Motion Replay Segments in Sports Video for Highlights Generation*. IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 3, pag. 1649–1652, Salt Lake City, Utah, SUA 2001.

[Papoulis 91] Papoulis. *Probability, Random Variables, and Stochastic Processes*. Mc Graw Hill, Inc., New-York, 3rd edition, 1991.

[Pfeiffer 96] S. Pfeiffer, R. Lienhart, S. Fisher & W. Effelsberg. *Abstracting Digital Movies Automatically*. Journal of Visual Communication and Image Representation, vol. 7, nr. 4, decembrie 1996.

[Pilu 97] M. Pilu. *On Using Raw MPEG Motion Vectors To Determine Global Camera Motion*. HP - Hewlett Packard, `http://www.hpl.hp.com/techreports/97/HPL-97-102.pdf`, august 1997.

[Pineau 05] J.B. Pineau. *Extraction des Objets Couleur en Mouvement des Séquences Vidéo*. LABRI UMR CNRS 5800, `http://www.labri.fr/ImageetSon/AIV`, 2005.

[Porter 00] S.V. Porter, M. Mirmehdi & B.T. Thomas. *Video Cut Detection Using Frequency Domain Correlation*. 15th International Conference on Pattern Recognition, vol. 3, pag. 413–416, Barcelona, Spain 2000.

[Porter 01] S.V. Porter, M. Mirmehdi & B.T. Thomas. *Detection and Classification of Shot Transitions*. 12th British Machine Vision Conference, pag. 73–82, septembrie 2001.

[Porter 03] S.V. Porter, M. Mirmehdi & B.T. Thomas. *A Shortest Path Representation for Video Summarization*. 12th International Conference on Image Analysis and Processing, pag. 460–465, septembrie 2003.

[QBIC 03] IBM QBIC. *Hermitage Museum - Query by Image Content*. `http://www.hermitage-museum.org`, 2003.

[Qian 99] R. Qian, N. Haering & I. Sezan. *A Computational Approach to Semantic Event Detection*. Computer Vision and Pattern Recognition, vol. 1, pag. 206, iunie 1999.

[Radhakrishnan 04] R. Radhakrishnan, A. Divakaran & Z. Xiong. *A Time Series Clustering Based Framework for Multimedia Mining and Summarization Using Audio Features*. ACM International Workshop on Multimedia Information Retrieval, pag. 157–164, New York, USA 2004.

[Rasheed 03] Z. Rasheed & M. Shah. *Scene Detection in Hollywood Movies and TV Shows*. IEEE Computer Vision and Pattern Recognition, vol. 2, pag. 343–351, Wisconsin, USA 2003.

[Ren 03] W. Ren & S. Singh. *Video Transition: Modeling and Prediction*. Pattern Analysis and Neural Networks, PANN, `http://www.dcs.ex.ac.uk/research/pann/pdf/pann_SS_089.PDF`, 2003.

[Reoxiang 94] L. Reoxiang, Z. Bing & M.L. Liou. *A New Three-Step Search Algorithm for Block Motion Estimation*. IEEE

Transactions on Circuits and Systems for Video Technology, vol. 4, nr. 4, pag. 438–442, august 1994.

[Research 05] Compaq Corporate Research. *Audio Search Using Speech Recognition.* `http://speechbot.research.compaq.com`, 2005.

[Rivlin 95] E. Rivlin & I. Weiss. *Local Invariants for Recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, nr. 3, pag. 226–238, 1995.

[Rong 04] J. Rong, W. Jin & L. Wu. *Key Frame Extraction Using Inter-Shot Information.* IEEE International Conference on Multimedia and Expo, vol. 1, Taipei, Taiwan 2004.

[Rosenblatt 62] F. Rosenblatt. *Principles of Neurodynamics.* Spartan, New York, 1962.

[Rowe 01] L.A. Rowe, D. Harley, P. Pletcher & S. Lawrence. *BIBS: A Lecture Webcasting System.* Berkley Multimedia Research Center, TR 2001-160, iunie 2001.

[Rumelhart 86] D.E. Rumelhart, G.E. Hinton & R.J. Williams. *Learning Internal Representations by Error Propagation.* Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press, Cambridge, vol. 1, pag. 318–362, 1986.

[Sanson 81] H. Sanson. *Motion Affine Models Identification and Application to Television Image Coding.* Visual Communication and Image Processing, vol. 1605, nr. 2, pag. 570–581, noiembrie 1981.

[Saraceno 98] C. Saraceno & R. Leonardi. *Identification of Story Units in AV Sequencies by Joint Audio and Video Processing.* IEEE International Conference on Image Processing, vol. 1, pag. 363–367, octombrie, Chicago, USA 1998.

[Saur 97] D.D. Saur, Y.P. Tan, S.R. Kulkarni & P.J. Ramadge. *Automated Analysis and Annotation of Basketball Video.* SPIE Symposium on Storage and Retrieval for Image and Video Databases V, vol. 3022, pag. 176–187, 1997.

[Scaringella 06] N. Scaringella, G. Zoia & D. Mlynek. *Automatic Genre Classification of Music Content*. IEEE Signal Processing Magazine, vol. 23, nr. 2, pag. 133–141, martie 2006.

[Schmid 97] C. Schmid & R. Mohr. *Local Grayvalue Invariants for Image Retrieval*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, nr. 5, pag. 530–535, 1997.

[Schneiderman 00] H. Schneiderman & T. Kanade. *A Statistical Method for 3D Object Detection Applied to Faces and Cars*. IEEE Computer Vision and Pattern Recognition, vol. 1, pag. 746–751, Hilton Head Island, SC, USA 2000.

[Shahraray 95] B. Shahraray. *Scene Change Detection and Content-Based Sampling of Video Sequences*. SPIE Conference on Digital Video Compression: Algorithms and Technologies, vol. 2419, pag. 2–13, februarie 1995.

[Shen 97] B. Shen. *HDH Based Compressed Video Cut Detection*. Visual 97, pag. 149–156, decembrie, San Diego, USA 1997.

[Smeulders 00] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta & R. Jain. *Content-Based Image Retrieval at the End of the Early Years*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, nr. 12, pag. 1349–1380, decembrie 2000.

[Smith 98] M.A. Smith & T. Kanade. *Video Skimming and Characterization Through the Combination of Image and Language Understanding*. International Workshop on Content-Based Access of Image and Video Databases, Bombay, India 1998.

[Smith 04] P. Smith, T. Drummond & R. Cipolla. *Layered Motion Segmentation and Depth Ordering by Tracking Edges*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, nr. 4, pag. 479–492, aprilie 2004.

[Smith 99] J.R. Smith. *VideoZoom Spatial-Temporal Video Browsing*. IEEE Transactions on Multimedia, vol. 1, iunie 99.

[Snoek 05a] C.G.M. Snoek & M. Worring. *Multimedia Event Based Video Indexing Using Time Intervals*. IEEE Transactions on Multimedia, vol. 7, nr. 4, august 2005.

[Snoek 05b] C.G.M. Snoek & M. Worring. *Multimodal Video Indexing: A Review of the State-of-the-art*. Multimedia Tool and Applications, vol. 25, nr. 1, pag. 5–35, 2005.

[Song 02] H.S. Song, I.K. Kim & N.I. Cho. *Scene Change Detection by Feature Extraction from Strong Edge Blocks*. SPIE Visual Communications and Image Processing, vol. 4671, pag. 784–792, 2002.

[Stricker 95] M. Stricker & M. Orengo. *Similarity of color images*. SPIE Conference on Storage and Retrieval for Image and Video Databases, vol. 2420, pag. 381–392, 1995.

[Su 05a] C-.W Su, H.-Y.M. Liao, H.-R. Tyan, K.-C. Fan & L.-H. Chen. *A Motion-Tolerant Dissolve Detection Algorithm*. IEEE Transactions on Multimedia, vol. 7, nr. 6, pag. 1106–1113, 2005.

[Su 05b] C.W. Su, H.R. Tyan, H.Y. Mark Liao & L.H. Chen. *A Motion-Tolerant Dissolve Detection Algorithm*. IEEE Transactions on Multimedia, vol. 7, nr. 6, pag. 1106–1113, 2005.

[Sun 00] X. Sun & M.S. Kankanhalli. *Video Summarization Using R-Sequences*. Journal of Real Time Imaging, vol. 6, pag. 449–459, 2000.

[Sundaram 02] H. Sundaram & S.-F. Chang. *Video Skims: Taxonomies and an Optimal Generation Framework*. IEEE International Conference on Image Processing, vol. 2, pag. 21–24, Rochester, USA 2002.

[Taniguchi 95] Y. Taniguchi, A. Akutsu, Y. Tonomura & H. Hamada. *An Intuitive and Efficient Access Interface to Real-Time Incoming Video Based on Automatic Indexing*. ACM Multimedia, pag. 25–33, San Francisco, California, United States 1995.

[Tardini 05] G. Tardini, C. Grana, R. Marchi & R. Cucchiara. *Shot Detection and Motion Analysis for Automatic MPEG-7*

Annotation of Sports Videos. 13th International Conference on Image Analysis and Processing, pag. 653–660, septembrie 2005.

[Timoner 01] S.J. Timoner & D.M. Freeman. *Multi-image gradient-based algorithms for motion estimation*. SPIE Optical Engineering, vol. 40, nr. 9, pag. 2003–2016, 2001.

[Tong 01] S. Tong & E. Chang. *Support Vector Machine Active Learning for Image Retrieval*. ACM Multimedia Conf., vol. 9, Ottawa, Canada 2001.

[Trecvid 08] Trecvid. *Video Retrieval Evaluation*. `http://www-nlpir.nist.gov/projects/trecvid/`, 2008.

[Trémeau 04] A. Trémeau, C. Fernandez-Maloigne & P. Bonton. *Image Numérique Couleur: De l'Acquisition au Traitement*. DUNOD ISBN 2-10-006843-1, 2004.

[Truong 00a] B.T. Truong & C. Dorai. *Automatic Genre Identification for Content-Based Video Categorization*. IEEE International Conference on Pattern Recognition, vol. 4, pag. 230–233, Barcelona, Spain 2000.

[Truong 00b] B.T. Truong, C. Dorai & S. Venkatesh. *New Enhancements to Cut, Fade, and Dissolve Detection Processes in Video Segmentation*. ACM Multimedia, pag. 219–227, noiembrie 2000.

[Truong 07] B.T. Truong & S. Venkatesh. *Video Abstraction: A Systematic Review and Classification*. ACM Transactions on Multimedia Computing, Communications and Applications, vol. 3, nr. 1, 2007.

[Turaga 98] D. Turaga & M. Alkanhal. *Search Algorithms for Block-Matching in Motion Estimation*. `http://www.ece.cmu.edu/~ee899/project/deepak_mid.htm`, 1998.

[Vasconcelos 00] N. Vasconcelos & A. Lippman. *Statistical Models of Video Structure for Content Analysis and Characterization*. IEEE Transactions on Image Processing, vol. 9, pag. 3–19, ianuarie 2000.

[Vendriga 01] J. Vendriga & M. Worring. *Evaluation of Logical Story Unit Segmentation in Video Sequences*. IEEE International Conference on Multimedia and Expo, pag. 1092–1095, Tokyo, Japan 2001.

[Venkataraman 99] A. Venkataraman. *Decision Trees*. `http://www.speech.sri.com/people/anand/771/html/node28.html`, 1999.

[Vermaak 02] J. Vermaak, P. Prez, M. Gangnet & A. Blake. *Rapid Summarization and Browsing of Video Sequences*. British Machine Vision Conference, vol. 1, pag. 424–433, Cardiff, UK 2002.

[Vertan 04] C. Vertan, B. Ionescu, I. Stefan & M. Ciuc. *Osteoporosis detection in calcaneum X-ray images by digital image processing: fractal and statistical approaches*. Interdisciplinary Applications of Fractal and Chaos Theory, Eds. R. Dobrescu, C. Vasilescu, Editura Academiei Române (ISBN 973-27-1070-5), 2004.

[Vertan 08] C. Vertan. *Prelucrarea şi Analiza Imaginilor Color*. Suport de curs, laboratorul LAPI - Laboratorul de Analiza şi Prelucrarea Imaginilor, Universitatea Politehnica din Bucureşti, `http://alpha.imag.pub.ro/ro/cursuri/paic/index.html`, 2008.

[Visibone 06] Visibone. *Webmaster Palette*. `http://www.visibone.com/colorlab`, 2006.

[Vogl 99] S. Vogl, K.Manske & M. Muhlhauser. *A VRML Approach to Web Video Browsing*. Multimedia Computing and Networking, pag. 276–285, San Jose, USA 1999.

[Wallin 01] H. Wallin, C. Christopoulos & F. Furesjo. *Robust parametric motion estimation for image mosaicing in the MPEG-7 standard*. IEEE International Conference on Image Processing, vol. 2, pag. 961–964, octombrie 2001.

[Wang 92] L.-X. Wang. *Fuzzy Systems are Universal Approximators*. IEEE Conference on Fuzzy Systems, pag. 1163–1170, San Diego, USA 1992.

[Wang 00] Y. Wang, Z. Liu & J.-C. Huang. *Multimedia Content Analysis Using Both Audio and Visual Clues*. IEEE Signal Processing Magazine, vol. 17, nr. 6, pag. 12–36, noiembrie 2000.

[Wang 01] J.Z. Wang, J. Li & G. Wiederhold. *SIMPLIcity: Semantic-Sensitive Integrated Matching for Picture Libraries*. IEEE Transactions on Pattern Analisys and Machine Intelligence, vol. 23, nr. 9, pag. 947–963, 2001.

[Welling 05] M. Welling. *Support Vector Machines*. Note de curs, University of Toronto, Department of Computer Science, Canada, `http://www.ics.uci.edu/~welling/classnotes/papers_class/SVM.pdf`, 2005.

[Wikipedia 08] Wikipedia. *The Free Online Encyclopedia*. `http://www.wikipedia.org/`, 2008.

[Witten 05] I.H. Witten & E. Frank. *Data Minning - Practical Machine Learning Tools and Techniques*. Elsevier, Morgan Kaufman Publishers, second edition, pag. 265–270, 2005.

[Wolf 96] W. Wolf. *Key Frame Selection by Motion Analysis*. IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pag. 1228–1231, 1996.

[Wyszecki 82] G. Wyszecki & W.S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley and sons, second edition, 1982.

[Xiong 97] W. Xiong, J. C.-M. Lee & R.-H. Ma. *Automatic Video Data Structuring Through Shot Partitioning and Key Frame Computing*. Machine Vision and Applications, vol. 10, nr. 2, pag. 51–65, 1997.

[Xiong 03] Z. Xiong, R. Radhakrishnan & A. Divakaran. *Generation of Sports Highlights Using Motion Activity in Combination With a Common Audio Feature Extraction Framework*. IEEE International Conference on Image Processing, vol. 1, Barcelona, Spain 2003.

[Yahiaoui 01] L. Yahiaoui, B. Merialdo & B. Huet. *Automatic Video Summarization*. CBMIR, 2001.

[Yao 01] A. Yao & J. Jin. *The Developing of a Video Metadata Authoring and Browsing System in XML.* ACM International Conference - Pan-Sydney Workshop on Visual Information Processing, vol. 2, Sydney, Australia 2001.

[Yeo 95] B.-L. Yeo & B. Liu. *Rapid Scene Analysis on Compressed Video.* IEEE Transactions on Circuits, Systems and Video Technology, vol. 5, pag. 533–544, decembrie 1995.

[Young 02] T. Young. *On the Theory of Light and Colors.* Philosophical Transactions of the Royal Society, nr. 91, 1802.

[Yu 97] H. Yu, G. Bozdagi & S. Harrington. *Feature-Based Hierarchical Video Segmentation.* IEEE International Conference on Image Processing, vol. 2, pag. 498–501, 1997.

[Yu 03] B. Yu, W.-Y. Ma, K. Nahrstedt & H.-J. Zhang. *Video Summarization Based on User Log Enhanced Link Analysis.* ACM Multimedia, pag. 382–391, Berkeley, USA 2003.

[Yu 04] X.-D. Yu, L. Wang, Q. Tian & P. Xue. *Multi-level Video Representation With Application to Keyframe Extraction.* International Conference on Multimedia Modeling, pag. 117–121, Brisban, Australia 2004.

[Yuan 95] Y. Yuan & M.J. Shaw. *Induction of Fuzzy Decision Trees.* Fuzzy Sets and Systems, vol. 69, pag. 125–139, 1995.

[Zabih 95] R. Zabih, J. Miller & K. Mai. *A Feature-Based Algorithm for Detecting and Classifying Scene Breaks.* ACM Multimedia, pag. 189–200, noiembrie, San Francesco, USA 1995.

[Zabih 99] R. Zabih, J. Miller & K. Mai. *A Feature-Based Algorithm for Detecting and Classification Production Effects.* Multimedia Systems, vol. 7, pag. 119–128, 1999.

[Zadeh 65] L.A. Zadeh. *Fuzzy Sets.* Information and Control, vol. 8, nr. 3, pag. 338–353, 1965.

[Zahariadis 96] T. Zahariadis & D. Kalivas. *A Spiral Search Algorithm For Fast Estimation Of Block Motion Vectors*. Signal Processing VIII, Theories and Applications. Eighth European Signal Processing Conference, vol. 2, pag. 1079–1082, 1996.

[Zaim 01] M. Zaim, A. El Ouaazizi & R. Benslimane. *Genetic Algorithms Based Motion Estimation*. Vision Interface Annual Conference, iunie, Ottawa, Canada 2001.

[Zeng 02] W. Zeng, W. Gao & D. Zhao. *Video Indexing by Motion Activity Maps*. IEEE International Conference on Image Processing, vol. 1, pag. 912–915, 2002.

[Zhang 93] H. Zhang, A. Kankanhalli & S.W. Smoliar. *Automatic Partitioning of Full-Motion Video*. Multimedia Systems, vol. 1, nr. 1, pag. 10–28, 1993.

[Zhang 94] H. Zhang, C.Y. Low, Y. Gong & S.W. Smoliar. *Video Parsing Using Compressed Data*. SPIE Image and Video Processing II, vol. 2182, pag. 142–149, februarie 1994.

[Zhang 97] H.J. Zhang, J. Wu, D. Zhong & S.W. Smoliar. *An Integrated System for Content-Based Video Retrieval and Browsing*. Pattern Recognition, vol. 30, nr. 4, pag. 643–658, 1997.

[Zhao 03] M. Zhao, J. Bu & C. Chen. *Audio and Video Combined for Home Video Abstraction*. IEEE International Conference on Acoustic, Speech and Signal Processing, vol. 5, pag. 620–623, Hong Kong, China 2003.

[Zhong 96] D. Zhong, H. Zhang & C. Chang. *Clustering Methods for Video Browsing and Annotation*. SPIE Storage and Retrieval for Still Image and Video Databases IV, vol. 2670, pag. 239–246, 1996.

[Zhong 97] D. Zhong & S.-F. Chang. *Spatio-temporal Video Search using the Object-Based Video Representation*. IEEE International Conference on Image Processing, vol. 1, pag. 1–12, 1997.

[Zhou 02] W. Zhou, S. Dao & C.-C.J. Kuo. *On-Line Knowledge and Rule-Based Video Classification System for Video Indexing and Dissemination.* Information Systems, vol. 27, nr. 8, 2002.

[Zhu 05] C.-Z. Zhu, T. Mei & X.-S. Hua. *Video Booklet - Natural Video Browsing.* ACM Multimedia, pag. 265 – 266, noiembrie, Singapore 2005.

[Zhuang 98] Y. Zhuang, Y. Rui, T.S. Huang & S. Mehrota. *Adaptive Key Frame Extraction Using Unsupervized Clustering.* IEEE International Conference on Image Processing, vol. 1, pag. 866–870, 1998.