

Facultatea de Electronică, Telecomunicații și Tehnologie Informației

AIM AI Multimedia Lab
https://www.aimmultimedialab.ro/

Universitatea Politehnică din București

Tehnici de analiză și clasificare automată a informației

Prof. dr. ing. Bogdan IONESCU
<https://bionescu.aimmultimedialab.ro/>

București, 2022

1

Plan Curs

- M1. Introducere (concept, aplicații)
- M2. Prelucrarea și reprezentarea datelor de intrare
- M3. Tehnici de clasificare ne-supervizată ("clustering")
- M4. Tehnici de clasificare supervizată ("classification")
- M5. Evaluarea performanței clasificatorilor

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

2

> M2. Prelucrarea și reprezentarea datelor de intrare

- 2.1. [Reprezentarea datelor]
- 2.2. [Descrierea conținutului]
- 2.3. [Normalizarea datelor]
- 2.4. [Decorelarea datelor]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

3

Reprezentarea datelor

> exemplu date meteo:

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicăată	nu	Nu
3	ploioasă	cald	ridicăată	nu	?
4	innorat	rece	normală	da	?

set de date

atribute

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

4

Reprezentarea datelor (cont.)

> exemplu date meteo (cont.):

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicăată	nu	Nu
3	ploioasă	cald	ridicăată	nu	Da
4	innorat	rece	normală	da	Nu

> prin **valori simbolice**, ex. Atributul *vreme*={"însorită", "ploioasă", "innorat"} sau *vânt*={"da", "nu"}, etc;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

5

Reprezentarea datelor (cont.)

> exemplu date meteo (cont.):

nr.	vreme	temperatură	umiditate	vânt	sport
1	1	30	50	15	1
2	0.9	39	90	1	0
3	0.5	30	90	0	0
4	0.1	17	50	17	1

> prin **valori numerice**, ex. atributul *vreme*=[0;1], *temperatură*=[15;40]°, *umiditate*=[40;100]%, *vânt*=[0;30]km/h, *sport*={0,1}; de regulă datele au o semnificație fizică (ex. grade, procent, etc).

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

6

Reprezentarea datelor (cont.)

> exemplu date meteo (cont.):

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	30	50	15	Da
2	însorită	39	90	1	Nu
3	ploioasă	30	90	0	Da
4	înnorat	17	50	17	Nu

> prin **valori mixte**, numeric - simbolic;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

7

7

Reprezentarea datelor (cont.)

> cazuri particulare:

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
0	0	0	0	0	10	0	0	0	0	0	1	0	0	0	0
2	0	0	0	0	0	0	7	0	0	0	0	0	6	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	12	2	0	0	0	0	0	0	0	0

> ce se observă în mod particular la aceste date?

> **"sparse data"** – "date rarefiate": foarte multe dintre valori sunt 0 iar valorile utile sunt împrăștiate; necesită metode de prelucrare adaptate datorită influențării modului de calcul al similarității;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

8

8

Reprezentarea datelor (cont.)

> cazuri particulare (cont.):

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
1		2	1	0	10	7	3	5		4	1	1	1	2	6
2	0	0	2	3	1		7	6	4	5	4	4	6	2	1
0	0	1	3	2	2	6	6	8	8		9	2	3		2
1	0		5	4	6	12	2	5	6	8	2	3	5	5	8

> ce se observă în mod particular la aceste date?

> **"missing data"**: datele lipsă au o semnificație aparte și pot surveni din diferite motive, ex. nu sunt înregistrate, nu sunt relevante (în afara scalei), nu sunt cunoscute, sunt voluntare (decizia de a nu măsura un atribut), etc.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

9

9

Reprezentarea datelor (cont.)

> cazuri particulare (cont.):

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
1	999	2	1	0	10	7	3	5	4	4	1	1	1	2	6
2	0	0	2	3	1	878	7	6	4	5	4	4	6	2	1
0	0	1	3	2	2	6	6	8	8	4	9	2	3	0	2
2	0	0	2	3	1	878	7	6	4	5	4	4	6	2	1

> ce se observă în mod particular la aceste date?

> **"innacurate data"**: date eronate ce pot surveni din erori umane, introduse deliberat (ex. nu se cunoaște valoarea) sau din cauza zgomotului de măsurare; vizibile ca "outliers" (deviază semnificativ de la plaja de valori uzuală) sau date duplicat.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

10

10

Reprezentarea datelor (cont.)

> cazuri particulare (cont.):

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
0.1	999	-10	0	0	10	-50	5	5	0.1	4	1	-99	1	2	64
0	630	-4	5	3	1	40	7	6	0.4	5	4	99	6	2	63
0.5	123	5	3	2	2	-30	6	8	0.2	4	9	0	3	2	65
1	546	10	1	3	1	46	7	6	0.5	5	4	40	6	2	61

> ce se observă în mod particular la aceste date?

> atributele au valori în **plaje foarte diferite** dar totuși sunt folosite unitar sub formă de descriptor; acest lucru va influența evaluarea similarității între acestea – normalizare?

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

11

11

Reprezentarea datelor (cont.)

> cazuri particulare (cont.):

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
1	1	2	3	5	8	13	4	17	21	21	42	45	90	91	90
2	3	5	8	9	9	18	27	26	53	52	99	99	99	99	99
0	1	1	2	3	5	8	13	21	34	55	89	90	4	3	4
1	2	3	4	5	6	7	8	9	10	11	11	12	14	14	16

> ce se observă în mod particular la aceste date?

> atributele tind să fie **corelate** între ele ceea ce înseamnă că anumite valori pot fi deduse unele din altele – reprezentare inefficientă – decorare?

concluzie: datele trebuie verificate preliminar (vizual, grafic)!

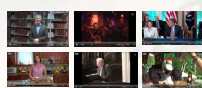
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

12

12

Descrerea conținutului datelor

> în practică, de cele mai multe ori informația de analizat nu poate fi reprezentată în forma în care a fost "înregistrată", ex. imagini:



date



[200 255 80 90]	[200 255 80 90]	[200 255 80 90]
[150 200 67 0]	[150 200 67 0]	[150 200 67 0]
[160 30 45 0]	[160 30 45 0]	[160 30 45 0]
[0 0 200 200]	[0 0 200 200]	[0 0 200 200]
[200 255 80 90]	[200 255 80 90]	[200 255 80 90]
[150 200 67 0]	[150 200 67 0]	[150 200 67 0]
[160 30 45 0]	[160 30 45 0]	[160 30 45 0]
[0 0 200 200]	[0 0 200 200]	[0 0 200 200]

informația înregistrată
(valori pixeli pe 8 biți)

> *motivație*: valorile nu sunt reprezentative pentru conținutul semantic al datelor; necesară o reprezentare cât mai discriminativă;

descriptor (observație, vector de caracteristici) = reprezentarea unei date într-o formă mai compactă, de regulă vectorială. Valorile vectorului reprezintă măsurători ale unor proprietăți definitorii ale datei respective.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

13

13

Descrierea conținutului datelor multimedia

[sursă platformă YouTube]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

14

14

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: imagini/video (revisited)

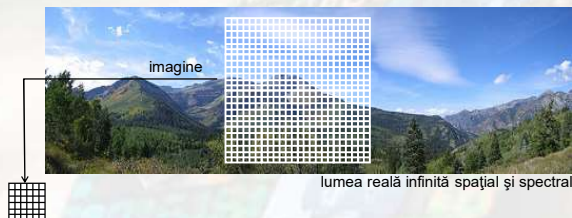


image digitală = ansamblu de valori plasate după o formă spațială într-un spațiu cel puțin bidimensional (matrice)

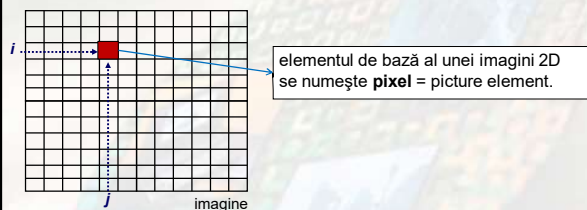
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

15

15

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: imagini/video (cont.)



fiecare pixel al imaginii este caracterizat prin:

- **valoare** → legată de "culoarea" transmisă (ex. alb, roșu, etc.)
- **poziție în imagine** → informație spațială.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

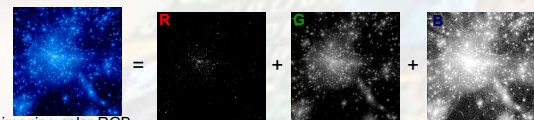
16

16

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: imagini/video (cont.)

↳ **imagini** = valoarea unui pixel este un vector:



↳ **video** = o succesiune temporală de imagini:



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

17

17

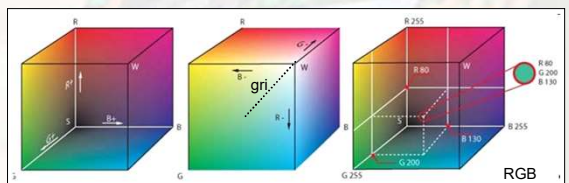
Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

1. informația de culoare;

- cea mai frecvent folosită pentru descrierea conținutului imaginilor; sistemul vizual uman este bazat pe prelucrarea informației de culoare (unde luminoase de diverse frecvențe).

> alegerea modului de reprezentare al culorilor: **Red-Green-Blue**



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

18

18

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

1. **informația de culoare (cont.);**

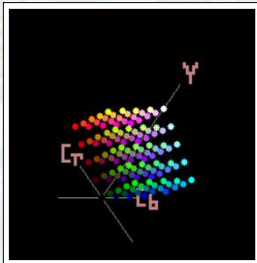
> alegerea modului de reprezentare al culorilor: **YCbCr** (separare intensitate de cromaticitate)

$$Y = 0,299 \cdot R + 0,587 \cdot G + 0,114 \cdot B$$

$$C_b = a_1 \cdot (R - Y) + b_1 \cdot (B - Y)$$

$$C_r = a_2 \cdot (R - Y) + b_2 \cdot (B - Y)$$

- **YIQ**: NTSC (National Television System Committee, 30 imagini/s),
- **YUV**: PAL (Phase Alternating Line, 50Hz întrețesut),
- **YDbDr**: SECAM (Séquentiel Couleur à Mémoire, 25 imagini/s)



[sursă imagine Wikipedia]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

19

19

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

1. **informația de culoare (cont.);**

> alegerea modului de reprezentare al culorilor: **L*a*b*** (distanța perceptuală dintre culori ~ distanța matematică)

$$L^* = \begin{cases} 116 \cdot y^{1/3} - 16 & \text{daca } y > 0.008856 \\ 903.3 \cdot y & \text{altfel} \end{cases}$$

$$a^* = 500 \cdot [f(x) - f(y)] \quad \text{unde: } x = \frac{X}{X_W}, y = \frac{Y}{Y_W}, z = \frac{Z}{Z_W}$$

$$b^* = 200 \cdot [f(y) - f(z)] \quad \text{iar } (X_W, Y_W, Z_W) = \text{alb de referință}$$

$$f(t) = \begin{cases} t^{1/3} & t > 0.008856 \\ 7.787 \cdot y + 0.137931 & \text{altfel} \end{cases}$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

20

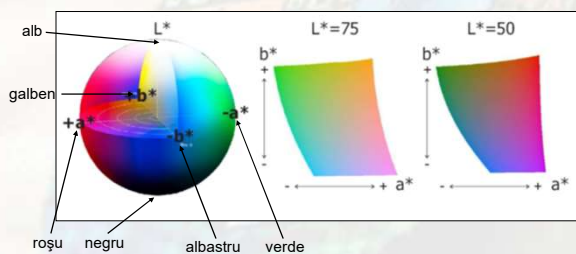
20

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

1. **informația de culoare (cont.);**

> alegerea modului de reprezentare al culorilor: **L*a*b*** (cont.)



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

21

21

Descrierea conținutului datelor multimedia (cont.)

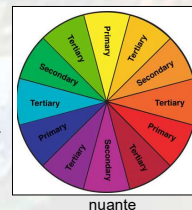
A. **informația vizuală:** surse de informație disponibile

1. **informația de culoare (cont.);**

> alegerea modului de reprezentare al culorilor: **HSV** (separare informație de nuanță, saturație și intensitate)

↳ în general proprietățile unei culori sunt definite de trei informații:

nuanță (hue) = un alt cuvânt pentru culoare, "pigmentul", culoarea pură, etc.



nuanțe

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

22

22

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

1. **informația de culoare (cont.);**

> alegerea modului de reprezentare al culorilor: **HSV** (cont.)

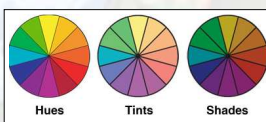
saturație (saturation) = intensitatea culorii, grad de puritate, cum arată în condiții de iluminare diferită.

saturație: max. → slab

intensitate (intensity, value) = gradul de luminozitate, de strălucire.

amestec cu gri a unei nuanțe pure → **ton de culoare**,

intensitate: >luminoase, <luminoase



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

23

23

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

1. **informația de culoare (cont.);**

> alegerea modului de reprezentare al culorilor: **HSV** (separare informație de nuanță, saturație și intensitate)

$$h_{HSV} = \begin{cases} 0 & \text{max} = \min \\ 60 \cdot \frac{g-b}{\max - \min} + 0 & \text{max} = r, g \geq b \\ 60 \cdot \frac{g-b}{\max - \min} + 360 & \text{max} = r, g < b \\ 60 \cdot \frac{b-r}{\max - \min} + 120 & \text{max} = g \\ 60 \cdot \frac{r-g}{\max - \min} + 240 & \text{max} = b \end{cases}$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

24

24

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

1. informația de culoare (cont.);

> alegerea modului de reprezentare al culorilor: **HSV** (cont.)

$$s_{HSV} = \begin{cases} 0 & \text{max} = 0 \\ \frac{\max - \min}{\max} & \text{altfel} \end{cases}$$

$$v_{HSV} = \max$$

unde: $r = \frac{R}{255}, g = \frac{G}{255}, b = \frac{B}{255}$

$$\max = \max\{r, g, b\},$$

$$\min = \min\{r, g, b\}$$

$$h_{HSV} \in [0; 360]$$

$$s_{HSV}, v_{HSV} \in [0; 1]$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

25

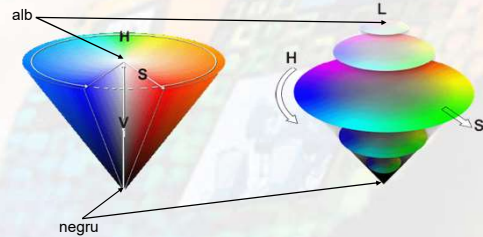
25

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

1. informația de culoare (cont.);

> alegerea modului de reprezentare al culorilor: **HSV** (cont.)



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

26

26

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

1. informația de culoare (cont.);

> descriptori uzuali:

- histograme de culoare calculate în diverse spații de culoare, histograme ponderate, culori predominante, varianța de culoare, parametri de intensitate, descrierea repartiției spațiale a culorilor, prezența culorii pielii (prezență umană);

$$h(D) = \frac{1}{X \cdot Y} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} \delta(A(x, y) - D), \quad \delta(x) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases}$$

unde D este un nivel de gri (sau culoare), $D=0, \dots, L-1$ (sau ia valori în paleta de culoare considerată), XY dimensiunea imaginii, $A()$ este funcția imagine.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

27

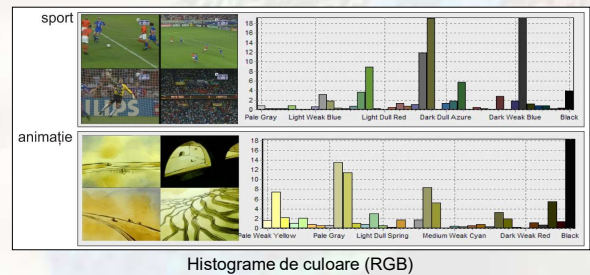
27

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

1. informația de culoare (cont.);

> descriptori uzuali (cont.):



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

28

28

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

1. informația de culoare (cont.);

> exemplu: histograma color



R	99	83	161	188	188
	79	66	123	210	164
	84	74	135	187	158
	82	76	116	170	227
	113	130	152	180	213
G	186	183	174	107	166
	177	174	172	130	100
	177	171	192	145	78
	178	166	170	155	104
	190	200	198	132	133
B	223	222	205	163	148
	219	217	217	170	169
	219	213	217	164	131
	219	212	188	131	83
	230	185	210	137	134

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

29

29

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

1. informația de culoare (cont.);

> exemplu: histograma color (cont.)

- 16.777.216 culori posibile (8x8x8 biți), cuantizare la un număr mai mic de valori;

- alegere n bini pentru fiecare axă: R, G și B ($n \times n \times n$ valori);

- exemplu: $n=3$,
bin 1 – [0 - 84],
bin 2 – [85 - 170],
bin 3 – [171 - 255].

R	99	83	161	188	188
	79	66	123	210	164
	84	74	135	187	158
	82	76	116	170	227
	113	130	152	180	213
G	186	183	174	107	166
	177	174	172	130	100
	177	171	192	145	78
	178	166	170	155	104
	190	200	198	132	133
B	223	222	205	163	148
	219	217	217	170	169
	219	213	217	164	131
	219	212	188	131	83
	230	185	210	137	134

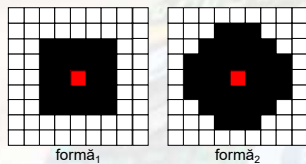
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

30

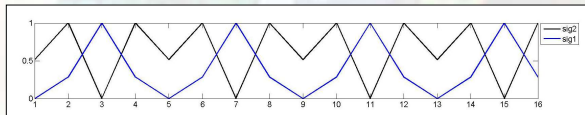
30

Descrierea conținutului datelor multimedia (cont.)

- A. **informația vizuală:** surse de informație disponibile
 2. **informația relativă la forme (cont.);**
 > exemplu: semnătură distanță față de centroid (cont.)



- simetrie formă se transformă în periodicitate semnătură;
- limitări?



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

37

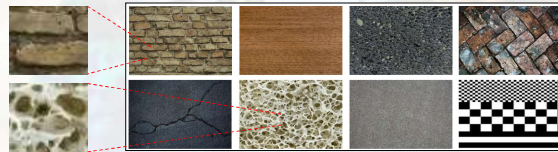
37

Descrierea conținutului datelor multimedia (cont.)

- A. **informația vizuală:** surse de informație disponibile
 3. **informația relativă la textură;**

- caracterizarea proprietăților materialelor prezente în imagini (atăt analiză de culoare cât și de contur).

textură = o regiune din imagine ce prezintă caracteristici omogene, precum un motiv de bază ce se repetă (domeniu spațial/frecvențial).



texturi

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

38

38

Descrierea conținutului datelor multimedia (cont.)

- A. **informația vizuală:** surse de informație disponibile
 3. **informația relativă la textură (cont.);**
 > descriptori uzuali:

- analiza *statistică* a distribuției spațiale a intensităților pixelilor din imagine; ex. probabilitatea de co-ocurență a unei anumite intensități în diverse direcții și distanțe față de un punct de referință;
- analiza proprietăților *geometrice* ale elementelor texturii, precum arie, formă, lungime și a modului de distribuție al acestora într-o anumită rețea;
- analiză pe bază de *modele* al căror parametri descriu proprietățile esențiale ale acestuia (modelare pe bază de puncte întunecate sau luminoase, tranziții verticale sau orizontale, linii, etc);
- analiză pe bază de filtre, ex. operatorii de derivare (Laplacian, Roberts) sau filtrele Gabor.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

39

39

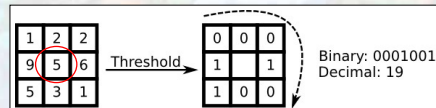
Descrierea conținutului datelor multimedia (cont.)

- A. **informația vizuală:** surse de informație disponibile
 3. **informația relativă la textură (cont.);**
 > exemplu: Local Binary Patterns (LBP)

p1: imaginea este împărțită în N macro-blocuri (ex. 16x16, 32x32);

p2: pentru fiecare macro-bloc, fiecare pixel este comparat cu cei 8 vecini astfel:

- intensitate $>$, se scrie valoare 0;
- intensitate $<$, se scrie valoare 1;



p3: pixelul curent va avea asociat astfel un șablon binar (pattern);

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

40

40

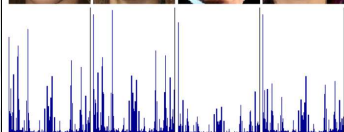
Descrierea conținutului datelor multimedia (cont.)

- A. **informația vizuală:** surse de informație disponibile
 3. **informația relativă la textură (cont.);**
 > exemplu: Local Binary Patterns (cont.) [PRACE Summer of HPC 2013]

p4: pentru macro-blocul curent se realizează o histogramă a frecvenței șabloanelor pixelilor (număr de valori 2^8);



p5: pentru imagine, LBP este dat de concatenarea tuturor histogramelor macro-blocurilor (dim. N^*2^8).



LBP pentru comparare imagini cu fețe

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

41

41

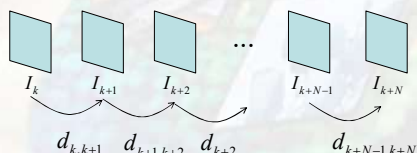
Descrierea conținutului datelor multimedia (cont.)

- A. **informația vizuală:** surse de informație disponibile
 4. **informația relativă la mișcare;**

- caracterizarea schimbărilor (spațiale) survenite într-o secvență de imagini (mișcare obiecte, mișcare scenă, etc).

> presupune detecția în prealabil a acestor schimbări:

(1) detecția de mișcare (exemplu diferență cadre succesive);



unde I_k este imaginea la momentul k ,
 $d_{k,k+1}$ este diferența dintre I_k și I_{k+1}

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

42

42

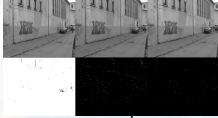
Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

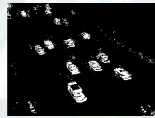
4. informația relativă la mișcare (cont.);

(1) detecția de mișcare (exemplu diferență cadre succesive; cont.);

decizie: dacă $d_{k,k+1}(n) > \tau$
 $\Rightarrow$ pixelul $n=(x,y)$ se mișcă;



exemplu



maskă, $\tau = 25$



maskă, $\tau = 50$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

43

43

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

4. informația relativă la mișcare (cont.);

(2) estimarea mișcării;



Exemplu de vectori de mișcare (imagine film "Matrix")

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

44

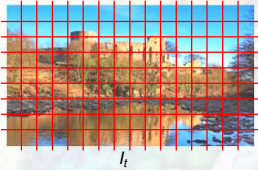
44

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

4. informația relativă la mișcare (cont.);

(2) estimarea mișcării (exemplu metode pe blocuri de pixeli);



I_t

- imaginea curentă analizată la momentul t , I_t , este împărțită în blocuri disjuncte de pixeli de dimensiune $B \times B$;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

45

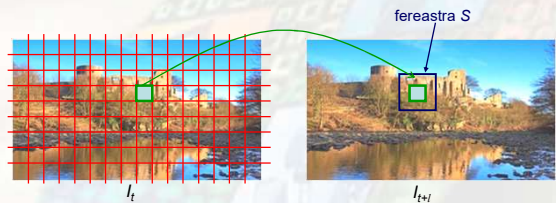
45

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

4. informația relativă la mișcare (cont.);

(2) estimarea mișcării (exemplu metode pe blocuri de pixeli; cont.);



- pentru fiecare bloc din I_t se caută noua sa poziție în imaginea următoare la momentul $t+1$ (căutarea se face doar într-o fereastră S de $(2B+1) \times (2B+1)$ pixeli);

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

46

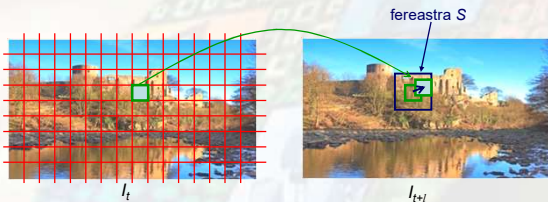
46

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

4. informația relativă la mișcare (cont.);

(2) estimarea mișcării (exemplu metode pe blocuri de pixeli; cont.);



- noua poziție a blocului curent este determinată pe baza minimizării unei funcții de cost, F_C , ce estimează eroarea de aproximare a blocului curent cu blocurile din fereastră S .

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

47

47

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

4. informația relativă la mișcare (cont.);

(2) estimarea mișcării (exemplu metode pe blocuri de pixeli; cont.);

↳ Mean Absolute Difference (MAD)

$$MAD(b_1, b_2) = \frac{1}{X \cdot Y} \sum_{x=1}^X \sum_{y=1}^Y |b_1(x, y) - b_2(x, y)|$$

unde b_1 și b_2 reprezintă două blocuri de pixeli iar X și Y dimensiunile imaginii.

↳ Mean Square Difference (MSD)

$$MSD(b_1, b_2) = \frac{1}{X \cdot Y} \sum_{x=1}^X \sum_{y=1}^Y [b_1(x, y) - b_2(x, y)]^2$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

48

48

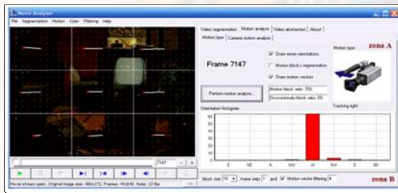
Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

4. informația relativă la mișcare (cont.);

(2) estimarea mișcării (exemplu metode pe blocuri de pixeli; cont.);

> demo:



a. HDuncomp: frame 17;
b. The Holiday frame 31623;
frame 45796;
frame 109038;
frame 1;
frame 7636.

[http://alpha.imag.pub.ro/VideoIndexingRP2/]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

49

49

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

4. informația relativă la mișcare (cont.);

> descriptori uzuali:

- traiectoria obiectelor din scenă,
- identificarea tipului de mișcare a camerei video (apropiere/departare, rotație, translație);
- activitate de mișcare prin cuantizarea varianței amplitudinii vectorilor de mișcare;
- imagini MHI de "istorie a mișcării" (Motion History Images);
- histograme de intensitate a mișcării, etc

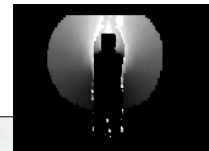


image MHI (normalizare)

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

50

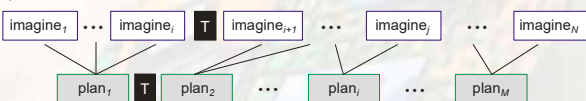
50

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

5. informația relativă la structura temporală;

- caracterizarea proprietăților structurale-temporale ale unei secvențe de imagini.



- image (frame) = "atomul" secvenței;
- plan video (video shot) = secvența continuă de imagini înregistrată de la pornirea camerei până la oprirea acesteia;
- tranziții video (video transition, T) = un efect vizual de scurtă durată ce face legătura între planurile video vecine temporale;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

51

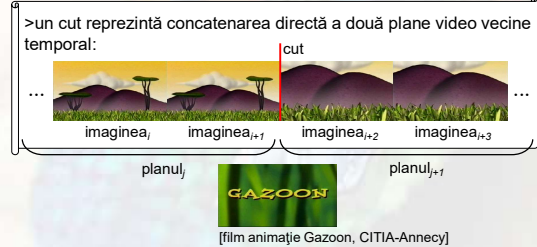
51

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

5. informația relativă la structura temporală (cont.);

> presupune segmentarea temporală în prealabil a secvenței = detecția tranzițiilor video (ex. cut, fade, dissolve, wipes, etc):



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

52

52

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

5. informația relativă la structura temporală (cont.);

> presupune segmentarea temporală în prealabil a secvenței = detecția tranzițiilor video, exemplu (cont.):



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

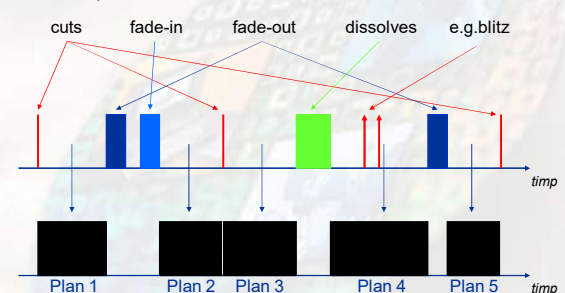
53

53

Descrierea conținutului datelor multimedia (cont.)

A. informația vizuală: surse de informație disponibile

5. informația relativă la structura temporală (cont.);



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

54

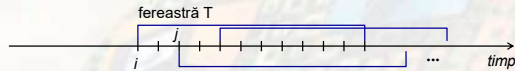
54

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

5. **informația relativă la structura temporală (cont.);**

> descriptori uzuali: activitate vizuală



$\xi_T(i)$ numărul de schimbări de plan ce au loc în fereastra temporală de dimensiune T (ex. 5s) începând cu cadrul la momentul i ;

> evaluând valorile lui $\xi_T(i)$ pentru întreaga secvență (ex. ferestre suprapuse cu un pas p), determinăm ritmul vizual mediu al secvenței:

$$\bar{v}_T = E\{\xi_T(i)\} \sim \text{tempo mediu.}$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

55

55

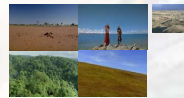
Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

5. **informația relativă la structura temporală (cont.);**

> descriptori uzuali: activitate vizuală (cont.)

$$\bar{v}_T = E\{\xi_T(i)\}$$



documentare

$$\bar{v}_T = 0.2^*$$



știri

$$\bar{v}_T = 0.24^*$$



muzică

$$\bar{v}_T = 0.56^*$$

[*valori medii pentru 30 de exemple de secvențe.]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

56

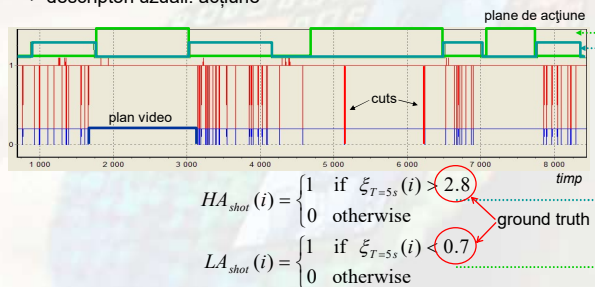
56

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

5. **informația relativă la structura temporală (cont.);**

> descriptori uzuali: acțiune



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

57

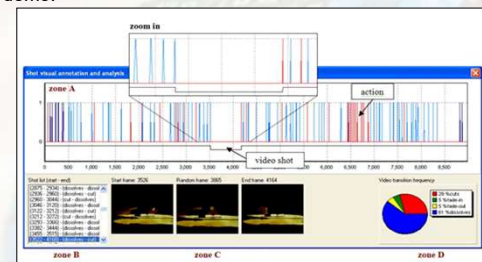
57

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

5. **informația relativă la structura temporală (cont.);**

> demo:



[http://alpha.imag.pub.ro/VideoIndexingRP2/]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

58

58

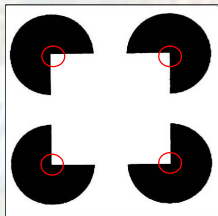
Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături;**

- un caz particular de caracterizare a proprietăților unei clase particulare de puncte de contur ("features").

> **punct de interes** = o regiune de pixeli (de dimensiuni reduse) a căror proprietăți o fac reprezentativă pentru înțelegerea conținutului structural al imaginii:



<http://webvision.med.utah.edu/book/>

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

59

59

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături (cont.);**

> **punct de interes (formalizare)**

- au o definiție matematică bine precizată;
- au o poziție bine definită în imagine;
- informația locală din jurul punctului de interes este bogată informațional (definite de context);
- invariante la perturbații locale și globale (deformări datorate transformărilor de perspectivă, schimbarea unghiului de vizualizare, schimbări de scală, rotații, translații, variații de iluminare, etc).

> descrierea punctelor de interes presupune **detectia în prealabil** a acestora și apoi reprezentarea informației specifice;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

60

60

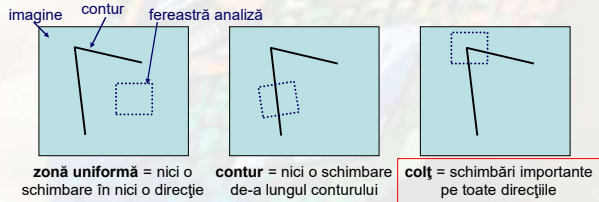
Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături (cont.);**

> detectori uzuali: Harris detector (detector de colțuri)

- bazat pe derivatele de ordin 1 ale imaginii: I_x - derivata pe oX și I_y - derivata pe oY;



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

61

61

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

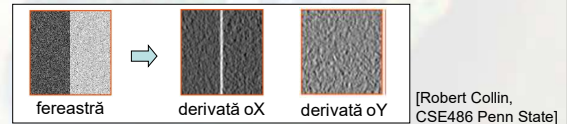
6. **informația relativă la trăsături (cont.);**

> detectori uzuali: Harris detector (cont.)

p1. pentru fiecare pixel din imagine, într-o fereastră, calculăm:

$$M = \sum_{x,y} w(x,y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}$$

unde $w(x,y)$ o funcție de ponderare ce poate fi uniformă, iar x,y iau valori în fereastra considerată (I_x - derivata pe oX, I_y - derivata pe oY).



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

62

62

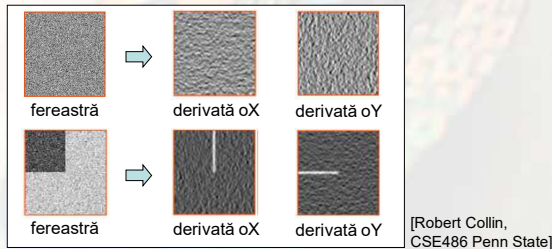
Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături (cont.);**

> detectori uzuali: Harris detector (cont.)

p1. pentru fiecare pixel din imagine, într-o fereastră, calculăm M.



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

63

63

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături (cont.);**

> detectori uzuali: Harris detector (cont.)

p2. estimăm gradul de variație al pixelilor în fereastra considerată:

$$R = \det(M) - k \cdot \text{Tr}^2(M)$$

unde k este o constantă iar $\text{Tr}()$ este operatorul "trace" ce returnează suma elementelor de pe diagonala principală;

- valori mari pentru colțuri;
- valori negative mari pentru contur;
- valori mici în modul pentru regiuni omogene.

p3. punctele de interes sunt obținute pe baza filtrării valorilor lui R:

dacă $(R > \tau) \ \&\& \ (R == \max \text{ local}) \Rightarrow \text{colț obiect}$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

64

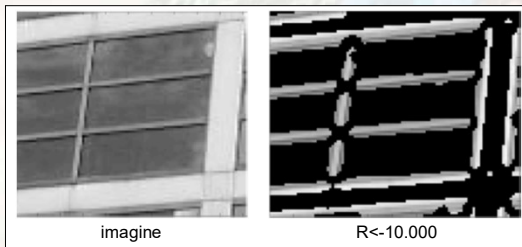
64

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături (cont.);**

> detectori uzuali: Harris detector (cont.)



[Robert Collin, CSE486 Penn State]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

65

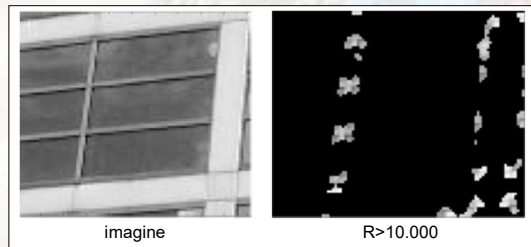
65

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături (cont.);**

> detectori uzuali: Harris detector (cont.)



[Robert Collin, CSE486 Penn State]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

66

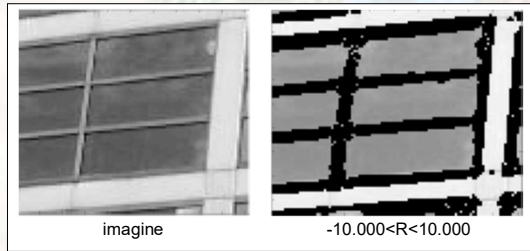
66

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături (cont.);**

> detectori uzuali: Harris detector (cont.)



[Robert Collin, CSE486 Penn State]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

67

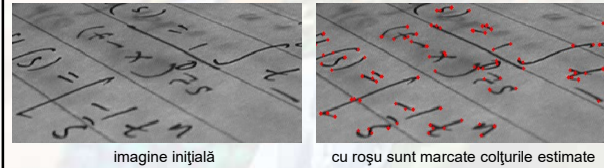
67

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături (cont.);**

> detectori uzuali: Harris detector (cont.)



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

68

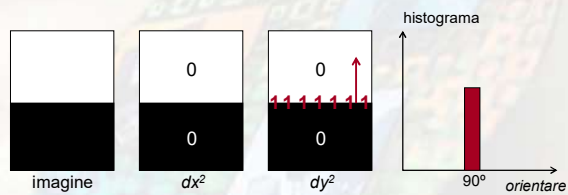
68

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături (cont.);**

> descriptori uzuali: Histograms of oriented Gradients (HoG)



$$d = \sqrt{dx^2 + dy^2}, \alpha = \arctan \frac{dy}{dx}$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

69

69

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

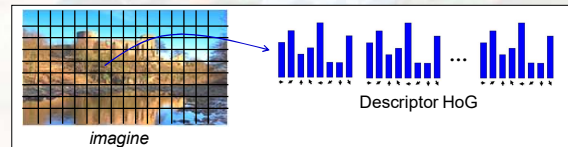
6. **informația relativă la trăsături (cont.);**

> descriptori uzuali: Histograms of oriented Gradients (HoG; cont.)

p1. imaginea este împărțită în blocuri disjuncte de pixeli;

p2. pentru fiecare bloc se calculează o histogramă de orientări ale gradientului pixelilor;

p3. descriptor = concatenarea tuturor histogramelor *normalizate*.



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

70

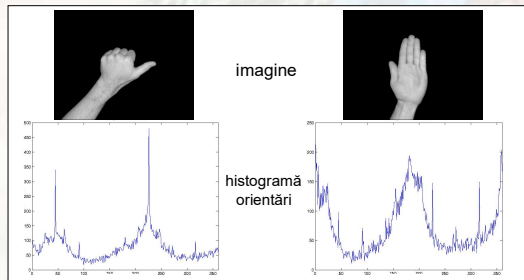
70

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături (cont.);**

> descriptori uzuali: Histograms of oriented Gradients (HoG; cont.)



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

71

71

Descrierea conținutului datelor multimedia (cont.)

A. **informația vizuală:** surse de informație disponibile

6. **informația relativă la trăsături (cont.);**

> detectori/descriptori uzuali (cont.)

Detectori:

- Harris Laplace (Harris reprezentat la diverse niveluri de scală);
- reprezentări "scale-space": Laplacian of Gaussian (LoG), Difference of Gaussian (DoG), Determinant of Hessian (DoH), Maximally Stable Extremum Regions (MSER), Scale Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF);
- reprezentări spațio-temporale: Harris 3D, Cuboid, Hessian 3D, SURF 3D, etc.

Descriptori:

- reprezentări Bag-of-VisualWords (BoW), Histograms of Optical Flow (HoF), Vector of Locally Aggregated Descriptors (VLAD), Fisher Kernel (FK), etc.

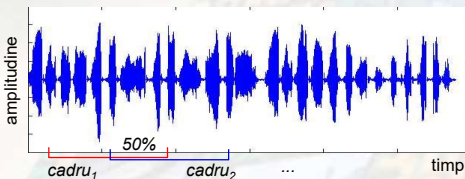
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

72

72

Descrierea conținutului datelor multimedia (cont.)

B. informația audio



sunet digital = o secvență de eșantioane (valori de amplitudine ale undelor sonore) înregistrate în timp.

cadru (audio) = o secvență temporală ce conține un anumit număr de eșantioane (ex. 1024 de eșantioane).

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

73

73

Descrierea conținutului datelor multimedia (cont.)

B. informația audio (cont.)

1. reprezentare temporală;

Descriptorii sunt calculați la nivel de cadru audio și apoi agregați la nivel de înregistrare folosind reprezentări statistice, ex. medie, varianță, median, "skewness", kurtosis (momente statistice).

> descriptori uzuali: Zero Crossing Rate (ZCR)

$$ZCR = \frac{1}{T-1} \sum_{t=1}^{T-1} II\{s_t s_{t-1} < 0\}$$

unde s este un semnal de durată T iar $II\{\cdot\}$ este operatorul care returnează valoarea 1 dacă argumentul este adevărat și 0 altfel.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

74

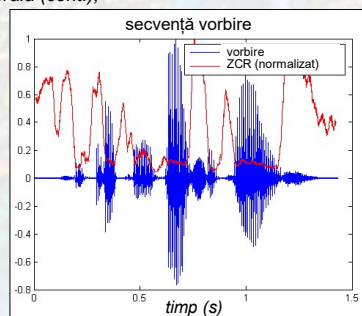
74

Descrierea conținutului datelor multimedia (cont.)

B. informația audio (cont.)

1. reprezentare temporală (cont.);

> descriptori uzuali: Zero Crossing Rate (ZCR; cont.)



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

75

75

Descrierea conținutului datelor multimedia (cont.)

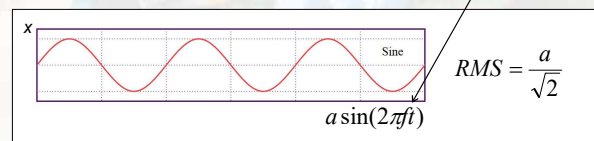
B. informația audio (cont.)

1. reprezentare temporală (cont.);

> descriptori uzuali: energia semnalului (Root Mean Square of Signal Energy - RMS)

$$RMS = \sqrt{\frac{1}{n} (x_1^2 + x_2^2 + \dots + x_n^2)}$$

unde x_i este valoarea semnalului la un moment dat.



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

76

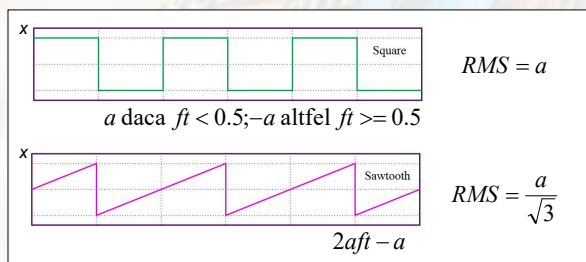
76

Descrierea conținutului datelor multimedia (cont.)

B. informația audio (cont.)

1. reprezentare temporală (cont.);

> descriptori uzuali: energia semnalului (RMS; cont.)



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

77

77

Descrierea conținutului datelor multimedia (cont.)

B. informația audio (cont.)

1. reprezentare temporală (cont.);

> descriptori uzuali (cont.): rata de absență a sunetului, coeficienții de autocorelație ai semnalului, etc.

2. reprezentare frecvențială;

> transformata Fourier 1D discretă (revisited)

> eșantionare:

$$u(m) = f(x_0 + m \cdot \Delta x)$$

unde $f(x)$ este semnalul continuu, x_0 reprezintă coordonata de la care începe eșantionarea (uzual $x_0=0$), Δx este pasul de eșantionare iar $m=0, \dots, N-1$ reprezintă coordonata discretă.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

78

78

Descrierea conținutului datelor multimedia (cont.)

C. informația textuală (cont.)

> descriptori uzuali: TF-IDF (cont.)

- **term frequency-inverse document freq.**, $tfidf(t, d, D)$:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

> sunt filtrate cuvintele care sunt comune (valoare mică idf);

> exemplu:

d_1 ="This is a sample."

d_2 ="This is another example and another situation. Example one and example two are relevant."

D ="This is a sample. This is another example and another situation. Example one and example two are relevant."

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

85

85

Descrierea conținutului datelor multimedia (cont.)

C. informația textuală (cont.)

> descriptori uzuali: TF-IDF (cont.)

> exemplu (cont.):

termen (t)	$tf(t, d_1)$
"this"	1
...	...
"sample"	1

termen (t)	$tf(t, d_2)$
"this"	1
...	...
"example"	3

$$idf("this", D) = \log \frac{2}{1} = 0$$

$$tfidf("this", d_1, D) = 1 \times 0 = 0 = tfidf("this", d_2, D)$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

86

86

Descrierea conținutului datelor multimedia (cont.)

C. informația textuală (cont.)

> descriptori uzuali: TF-IDF (cont.)

> exemplu (cont.):

termen (t)	$tf(t, d_1)$
"this"	1
...	...
"sample"	1

termen (t)	$tf(t, d_2)$
"this"	1
...	...
"example"	3

$$idf("example", D) = \log \frac{2}{1} = 0.3$$

$$tfidf("example", d_2, D) = 3 \times 0.3 = 0.9$$

$$tfidf("example", d_1, D) = 0 \times 0.3 = 0$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

87

87

Descrierea conținutului datelor multimedia (cont.)

C. informația textuală (cont.)

> descriptori uzuali: TF-IDF (cont.)

> *abordare practică*:

- pentru datele de intrare (=corpus) se extrag cuvintele;
- pentru fiecare cuvânt dintr-o instanță (=dată) se calculează $tfidf$;
- descriptorul este dat de concatenarea valorilor $tfidf$ (se folosește același set de cuvinte);

> ce dimensiune are descriptorul rezultat?

> ce particularitate au datele reprezentate în acest fel?
(indicație - setul de cuvinte este același pentru toate datele, indiferent de dimensiunea acestora)

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

88

88

Descrierea conținutului datelor multimedia (cont.)

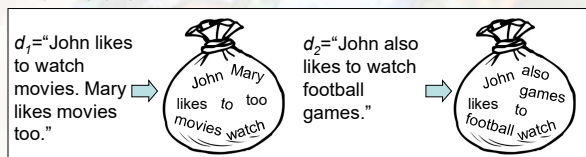
C. informația textuală (cont.)

> descriptori uzuali: Bag-of-Words

> permite reprezentarea datelor de dimensiuni diferite cu descriptori de dimensiune fixă;

> abordare:

1. datele de intrare sunt reprezentate drept mulțimi de cuvinte ("pungi"), ignorând astfel semantica acestora;



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

89

89

Descrierea conținutului datelor multimedia (cont.)

C. informația textuală (cont.)

> descriptori uzuali: Bag-of-Words (cont.)

2. pe baza datele de intrare se construiește un dicționar comun (= reuniunea exclusivă a cuvintelor dintre "pungile" de cuvinte);

dicționar:
1 – "John";
2 – "likes";
3 – "to";
4 – "watch";
5 – "movies";
6 – "also";
7 – "football";
8 – "games";
9 – "Mary";
10 – "too".

3. descriptorul este dat de reprezentarea sub formă de histogramă a frecvenței de apariție a cuvintelor din dicționar:

d_1 ="John likes to watch movies. Mary likes movies too."

$desc_{d1} = [1 \ 2 \ 1 \ 1 \ 2 \ 0 \ 0 \ 1 \ 1]$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

90

90

Descrierea conținutului datelor multimedia (cont.)

C. informația textuală (cont.)

> descriptori uzuali: Bag-of-Words (cont.)

2. pe baza datele de intrare se construiește un dicționar comun (= intersecția cuvintelor dintre "pungile" de cuvinte);

dicționar:
1 – "John";
2 – "likes";
3 – "to";
4 – "watch";
5 – "movies";
6 – "also";
7 – "football";
8 – "games";
9 – "Mary";
10 – "too".

3. descriptorul este dat de reprezentarea sub formă de histogramă a frecvenței de apariție a cuvintelor din dicționar (cont.):

d_2 = "John also likes to watch football games."

$desc_{d2} = [1 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 0]$

> se poate aplica acest principiu la imagini?

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

91

91

Descrierea conținutului datelor multimedia (cont.)

Nivelul semantic de reprezentare al informației

> descriere de nivel scăzut ("low-level")

- descrieri numerice (~sintactica datelor);
- exemplu: histograme, vectori de mișcare, etc.

> descriere de nivel intermediar ("mid-level")

- descrieri simbolice;
- exemplu: denumirea culorilor, detectarea prezenței umane în scenă, etc.

> descriere de nivel semantic ("high-level")

- de regulă descrieri textuale care codează interpretarea datelor pentru a servi unei aplicații specifice;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

92

92

Descrierea conținutului datelor multimedia (cont.)

Nivelul semantic de reprezentare al informației (cont.)



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

93

93

Normalizarea datelor

> descriptori diferiți implică plaje de valori diferite;

$$desc = [a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_m, c_1, c_2, \dots, c_p]$$

desc₁ desc₂ desc₃

date	0.2	0	...	0.9	0	5	...	10	-5	5	...	0
	0	0.6	...	1	10	0	...	2	4	3	...	-5
	0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
	0.1	1	...	0	1	1	...	0	1	-2	...	3
	1	0.4	...	0.5	4	10	...	4	5	4	...	5

> care este plaja de valori? a_i - [0 ; 1], b_i - [0 ; 10], c_i - [-5 ; 5]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

94

94

Normalizarea datelor (cont.)

> compararea (ex. evaluare similaritate) datelor cu valori în același interval este mai eficientă (corectă) -> **normalizare**:

• **normalizarea min-max;**

$$desc = [x_1, x_2, \dots, x_n]$$

$$x'_i = \frac{x_i - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\}}$$

unde $i=1, \dots, n$, $\max\{x_i\}$ și $\min\{x_i\}$ reprezintă operatorii ce returnează valoarea minimă și respectiv maximă a tuturor valorilor descriptorilor (instanțelor) pentru atributul x_i .

> care este plaja de valori rezultată? [0;1]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

95

95

Normalizarea datelor (cont.)

• **normalizarea min-max (cont.);**

$$desc = [x_1, x_2, \dots, x_n], x'_i = \frac{x_i - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\}}$$

date	0.2	0	...	0.9	0	5	...	10	-5	5	...	0
	0	0.6	...	1	10	0	...	2	4	3	...	-5
	0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
	0.1	1	...	0	1	1	...	0	1	-2	...	3
	1	0.4	...	0.5	4	10	...	4	5	4	...	5

$x'_i = \frac{0.6 - 0}{1 - 0} = 0.6$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

96

96

Normalizarea datelor (cont.)

- normalizarea min-max (cont.);

$$desc = [x_1, x_2, \dots, x_n], x'_i = \frac{x_i - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\}}$$

date

0.2	0	...	0.9	0	5	...	10	-5	5	...	0
0	0.6	...	1	10	0	...	2	4	3	...	-5
0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
0.1	1	...	0	1	1	...	0	1	-2	...	3
1	0.4	...	0.5	4	10	...	4	5	4	...	5

$x'_i = \frac{9-0}{10-0} = 0.9$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

97

97

Normalizarea datelor (cont.)

- normalizarea min-max (cont.);

$$desc = [x_1, x_2, \dots, x_n], x'_i = \frac{x_i - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\}}$$

date

0.2	0	...	0.9	0	5	...	10	-5	5	...	0
0	0.6	...	1	10	0	...	2	4	3	...	-5
0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
0.1	1	...	0	1	1	...	0	1	-2	...	3
1	0.4	...	0.5	4	10	...	4	5	4	...	5

$x'_i = \frac{-2+5}{5+5} = 0.3$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

98

98

Normalizarea datelor (cont.)

- normalizarea min-max (cont.);

Date

0.2	0	...	0.9	0	5	...	10	-5	5	...	0
0	0.6	...	1	10	0	...	2	4	3	...	-5
0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
0.1	1	...	0	1	1	...	0	1	-2	...	3
1	0.4	...	0.5	4	10	...	4	5	4	...	5

Date normalizate

0.2	0	...	0.9	0	0.5	...	1	0	1	...	0.5
0	0.6	...	1	1	0	...	0.2	0.9	0.8	...	0
0.1	0.2	...	0.1	0.5	0.9	...	1	0.7	0	...	0.6
0.1	1	...	0	0.1	0.1	...	0	0.6	0.3	...	0.8
1	0.4	...	0.5	0.4	1	...	0.4	1	0.9	...	1

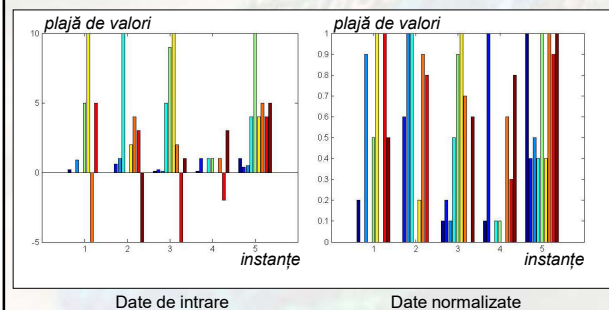
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

99

99

Normalizarea datelor (cont.)

- normalizarea min-max (cont.);



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

100

100

Normalizarea datelor (cont.)

- normalizarea z-score;

$$desc = [x_1, x_2, \dots, x_n]$$

$$x'_i = \frac{x_i - \text{medie}\{x_i\}}{\sigma\{x_i\}}$$

unde $i=1, \dots, n$, $\text{medie}\{x_i\}$ și $\sigma\{x_i\}$ reprezintă operatorii ce returnează valoarea medie și respectiv abaterea pătratică medie a tuturor valorilor descriptorilor (instanțelor) pentru atributul x_i .

> care este plaja de valori rezultată? medie 0 și dispersie 1

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

101

101

Normalizarea datelor (cont.)

- normalizarea z-score (cont.);

$$desc = [x_1, x_2, \dots, x_n], x'_i = \frac{x_i - \text{medie}\{x_i\}}{\sigma\{x_i\}}$$

date

0.2	0	...	0.9	0	5	...	10	-5	5	...	0
0	0.6	...	1	10	0	...	2	4	3	...	-5
0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
0.1	1	...	0	1	1	...	0	1	-2	...	3
1	0.4	...	0.5	4	10	...	4	5	4	...	5

$x'_i = \frac{0.6-0.44}{0.3847} = 0.4159$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

102

102

Normalizarea datelor (cont.)

- normalizarea z-score (cont.);

$$desc = [x_1, x_2, \dots, x_n], x'_i = \frac{x_i - \text{medie}\{x_i\}}{\sigma\{x_i\}}$$

date

0.2	0	...	0.9	0	5	...	10	-5	5	...	0
0	0.6	...	1	10	0	...	2	4	3	...	-5
0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
0.1	1	...	0	1	1	...	0	1	-2	...	3
1	0.4	...	0.5	4	10	...	4	5	4	...	5

$$x'_i = \frac{-2 - 1}{4.3012} = -0.6975$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

103

103

Normalizarea datelor (cont.)

- normalizarea z-score (cont.);

medie=0
σ=1

0.2	0	...	0.9	0	5	...	10	-5	5	...	0
0	0.6	...	1	10	0	...	2	4	3	...	-5
0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
0.1	1	...	0	1	1	...	0	1	-2	...	3
1	0.4	...	0.5	4	10	...	4	5	4	...	5
-0.2	-1.1	...	0.9	-1	0	...	1	-1.6	0.9	...	-0.2
-0.7	0.4	...	1.1	1.5	-1.1	...	-0.7	0.7	0.5	...	-1.5
-0.4	-0.6	...	-0.9	0.2	0.9	...	1	0.1	-1.4	...	0
-0.4	1.5	...	-1.1	-0.8	-0.9	...	-1.1	-0.1	-0.7	...	0.6
1.8	-0.1	...	0	0	1.1	...	-0.3	0.9	0.7	...	1.1

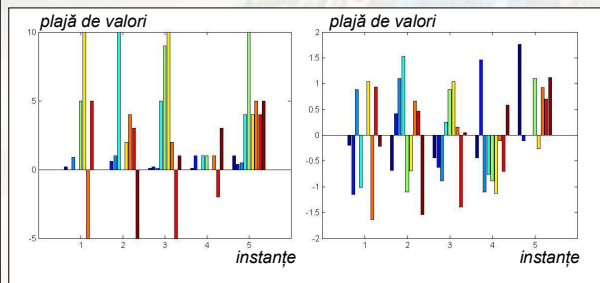
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

104

104

Normalizarea datelor (cont.)

- normalizarea z-score (cont.);



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

105

105

Normalizarea datelor (cont.)

- normalizarea zecimală;

$$desc = [x_1, x_2, \dots, x_n]$$

$$x'_i = \frac{x_i}{10^p}, p = \log_{10}(\max |x_i|)$$

unde $i=1, \dots, n$, $\max\{x_j\}$ și $|x_i|$ reprezintă operatorii ce returnează valoarea maximă și respectiv absolută a tuturor valorilor descriptorilor (instanțelor) pentru atributul x_i , iar p este cel mai mic întreg pentru care $\max(|x'_i|) \leq 1$.

> care este plaja de valori rezultată? mișcă punctul zecimal, $|x'_i| \leq 1$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

106

106

Normalizarea datelor (cont.)

- normalizarea zecimală (cont.);

$$desc = [x_1, x_2, \dots, x_n], x'_i = \frac{x_i}{10^p}, p = \log_{10}(\max |x_i|)$$

date

0.2	0	...	0.9	0	5	...	10	-5	5	...	0
0	0.6	...	1	10	0	...	2	4	3	...	-5
0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
0.1	1	...	0	1	1	...	0	1	-2	...	3
1	0.4	...	0.5	4	10	...	4	5	4	...	5

$$x'_i = \frac{0.6}{10^0} = 0.6$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

107

107

Normalizarea datelor (cont.)

- normalizarea zecimală (cont.);

$$desc = [x_1, x_2, \dots, x_n], x'_i = \frac{x_i}{10^p}, p = \log_{10}(\max |x_i|)$$

date

0.2	0	...	0.9	0	5	...	10	-5	5	...	0
0	0.6	...	1	10	0	...	2	4	3	...	-5
0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
0.1	1	...	0	1	1	...	0	1	-2	...	3
1	0.4	...	0.5	4	10	...	4	5	4	...	5

$$x'_i = \frac{9}{10^1} = 0.9$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

108

108

Normalizarea datelor (cont.)

- normalizarea zecimală (cont.);

$$desc = [x_1, x_2, \dots, x_n], x'_i = \frac{x_i}{10^p}, p = \log_{10}(\max |x_i|)$$

date

0.2	0	...	0.9	0	5	...	10	-5	5	...	0
0	0.6	...	1	10	0	...	2	4	3	...	-5
0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
0.1	1	...	0	1	1	...	0	1	-2	...	3
1	0.4	...	0.5	4	10	...	4	5	4	...	5

$$x'_i = \frac{-2}{10^{0.699}} = \frac{-2}{5} = -0.4$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

109

109

Normalizarea datelor (cont.)

- normalizarea zecimală (cont.);

Date

0.2	0	...	0.9	0	5	...	10	-5	5	...	0
0	0.6	...	1	10	0	...	2	4	3	...	-5
0.1	0.2	...	0.1	5	9	...	10	2	-5	...	1
0.1	1	...	0	1	1	...	0	1	-2	...	3
1	0.4	...	0.5	4	10	...	4	5	4	...	5

Date normalizate

0.2	0	...	0.9	0	0.5	...	1	-1	1	...	0
0	0.6	...	1	1	0	...	0.2	0.8	0.6	...	-1
0.1	0.2	...	0.1	0.5	0.9	...	1	0.4	-1	...	0.2
0.1	1	...	0	0.1	0.1	...	0	0.2	-0.4	...	0.6
1	0.4	...	0.5	0.4	1	...	0.4	1	0.8	...	1

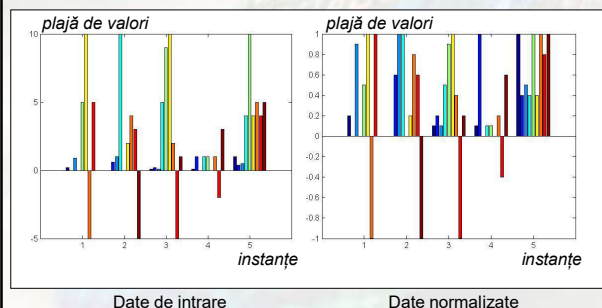
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

110

110

Normalizarea datelor (cont.)

- normalizarea zecimală (cont.);



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

111

111

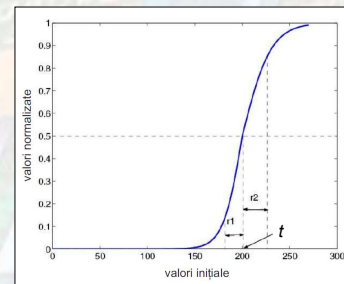
Normalizarea datelor (cont.)

- normalizarea cu funcții sigmoide;

$$desc = [x_1, x_2, \dots, x_n]$$

$$x'_i = \left[1 + e^{-\frac{2(x_i - t)}{r}} \right]^{-1}$$

unde $i=1, \dots, n$, t este de regulă valoarea medie a distribuției valorilor descriptorului iar $r=r_1$ dacă $x_i < t$ și $r=r_2$ altfel.



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

112

112

Decorelarea datelor

> valorile atributele tind să fie corelate între ele ceea ce înseamnă că anumite valori pot fi deduse unele din altele -> **reprezentare ineficientă**;

> decorelarea datelor prin **Principal Component Analysis (PCA)**;

- dispersie ("variance"):

$$\text{var}(X) = \sigma^2 = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - \bar{x})^2$$

unde $x_i, i=0, \dots, n-1$, reprezintă un set de valori ai unui parametru (vector X) iar $\bar{x}$ este valoarea medie a setului.

> măsură a dispersiei valorilor față de valoarea medie;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

113

113

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

- covarianță:

> măsură a variației a două dimensiuni față de medii, una în funcție de cealaltă;

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - \bar{x})(y_i - \bar{y})$$

unde X și Y sunt doi parametri, fiecare cu n valori măsurate (ex. două dimensiuni/atribute ale unui descriptor multi-dimensional).

- valori pozitive: X și Y cresc sau descresc împreună;
- valori negative: fie X crește iar Y scade sau invers;
- valoare 0: X este independent de Y .

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

114

114

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

• *matrice de covarianță (ex. 3 dimensiuni):*

$$S = \begin{bmatrix} \text{cov}(X, X) & \text{cov}(X, Y) & \text{cov}(X, Z) \\ \text{cov}(Y, X) & \text{cov}(Y, Y) & \text{cov}(Y, Z) \\ \text{cov}(Z, X) & \text{cov}(Z, Y) & \text{cov}(Z, Z) \end{bmatrix}$$

unde X , Y și Z sunt trei parametri, fiecare cu n valori măsurate (ex. trei dimensiuni/atribute ale unui descriptor multi-dimensional).

- pe diagonală $\text{cov}(X, X) = \text{var}(X)$, dispersii;
- $\text{cov}(Y, X) = \text{cov}(X, Y)$, matrice simetrică.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

115

115

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

• *vectori și valori proprii:*

$$A \cdot v = \lambda \cdot v$$

unde A este o matrice de dimensiune $n \times n$, v este un vector de dimensiune $n \times 1$ iar λ este un scalar;

- λ reprezintă valoarea proprie a lui A ;
- v reprezintă vectorul propriu al lui A .

> având matricea A , vectorii și valorile proprii se calculează ca soluție a ecuației:

$$\det(A - \lambda \cdot I) = 0 \quad \text{unde } I \text{ este matricea unitate iar } \det(.) \text{ returnează determinantul unei matrice.}$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

116

116

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

• *vectori și valori proprii (cont.): exemplu*

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$

$$\det(A - \lambda \cdot I) = \det\left(\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} - \lambda \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) =$$

$$\det\left(\begin{bmatrix} -\lambda & 1 \\ -2 & -3-\lambda \end{bmatrix}\right) = \lambda^2 + 3\lambda + 2 \quad \begin{cases} \lambda_1 = -1 \\ \lambda_2 = -2 \end{cases}$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

117

117

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

• *vectori și valori proprii (cont.): exemplu*

$$(A - \lambda \cdot I) \cdot v = 0$$

$$(A - \lambda_1 \cdot I) \cdot v_1 = 0$$

$$\begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix} \cdot \begin{bmatrix} v_{1,1} \\ v_{1,2} \end{bmatrix} = 0 \quad \Rightarrow \quad v_1 = k_1 \begin{bmatrix} +1 \\ -1 \end{bmatrix}$$

$$\begin{cases} v_{1,1} + v_{1,2} = 0 \\ -2v_{1,1} - 2v_{1,2} = 0 \end{cases} \quad \text{similar: } v_2 = k_2 \begin{bmatrix} +1 \\ -2 \end{bmatrix}$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

118

118

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

obiectiv: determinarea unei modalități prin care cele n atribute de intrare să fie modificate astfel încât să obținem cea mai reprezentativă descriere a datelor curente;

- reprezentative =

- nu depind de celelalte atribute (decorelate);
- nu sunt perturbate de zgomot;
- prezintă o variație importantă a valorilor.

- cum putem reprezenta datele astfel încât să obținem ceea ce ne dorim? -> transformare a bazei de reprezentare;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

119

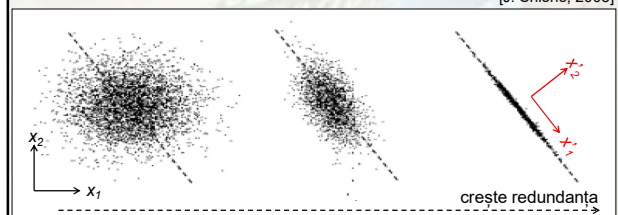
119

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

- de ce este importantă varianța:

[J. Shlens, 2003]



- în care dintre cazuri avem o redundanță maximă a datelor?
- ce transformare a bazei permite maximizare varianță?

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

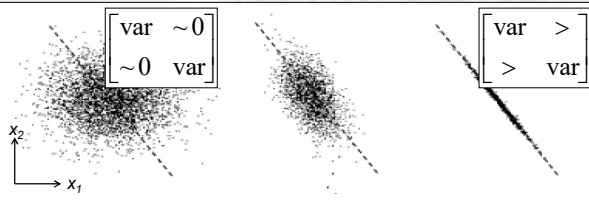
120

120

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

[J. Shlens, 2003]



- cum arată matricea de covarianță pentru cazurile de mai sus?

$$\begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) \end{bmatrix}$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

121

121

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

algoritm: găsirea acelei transformări care permite diagonalizarea matricei de covarianță (= maximizare dispersie valori – valori diagonală; minimizare redundanță – valori de lângă diagonală);

$$X' = P \cdot X$$

unde X sunt datele de intrare, P este matricea transformării care este determinată astfel încât să diagonalizeze matricea de covarianță S a lui X ;

> soluția este dată de P având ca linii vectorii proprii ai matricei de covarianță a lui X ;

> vectorii proprii definesc o nouă bază ortonormată.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

122

122

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

> algoritm (exemplu numeric):

p1: pregătire date de intrare;

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
1	1	2	3	5	8	13	4	17	21	21	42	45	90	91	90
2	3	5	8	9	9	18	27	26	53	52	99	99	99	99	99
0	1	1	2	3	5	8	13	21	34	55	89	90	4	3	4
1	2	3	4	5	6	7	8	9	10	11	11	12	14	14	16

matrice X^T de intrare (T reprezintă transpusa)

*pentru a respecta sensul matematic, în calculele următoare datele sunt reprezentate ca având atributele pe linii și instanțele pe coloane.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

123

123

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

> algoritm (exemplu numeric):

p2: extragere valori medii -> medie 0;

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
0	-0.75	-0.75	-1.25	-0.5	1	1.5	-9	-1.25	-8.5	-13.75	-18.25	-16.5	38.25	39.25	37.75
1	1.25	2.25	3.75	3.5	2	6.5	14	7.75	23.5	17.25	38.75	37.5	47.25	47.25	46.75
-1	-0.75	-1.75	-2.25	-2.5	-2	-3.5	0	2.75	4.5	20.25	28.75	28.5	-47.75	-48.75	-48.25
0	0.25	0.25	-0.25	-0.5	-1	-4.5	-5	-9.25	-19.5	-23.75	-49.25	-49.5	-37.75	-37.75	-36.25

matrice X^T de intrare (centrată)

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

124

124

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

> exemplu numeric (cont.):

p3: calcul matrice de covarianță a lui X ;

p4: calcul valori și vectori proprii matrice de covarianță X ;

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}
8042.97	3668.96	67.64	0	0	0	0	0	0	0	0	0	0	0	0	0

p5: ordonare vectori proprii in ordinea descrescătoare a varianței (= a valorilor proprii λ) -> matrice P ;

p6: transformare X prin P -> $X'=PX$.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

125

125

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

> exemplu numeric (cont.):

p6: transformare X prin P
-> $X'=PX$ (cont.).

$$\text{cov}(X') = \begin{bmatrix} 8.04 & 0 & 0 & \dots & 0 \\ 0 & 3.67 & 0 & \dots & 0 \\ 0 & 0 & 0.07 & \dots & 0 \\ \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
51.28	-51.69	-8.98	0	0	0	0	0	0	0	0	0	0	0	0	0
99.27	28.91	7.34	0	0	0	0	0	0	0	0	0	0	0	0	0
-62.84	71.55	-4.95	0	0	0	0	0	0	0	0	0	0	0	0	0
-87.71	-48.76	6.60	0	0	0	0	0	0	0	0	0	0	0	0	0

matrice $(X')^T$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

126

126

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

> exemplu numeric (cont.):

p7: reducere dimensiuni prin eliminare atribute cu variabilitate scăzută;

$$\text{var}_1 = \frac{\lambda_1}{\sum_i \lambda_i}$$

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
51.28	-51.69	-8.98	0	0	0	0	0	0	0	0	0	0	0	0	0
99.27	28.91	7.34	0	0	0	0	0	0	0	0	0	0	0	0	0
-62.84	71.55	-4.95	0	0	0	0	0	0	0	0	0	0	0	0	0
-87.71	-48.76	6.60	0	0	0	0	0	0	0	0	0	0	0	0	0

→ păstrează 68.27% din variabilitatea datelor inițiale;
Cum am calculat?

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

127

127

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

> exemplu numeric (cont.):

p7: reducere dimensiuni prin eliminare atribute cu variabilitate scăzută (cont.);

$$\text{var}_{1\&2} = \frac{\lambda_1 + \lambda_2}{\sum_i \lambda_i}$$

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
51.28	-51.69	-8.98	0	0	0	0	0	0	0	0	0	0	0	0	0
99.27	28.91	7.34	0	0	0	0	0	0	0	0	0	0	0	0	0
-62.84	71.55	-4.95	0	0	0	0	0	0	0	0	0	0	0	0	0
-87.71	-48.76	6.60	0	0	0	0	0	0	0	0	0	0	0	0	0

→ păstrează 99.43% din variabilitatea datelor;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

128

128

Decorelarea datelor (cont.)

> Principal Component Analysis (cont.);

> exemplu numeric (cont.):

p7: reducere dimensiuni prin eliminare atribute cu variabilitate scăzută (cont.);

$$\text{var}_{1\&2\&3} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{\sum_i \lambda_i}$$

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
51.28	-51.69	-8.98	0	0	0	0	0	0	0	0	0	0	0	0	0
99.27	28.91	7.34	0	0	0	0	0	0	0	0	0	0	0	0	0
-62.84	71.55	-4.95	0	0	0	0	0	0	0	0	0	0	0	0	0
-87.71	-48.76	6.60	0	0	0	0	0	0	0	0	0	0	0	0	0

→ păstrează 100% din variabilitate;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

129

129

Decorelarea datelor (cont.)

> "albirea" datelor (whitening);

> în urma PCA obținem decorelarea atributelor și astfel o matrice de covarianță diagonală;

> valorile dispersiei nu sunt constante și nici egale astfel că datele nu sunt distribuite uniform;

idee: obținerea unei distribuții uniforme a datelor; **whitening** = transformarea datelor astfel încât să aibă o matrice de covarianță diagonală și de dispersii constante, egale cu 1;

$$x_i^w = \frac{1}{\sqrt{\lambda_i + \varepsilon}} \cdot x_i'$$

unde x_i' reprezintă valorile unui atribut în urma PCA, λ_i este valoarea proprie (matrice de covarianță) ce corespunde acestuia iar ε este un parametru de regularizare;

Care este rolul lui ε ? evitare /0.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

130

130

Decorelarea datelor (cont.)

> "albirea" datelor (whitening; cont.);

$$x_i^w = \frac{1}{\sqrt{\lambda_i + \varepsilon}} \cdot x_i'$$

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
51.28	-51.69	-8.98	0	0	0	0	0	0	0	0	0	0	0	0	0
99.27	28.91	7.34	0	0	0	0	0	0	0	0	0	0	0	0	0
-62.84	71.55	-4.95	0	0	0	0	0	0	0	0	0	0	0	0	0
-87.71	-48.76	6.60	0	0	0	0	0	0	0	0	0	0	0	0	0

 λ_1 λ_2 λ_3 matrice $(X')^T$

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
0.57	-0.85	-1.09	0	0	0	0	0	0	0
1.10	0.47	0.89	0	0	0	0	0	0	0
-0.70	1.18	-0.60	0	0	0	0	0	0	0
-0.97	-0.80	0.80	0	0	0	0	0	0	0

$$\text{cov}(X^w) = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

matrice $(X^w)^T$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

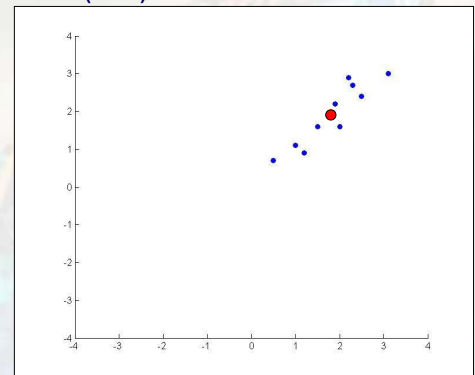
131

131

Decorelarea datelor (cont.)

Date de intrare

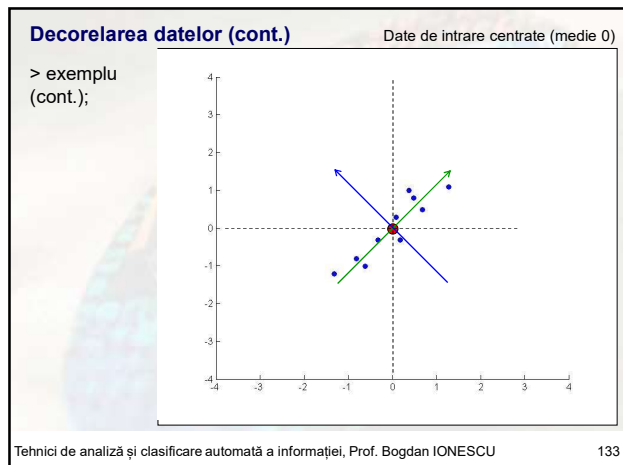
> exemplu;



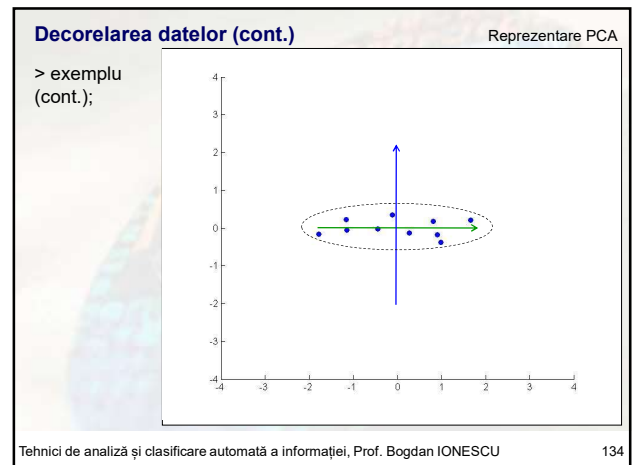
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

132

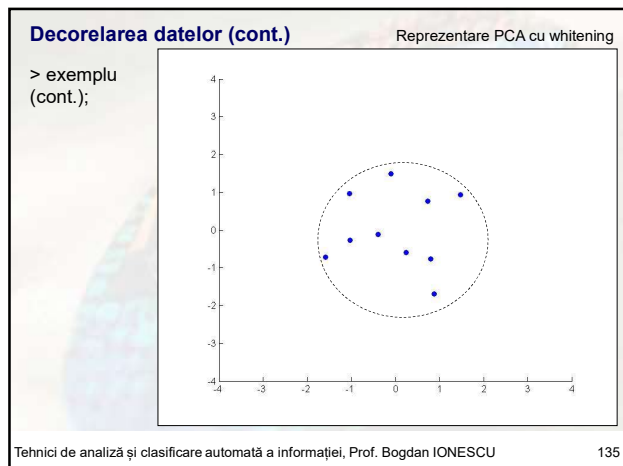
132



133



134



135



136