Multimodal violence detection in Hollywood movies: State-of-the-art and Benchmarking

Claire-Hélène Demarty, Cédric Penet, Bogdan Ionescu, Guillaume Gravier, Mohammad Soleymani

Abstract This paper introduces a benchmark evaluation targeting the detection of violent scenes in Hollywood movies. The evaluation was implemented in 2011 and 2012 as an affect task in the framework of the international MediaEval benchmark initiative. We report on these two years of evaluation, providing a detailed description of the dataset created, describing the state-of-the-art by studying the results achieved by participants and providing a detailed analysis of two of the best performing multimodal systems. We elaborate on the lessons learned after two years to provide insight on future work emphasizing multimodal modeling and fusion.

1 Introduction

Detecting violent scenes in movies appears as an important feature in various use cases related to video on demand and child protection against offensive content. In the framework of the MediaEval benchmark initiative, we have developed a large dataset for this task and assessed various approaches via comparative evaluations.

Claire-Hélène Demarty
Technicolor, R&D France, 975 av. des Champs Blancs 35576 Cesson Sévigné Cedex, e-mail: claire-helene.demarty@technicolor.com

Cédric Penet
Technicolor, R&D France, 975 av. des Champs Blancs 35576 Cesson Sévigné Cedex, e-mail: penetcedric@gmail.com

Bogdan Ionescu
LAPI, University Politehnica of Bucharest, 061071 Romania, e-mail: bionescu@imag.pub.ro

Guillaume Gravier
IRISA & INRIA Rennes, 35042 Rennes Cedex, France, e-mail: guig@irisa.fr

Mohammad Soleymani
iBUG, Imperial College London, SW7 2AZ, UK, e-mail: m.soleymani@imperial.ac.uk

MediaEval[1] is a benchmarking initiative dedicated to evaluating new algorithms for multimedia access and retrieval. MediaEval emphasizes the multimodal character of the data (speech, audio, visual content, tags, users, context, etc). As a track of MediaEval, the Affect Task - Violent Scenes Detection - involves automatic detection of violent segments in movies. The challenge derives from a use case at the company Technicolor[2]. Technicolor is a provider of services in multimedia entertainment and solutions, in particular, in the field of helping users select the most appropriate content according to, for example, their profile. In this context, a particular use case arises which involves helping users choose movies that are suitable for children in their family, by previewing the parts of the movies (i.e., scenes or segments) that include the most violent moments [10].

Such a use case raises several substantial difficulties. Among them, the subjectivity that will occur during the selection of those violent moments is certainly the most important one. Indeed the definition of a violent event remains highly subjective and dependent on the viewers, their culture, their gender. Agreeing on a common definition of a violent event is not easy, which explains why each work related to violence in the literature exhibits a different definition. The semantic nature of the events to retrieve also contributes to the difficulty of the task, as it entails a huge semantic gap between features and interpretation. Due to the targeted content (i.e., Hollywood movies) and the nature of the events, multimodality is also an important characteristic of the task, which stresses its ambitious and challenging nature even more.

The choice of the targeted content raises additional challenges which are not addressed in similar evaluation tasks, for example in the TRECVid Surveillance Event Detection or Multimedia Event Detection Evaluation Tracks[3]. Indeed, systems will have to cope with content of very different genres that may contain special editing effects, which may alter the events to detect.

In the literature, violent scene detection in movies has received very little attention so far. Moreover, comparing existing results is impossible because of the different definitions of violence adopted. As a consequence of the differences in the definition of violence, methods suffer from a lack of standard, consistent and substantial datasets. The Affect task of MediaEval constitutes a first attempt to address all these needs and establish a standard with state-of-the-art performance for future reference.

This paper provides a thorough description of the Violent Scene Detection (VSD) dataset and reviews the state-of-the-art for this task. The main contributions in this regard can be summarized with:

- the proposal of a definition of violence in movies and its validation in the community,

[1] http://www.multimediaeval.org/

[2] http://www.technicolor.com/

[3] http://www.nist.gov/itl/iad/mig/sed.cfm

- the design of a comprehensive dataset of 18 Hollywood movies annotated for violence and for concepts related to violence. Insights about annotation challenges are also provided;
- a detailed description of the state-of-the-art in violence detection;
- a comparison of the systems that competed in the 2011 and 2012 benchmarks and the description of two of the best performing systems.

The paper is organized as follows. Section 2 reviews previous research on violence detection in videos. Section 3 provides an overview of the violent scene detection task after two years of implementation within the MediaEval benchmarking initiative. Section 4 reports the results of the benchmark with a short comparative description of the competing systems. Section 5 provides an in-depth description of two of the best ranked systems with an explicit focus on the contribution of the multimodal information fusion.

2 A review of the literature

Automatically detecting violent scenes in movies received very limited attention prior to the establishment of the MediaEval violence detection task [20].

A closely related problem is action recognition focusing on detecting human violence in real-world scenarios. Datta *et al.* [9] proposed an hierarchical approach for detecting distinct violent events involving two people, e.g., fist fighting, hitting with objects, kicking. They computed the motion trajectory of image structures, i.e., acceleration measure vector and its jerk. Their method was validated on 15 short sequences including around 40 violent scenes. Another example is the approach in [40] which aims at detecting instances of aggressive human behavior in public environments. The authors used a Dynamic Bayesian Network (DBN) as a fusion mechanism to aggregate aggression scene indicators, e.g., “scream”, “passing train” or “articulation energy”. Evaluation is carried out using 13 clips featuring various scenarios, such as “aggression towards a vending machine” or “supporters harassing a passenger”.

Sports videos were also used for violence detection, usually relying on the bag of visual words (BoVW) representation. For instance, [31] addresses fight detection using BoVW along with space-time interest points and motion scale-invariant feature transform (MoSIFT) features. The authors evaluated their method on 1,000 clips containing different actions from ice hockey videos labeled at the frame level. The highest reported detection accuracy is near 90 %. A similar experiment is the one in [38] that used BoVW with local spatio-temporal features, for sports and surveillance videos. Experiments show that motion patterns tend to provide better performance than spatio-visual descriptors.

One of the early approaches targeting broadcast videos is from Nam *et al.* [30] where violent events were detected using multiple audio-visual signatures, e.g., description of motion activity, blood and flame detection, and violence/non-violence classification of the soundtrack and characterization of sound effects. Only qualitative validations were reported. More recently, Gong *et al.* [16] used shot length, motion activity, loudness, speech, light, and music as features for violence detection. A modified semi-supervised learning model was employed for detection and evaluated on 4 Hollywood movies, achieving a F-measure of 0.85 at best. Similarly, Giannakopoulos *et al.* [13] used various audio-visual features for violence detection in movies, e.g., spectrogram, chroma, energy entropy, Mel-Frequency Cepstral Coefficients (MFCC), average motion, motion orientation variance, measure of the motion of people or faces in the scene. Modalities were combined by a meta-classification architecture that classified mid-term video segments as "violent" or "non-violent". Experimental validation was performed on 50 video segments ripped from 10 different movies (totaling 150 minutes) with F-measures up to 0.58. Lin and Wang [26] proposed a violent shot detector that used a modified probabilistic Latent Semantic Analysis (pLSA). Audio features as well as visual concepts such as motion, flame, explosion and blood were employed. Final integration was achieved though a co-training scheme, typically used when dealing with small amounts of training data and large amounts of unlabeled data. Experimental validation was conducted on 5 movies showing an average F-measure of 0.88.

Most of the approaches are naturally multimodal, exploiting both the image and sound tracks. However, a few works approached the problem based on a single modality. For example, [7] used Gaussian mixture models (GMM) and hidden Markov models (HMM) to model audio events over time series. They considered the presence of gunplay and car racing with audio events such as "gunshot", "explosion", "engine", "helicopter flying", "car braking", and "cheers". Validation was performed on a very restrained data set, containing excerpts of 5 minutes extracted from 5 movies, leading to an average F-measure of up to 0.90. In contrast, [5] used only visual concepts such as face, blood, and motion information to determine whether an action scene had violent content or not. The specificity of their approach is in addressing more semantics-bearing scene structures of video rather than simple shots.

In general, most of the existing approaches focus more or less on finding the correct concepts that can be translated into violence in general and their findings are bounded by the size of the dataset and the definition of violence. Because of the high variability of violent events in movies, no common and objective enough definition for violent events was ever proposed to the community, even when restricting to physical violence. On the contrary, each piece of work dealing with the detection of violent scenes provides its own definition of the violent events to detect. For instance, [5] targeted "a series of human actions accompanied with bleeding", [38, 31] looked for "scenes containing fights, regardless of context and number of people involved". In [13], the following definition is used: "behavior by persons against persons that intentionally threatens, attempts, or actually inflicts physical harm". In [16], authors were interested in "fast paced scenes which contain explosions, gunshots and person-on-person fighting". Moreover, violent scenes and action scenes are often mixed up in the past as in [6, 16].

The lack of a common definition and the resulting absence of a reference and substantial dataset has made it so far very difficult to compare methods which were sometimes developed for a very specific type of violence. This is precisely the fault that we attempt to correct with the MediaEval violent scene detection task, by creating a benchmark based on a clear and generalizable definition of violence to advance the state-of-the-art on this topic.

3 Affect task description

The 2011 and 2012 Affect Task required participants to deploy multimodal approaches to automatically detect portions of movies depicting violence. Though not a strict requirement, we tried to emphasize multimodality for several reasons. First, videos are multimodal. Second, violence might be present in all modalities though not necessarily at the same time. This is clearly the case for images and soundtracks. Violence might also be reflected in subtitles though verbal violence was not considered. In spite of a definition of violence limited to physical violence, single modality approaches were bound to be suboptimal and most participants ended up using visual and audio features.

The key for creating a corpus for comparative evaluation clearly remains a general definition of the notion of violence which eases annotation while encompassing a large variety of situations. We discuss here the notion of violence and justify the definition that was adopted before describing the data set and evaluation rules.

3.1 Towards a definition of violence

The notion of violence remains highly subjective as it depends on viewers. The World Health Organization (WHO) [1] defines violence as: "*The intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation*". According to the WHO, three types of violence can be distinguished, namely, self-inflicted, interpersonal, and collective [23]. Each category is divided according to characteristics related to the setting and nature of violence, e.g., physical, sexual, psychological, and deprivation or neglect.

In the context of movies and television, Kriegel [22] defines violence on TV as an "*unregulated force that affects the physical or psychological integrity to challenge the humanity of an individual with the purpose of domination or destruction*".

These definitions only focus on intentional actions and, as such, do not include accidents, which are of interest in the use case considered, as they also result in potentially shocking gory and graphic scenes, e.g., a bloody crash. We therefore adopted an extended definition of violence that includes accidents while being as objective as possible and reducing the complexity of the annotation task. In MediaEval 2011 and 2012, violence is defined as "*physical violence or accident resulting* *in human injury or pain*". Violent events are therefore limited to physical violence, verbal or psychological violence being intentionally excluded.

Even though we attempted to narrow the field of violent events down to a set of events as objectively violent as possible, there are still some borderline cases. First of all, sticking to this definition leads to the rejection of some shots in which the results of some physical violence are shown but not the violent act itself. For example, shots in which one can see a dead body with a lot of injuries and blood were not annotated as violent. On the contrary, a character simply slapping another one in the face is considered as a violent action according to the task definition. Other events defined as "intent to kill", in which one sees somebody shooting somebody else for example with the clear intent to kill, but the targeted person escapes with no injury, were also discussed and finally not kept in the violent set. On the contrary, scenes where the shooter is not visible but where shooting at someone is obvious from the audio, e.g., one can hear the gunshot possibly with screams afterward, were annotated as violent. Interestingly, such scenes emphasize the multimodal characteristic of the task. Shots showing actions resulting in pain but with no intent to be violent or, on the contrary, with the aim of helping rather than harming, e.g., segments showing surgery without anesthetics, fit into the definition and were therefore deemed violent.

Another borderline case keenly discussed were the events such as shots showing the destruction of a whole city or the explosion of a moving tank. Technically speaking, these shots do not show any proof of people death or injury, though one can reasonably assume that the city or the tank were not empty at the time of destruction. Consequently, such cases, where pain or injury is implicit, were annotated as violent. Finally, shots showing the violent action and the result of the action itself happen to be separated by several non violent shots. In this case, the entire segment was annotated as violent if the duration between the two violent shots (action and result) was short enough (less than two seconds).

3.2 Data description

In line with the use case considered, the dataset consisted of Hollywood movies from a comprehensive range of genres, from extremely violent to movies without violence. In 2011, 15 movies were considered and completed by 3 additional movies in 2012. From these 18 movies, 12 were designated as development data[4] in 2011. The three movies used as test set[5] in 2011 where shifted to the development set in 2012 where three additional movies were provided for evaluation. The list of movies, along with some characteristics, is given in Table 1.

The development dataset represents a total of 26,108 shots in 2012—as given by automatic shot segmentation—for a total duration of 102,851 seconds. Violent

[4] the development data is intended for designing and training the approaches.

[5] the test set data is intended for the official benckmarking.

Table 1: Movie dataset (2011 dev. set: first 12 movies; 2011 test set: following 3 movies. 2012 dev. set: first 15 movies; 2012 test set: last three movies). Dur: duration in seconds; Sh: number of shots; V-Dur: violent shot duration proportion (%); V-Sh: Violent shot proportion (%).

2012	**2011**	**Movie**	**Dur**	**Sh**	**V-Dur**	**V-Sh**
Dev. set	Dev. set	Armageddon	8680.16	3562	10.16	11.0
		Billy Elliot	6349.44	1236	5.14	4.21
		Eragon	5985.44	1663	11.02	16.6
		Harry Potter 5	7953.52	1891	9.73	12.69
		I am Legend	5779.92	1547	12.45	19.78
		Leon	6344.56	1547	4.3	7.24
		Midnight Express	6961.04	1677	7.28	11.15
		Pirates Carib. 1	8239.4	2534	11.3	12.47
		Reservoir Dogs	5712.96	856	11.55	12.38
		Saving Private Ryan	9751.0	2494	12.92	18.81
		The Sixth Sense	6178.04	963	1.34	2.80
		The Wicker Man	5870.44	1638	8.36	6.72
		Total	**83805.9**	**21608**	**9.02**	**14.8**
	Test set	Kill Bill	6370.4	1597	17.47	23.98
		The Bourne Identity	6816.0	1995	7.61	9.22
		The Wizard of Oz	5859.2	908	5.51	5.06
		Total	**19045.6**	**4500**	**11.55**	**13.62**
	Total		**102851.5**	**26108**	**9.25**	**12.27**
Test set		Dead Poets Society	7413.2	1583	1.5	2.14
		Fight Club	8005.7	2335	13.51	13.27
		Independance Day	8834.3	2652	9.92	13.98
	Total		**24253.2**	**6570**	**8.53**	**10.88**

content corresponds to 9.25% of the total duration and 12.27% of the shots, highlighting the fact that violent segments are not so scarce in this database. We tried to respect the genre distribution (from extremely violent to non violent) both in the development and test sets. This appears in the statistics, as some movies such as *Billy Elliot* or *The Wizard of Oz* contain a small proportion of violent shots (around 5%). The choice we made for the definition of violence impacts the proportion of annotated violence in some movies such as *The Sixth Sense* where violent shots amount to only 2.8% of the duration. However, the movie contains several shocking scenes of dead people which do not fit the definition of violence that we adopted. In a similar manner, psychological violence, such as what may be found in *Billy Elliot*, was also not annotated, which also explains the small number of violent shots in this particular movie.

The violent scenes dataset was created by seven human assessors. In addition to segments containing physical violence according to the definition adopted, annotations also include high-level concepts potentially related to violence for the visual and audio modalities, highlighting the multimodal character of the task.

The annotation of violent segments was conducted using a 3 step process, with the same so-called "master annotators" for all movies. A first master annotator extracted all violent segments. A second master annotator reviewed the annotated segments and possibly missed segments according to his/her own judgment. Disagreements were discussed on a case by case basis, the third master annotator making the final decision in case of an unresolved disagreement. Each annotated violent segment contained a single action, whenever possible. In the case of overlapping actions, the corresponding global segment was proposed as a whole. This was indicated in the annotation files by adding the tag "multiple action scene". The boundaries of each violent segment were defined at the frame level, i.e., indicating the start and end frame numbers.

The high-level video concepts were annotated through a simpler process, involving only two annotators. Each movie was first processed by an annotator and then reviewed by one of the master annotators.

Seven visual concepts are provided: *presence of blood, fights, presence of fire, presence of guns, presence of cold weapons, car chases and gory scenes*. For the benchmark, participants had the option to carry out detection of the high-level concepts. However, concept detection is not among the task's goals and these high-level concept annotations were only provided on the development set. Each of these high-level concepts followed the same annotation format as for violent segments, i.e., starting and ending frame numbers and possibly some additional tags which provide further details. For blood annotations, a tag in each segment specifies the proportion of the screen covered in blood. Four tags were considered for fights: only two people fighting, a small group of people (roughly less than 10), large group of people (more than 10), distant attack (i.e., no real fight but somebody is shot or attacked at distance). As for the presence of fire, anything from big fires and explosions to fire coming out of a gun while shooting, a candle, a cigarette lighter, a cigarette, or sparks was annotated, e.g., a space shuttle taking off also generates fire and receives a fire label. An additional tag may indicate special colors of the fire (i.e., not yellow or orange). If a segment of video showed the presence of firearms (respectively cold weapons) it was annotated by any type of (parts of) guns (respectively cold weapons) or assimilated arms. Annotations of gory scenes are more difficult. In the present task, they are indicating graphic images of bloodletting and/or tissue damage. It includes horror or war representations. As this is also a subjective and difficult notion to define, some additional segments showing disgusting mutants or creatures are annotated as gore. In this case, additional tags describing the event/scene are added.

For the audio modality, three audio concepts were annotated, namely, *gunshots, explosions, screams*. Those concepts were extracted using the English audio tracks. Contrary to what is done for the video concepts, audio segments are identified by start and end times in seconds. Additional tags may be added to each segment to distinguish different types of sub-concepts. For instance, distinction was made between gunshots and cannon fires. All kinds of explosions were annotated, even magic explosions as well as explosions resulting from shells or cannonballs in cannon fires. Last, scream annotations are also provided, however for 9 movies only, in which anything from non verbal screams to what was called "effort noise" was extracted, as long as the noise came from a human or a humanoid. Effort noises were separated from the rest, by the use of two different tags in the annotation.

In addition to the annotation data, automatically generated shot boundaries with their corresponding key frames, as detected by Technicolor's software, were also provided with each movie.

3.3 Evaluation rules

Due to copyright issues, the video content was not distributed and participants were required to buy the DVDs. Participants were allowed to use all information automatically extracted from the DVDs, including visual and auditory material as well as subtitles. English was the chosen language for both the audio and subtitles channels. The use of any other data, not included in the DVD (web sites, synopsis, etc.) was not allowed.

Two types of runs were initially considered in the task, a mandatory shot classification run and an optional segment detection one. The shot classification run consisted in classifying each shot provided by Technicolor's shot segmentation software as violent or not. Decisions were to be accompanied by a confidence score where the higher the score, the more likely the violence. Confidence scores were optional in 2011 and compulsory in 2012 because of the chosen metric. The segment detection run involved detection of the violent segment boundaries, regardless of the shot segmentation provided.

System comparison was based on different metrics in 2011 and 2012. In 2011, performance was measured using a detection cost function weighting false alarms (FA) and missed detections (MI), according to

$$C = C_{fa} \cdot P_{fa} + C_{miss} \cdot P_{miss} \tag{1}$$

where the costs $C_{fa} = 1$ and $C_{miss} = 10$ were arbitrarily defined to reflect (a) the prior probability of the situation and (b) the cost of making an error. P_{fa} and P_{miss} are the estimated probabilities of respectively false alarms (false positive) and missed detections (false negative) given the system's output and the reference annotation. In the shot classification, the FA and MI probabilities were calculated on a per shot basis while in the segment level run, they were computed on a per unit of time basis, i.e., durations of both references and detected segments are compared. This cost function is called "MediaEval cost" in all that follows.

Experience taught us that the MediaEval detection cost was too strongly biased towards low missed detection rates, leading to systems hardly reaching cost values lower than 1 and therefore worse than a naive system classifying all shots as violent. We therefore adopted the Mean Average Precision (MAP) computed over the first 100 top-ranked violent segments as evaluation metric. Note that this measure is also well adapted to the search-related use case that serves as a basis for our work.

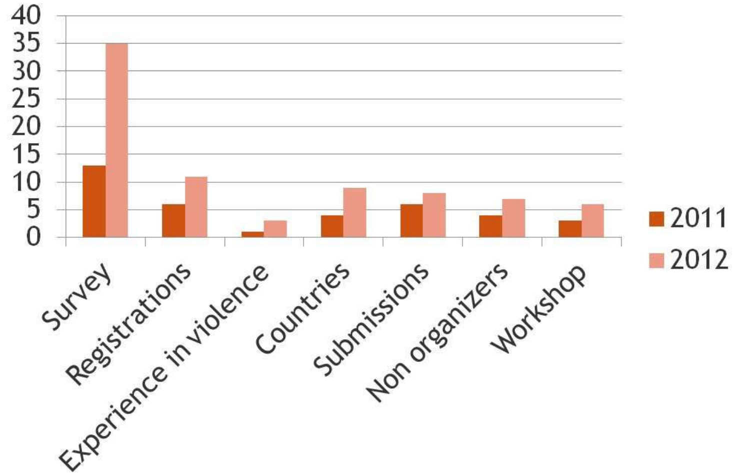


Fig. 1: Evolution of the participation to the task between 2011 and 2012.

We also report detection error tradeoff curves, showing P_{fa} as a function of P_{miss} given a segmentation and the confidence score for each segment, to compare potential performance at different operating points. Note that in the segment detection run, DET curves are possible only for systems returning a dense segmentation (a list of segments that spans the entire video): segments not present in the output list are considered as non violent for all thresholds.

4 Results

In 2011, the Affect Task on Violent Scenes Detection was proposed in MediaEval as a pilot for the first year. Thirteen teams, corresponding to 16 research groups considering joint submission proposals, declared interest in the task. Finally, six teams registered and completed the task, representing four different countries, for a grand total of 29 runs submitted. These figures show the interest for the task for this first year. This was confirmed in 2012, with the registration of 11 teams, of which 8 crossed the final line, by sending 36 runs for the evaluation. Interest is also emphasized by the wide geographic coverage area of teams. Interestingly, the multimodal aspect of the task shows in the fact that participants come from different communities, namely the audio and image processing communities. A more detailed evolution of the task for these two years is summarized in Figure 1.

Official results are reported in Table 2. Despite the change of official metric between 2011 and 2012, MAP values were also computed on the 2011 submissions. Similarly, the MediaEval cost is reported for 2012. It should nevertheless be noted that these two metrics imply different tunings of the systems (towards low precision rate for the MediaEval cost, and on the contrary towards high precision for the

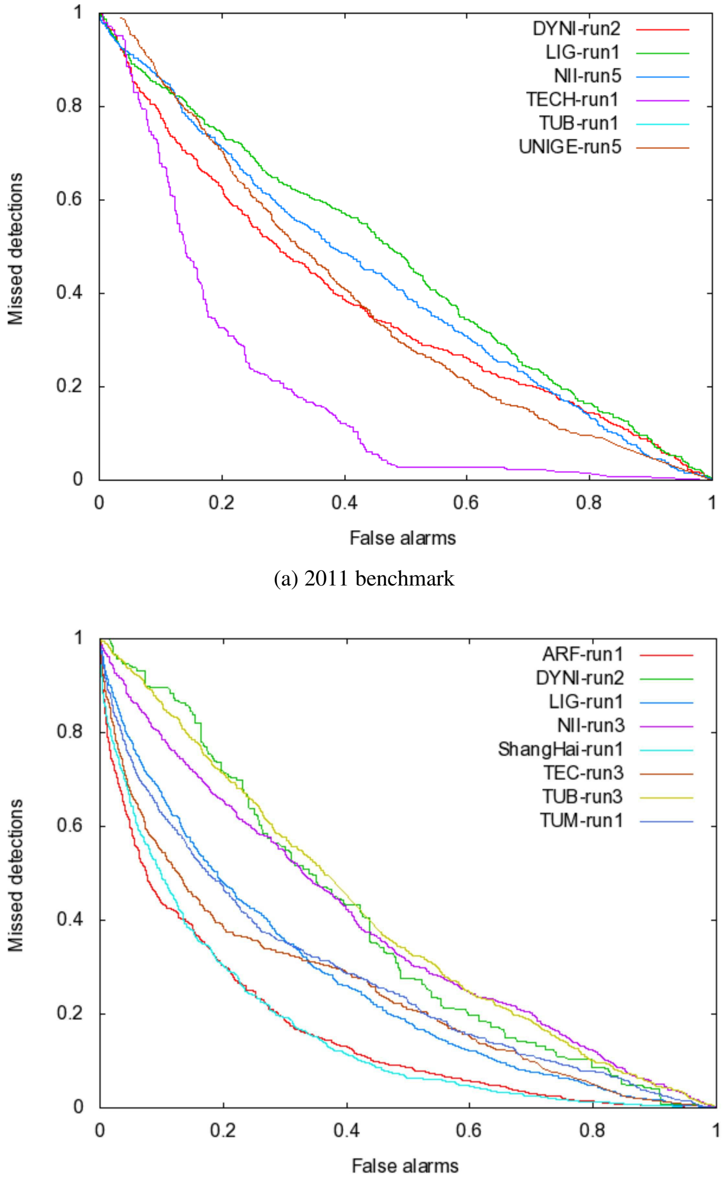
MAP), meaning that metric values should be compared cautiously, as systems were not optimized in the same way.

Table 2: Official results of the 2011 and 2012 Affect task evaluation at MediaEval. In 2011, we report in plain figure results from the best run according to the MediaEval cost and indicate in parenthesis results corresponding to the best run according to the mean average precision. Team names indicated with ”*” correspond to the task organizers.

team	country	MAP@20	MAP@100	Med. cost
			2011 benchmark	
ARF	Austria-Romania-France	—	—	—
DYNI	France	13.81 (*31.22*)	18.33 (*19.07*)	6.46 (*7.57*)
LIG	France	23.87 (*23.87*)	18.01 (*18.01*)	7.93 (*7.93*)
NII	Japan	40.73 (*33.14*)	24.78 (*27.71*)	1 (*1*)
Shanghai-Hongkong	China	—	—	—
TEC*	France-UK	33.33 (*44.94*)	21.89 (*40.58*)	0.76 (*0.89*)
TUB	Germany	4.69 (*4.69*)	14.29 (*14.29*)	1.26 (*1.26*)
TUM	Germany-Austria	—	—	—
UNIGE*	Switzerland	29.28 (*29.28*)	24.57 (*24.57*)	2.00 (*2.83*)
			2012 benchmark	
ARF	Austria-Romania-France	70.08	65.05	3.56
DYNI	France	0	12.44	7.96
LIG	France	28.64	31.37	4.16
NII	Japan	40.07	30.82	1.28
Shanghai-Hongkong	China	73.6	62.38	5.52
TEC*	France-UK	66.89	61.82	3.56
TUB	Germany	35.92	18.53	4.2
TUM	Germany-Austria	50.42	48.43	7.83
UNIGE*	Switzerland	—	—	—

In 2011 and 2012, all participants submitted predominantly runs for the shot classification task. Only the ARF team submitted one segment-level run in 2012. Results show a substantial improvement between 2011 and 2012. Although the overall performances of the proposed systems in 2011 were not good enough to satisfy the requirements of a real-life commercial system, in 2012 three systems reached MAP@100 values above 60%, leading to the conclusion that research still needs to be conducted on this subject, nevertheless state-of-the-art systems already show convincing performances.

Detection error trade-off curves, obtained from the confidence values provided by participants, are given in Figures 2a and 2b for the best run of each participant according to the official metric for the year considered. Clearly, ordering of the systems differs according to the operating point. Once again the direct comparison of the 2011 and 2012 curves is to be considered with caution. Nevertheless, improvements can be observed between the two years. Whereas in 2011, only one participant reached at best a false alarm rate of 20 % for a missed detection rate of



(a) 2011 benchmark

(b) 2012 benchmark

Fig. 2: Detection error trade-off curves for all participants in 2011 and 2012.

about 25 %, in 2012, at least two participants have similar results and three more additional teams have fair results.

Analyzing the 2011 submissions, three different systems categories can be distinguished. Two participants (NII [25] and LIG [36]) treated the problem of violent scene detection as a concept detection problem, applying generic systems developed for TRECVid evaluations to violent scene detection, potentially with specific tuning. Both sites used classic video only features, computed on the key frames provided, based on color, textures, edges, either local (interest points) or global, and classic classifiers. One participant (DYNI [14]) proposed a classifier-free technique exploiting only two low-level audio and video features, computed on each successive frame, both measuring the activity within a shot. After a late fusion process, decisions were taken by comparison with a threshold. The last group of participants (TUB [3] , UGE [15] and TI [32]) built dedicated supervised classification systems for the task of violent scene detection. Different classifiers were used from SVM, Bayesian networks to linear or quadratic discriminant analysis. All used multimodal features, either audio-video or audio-video-textual features (UGE). Features were computed globally for each shot (UGE, TI) or on the provided key frames (TUB).

In 2012, systems were all supervised classification systems; LIG [11] and NII [24] went on with some improved versions of their generic systems dedicated to concept detection, while others implemented dedicated versions of such systems for the task of violent scene detection. Chosen classifiers were mostly SVM, with some exceptions for neural networks and Bayesian networks. It should be noted that most participants [11, 37, 34, 2, 21, 12] voted for multimodal (audio+video) systems and that multimodality seems to help the performance of such systems. Globally classic low-level audio (MFCC, zero-crossing rate, asymetry, roll-off, etc) and video (color histograms, texture-related, Scale Invariant Feature Transform-like, Histograms of Oriented Gradients, visual activity, etc) features were extracted. One exception may be noted with the use of multi-scale local binary pattern histogram features by DYNI [29]. Added to those classical features, audio and video mid-concept detection was also used for this second year [24, 37, 11, 21], thanks to the annotated high-level concepts. Such mid-level concepts, especially used in a two-step classification scheme [37], seem to be promising.

Based on these results, one may draw some tentative conclusions about the global characteristics that were more likely to be useful for violence detection. Local video features (SIFT-like) did not add a lot of information to the systems. On the contrary, taking advantage of different modalities seems to improve performance, especially when modalities are merged using late fusion. Although results do not prove their impact in one way or another, it also seems of interest to use temporal integration. This was carried out in different manners in the systems, either by using contextual features, i.e., features at different times, or by temporal smoothing or aggregation of the decisions at the output of the chain. Using intermediate concept detection with high-level concepts related to violence such as those provided in the task seems to be rewarding.

5 Multimodal approaches

Progress achieved between 2011 and 2012 can probably be explained by two main factors. Data availability is undoubtedly the first one, along with experience on the task. Exploiting multimodal features is also one of the keys. While many systems made very limited use of multiple modalities in 2011, multimodal integration became more widely spread, mostly relying on the audio and visual modalities.

We provide here details for two multimodal systems which competed in 2012, namely the ARF system based on mid-level concepts detected from multimodal input and the Technicolor/IRISA system which directly exploits a set of low-level audio and visual features.

5.1 A mid-level concept fusion approach

We describe the approach developed by the ARF team [37, 20], relying on fusing mid-level concept predictions inferred from low-level features by employing a bank of multi-layer perceptron classifiers featuring a dropout training scheme.

The motivation of this approach lies in the high variability in appearance of violent scenes in movies and the low amount of training data that is usually available. In this scenario, training a classifier to predict violent frames directly from visual and auditory features seems rather difficult. The system proposed by ARF team uses the task provided a high-level concept ground-truth to infer mid-level concepts as an intermediate step towards the final violence detection goal, thus attempting to limit the semantic gap. Experiments proved that predicting mid-level concepts from low-level features should be more feasible than directly predicting all forms of violence.

5.1.1 Description of the system

Violence detection is first carried out at frame level by classifying each frame as being violent or non violent. Segment level prediction (shot level or arbitrary length) is then determined by a simple aggregation of frame level decisions. Given the complexity of this task, i.e., labeling of individual frames rather than video segments (ca. 160,000 frames per movie), the classification is tackled by exploiting the inherent parallel architecture of neural networks. The system involves several processing steps as illustrated in Figure 3.

Multimodal features. Firstly, raw video data is converted into content descriptors whose objective is to capture meaningful properties of the auditory-visual information. Feature extraction is carried out at the frame level. Given the specificity of the task, the system was tested using audio, color, feature description and temporal structure information which is specific both for violence-related concepts as well

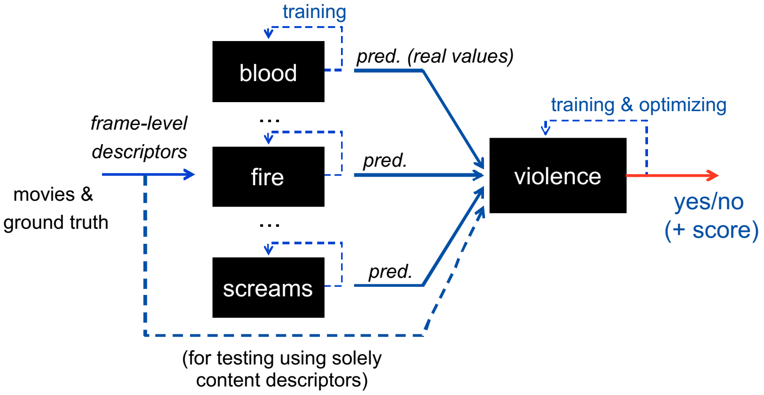


Fig. 3: Description of ARF teams's system developed for MediaEval 2012 (black boxes refer to classifiers).

as for the violent content itself. Results reported in 2012 were obtained with the following descriptors:

- audio descriptors (196 dimensions) consist of general purpose descriptors: linear prediction coefficients, line spectrum pairs, MFCCs, zero-crossing rate, and spectral centroid, flux, rolloff, and kurtosis, augmented with the variance of each feature over a window of 0.8 s around the current frame[6];
- color descriptors (11 dimensions) using the color naming histogram proposed in [39] which maps colors to 11 universal color names (“black”, “blue”,“brown”, “grey”, “green”, “orange”, “pink”, “purple”, “red”, “white”, and “yellow”);
- visual features (81 dimensions) which consist of the 81-dimensional Histogram of Oriented Gradients [28];
- temporal structure (1 dimension) derives a measure of visual activity. The cut detector in [19] that measures visual discontinuity by means of a difference between color histograms of consecutive frames, was modified to account for a broader range of significant visual changes. For each frame it determines the number of detections in a certain time window centered at the current frame. High values of this measure will account for important visual changes that are typically related to action.

Neural network classification. Both at the concept level and at the violence level, classification is carried out with a neural network, namely a multi-layer perceptron with a single hidden layer of 512 logistic sigmoid units. Network is trained by gradient descent on the cross-entropy error with backpropagation [35], using the recent idea in [18] to improve generalization: For each presented training case, a fraction of input and hidden units is omitted from the network and the remaining weights are scaled up to compensate. The set of dropped units is chosen at random for each

[6] the Yaafe toolkit for audio feature extraction was used.

presentation of a training case, such that many different combinations of units will be trained during an epoch.

Concept detection consists of a bank of perceptrons that are trained to respond to each of the targeted violence-related concepts, such as presence of “fire”, presence of “gunshots”, or “gory” scenes (see Section 3.2). As a result, a concept prediction value in $[0,1]$ is obtained for each concept. These values are used as inputs to a second classifier, acting as a final fusion scheme to provide values for the two classes “violence” and “non violence” on a frame-by-frame basis. For all classifiers, parameters were trained using reference annotations coming along with the data.

Violence classification. Frame prediction of violence for the unlabeled data is given by the system’s output when fed with the new data descriptors. As prediction is provided at frame level, aggregation into segments is performed by assigning a violence score corresponding to the highest predictor output for any frame within the segment. The segments are then tagged as “violent” or “non-violent” depending on whether their violence score exceeds a certain threshold (determined in the training step of the violence classifier).

5.1.2 Results

Results are evaluated on the shot classification task and on the segment detection one.

Shot level classification. To highlight the contributions of the concept fusion scheme, different feature combinations were tested, namely: ARF-(c) uses as features only mid-level concept predictions for violence detection; ARF-(a) uses only audio descriptors, i.e., the violence classifier is trained directly on features instead of using the concept prediction outputs; ARF-(v) uses only visual features; ARF-(av) uses only audio-visual features; finally, ARF-(avc) uses all concepts and audio-visual features using an early fusion aggregation of concept predictions and features.

Results on the 2012 benchmark, reported in Table 3, exhibited a F-measure of 49.9 which placed the system among the top systems. The lowest discriminative power is achieved using only visual descriptors (ARF-(v)), with an F-measure of 35.6. Compared to visual features, audio features seem to show better descriptive power, providing an F-measure of 46.3. The combination of descriptors (early fusion) tends to reduce their efficiency and yields lower performance than the use of concepts alone, e.g., audio-visual (ARF-(av)) yields an F-measure of 44.6, while audio-visual-concepts (ARF-(avc)) achieve 42.4.

Figure 4 details the precision-recall curves for this system. The use of concepts fusion scheme (red line) proved again to provide significantly higher recall than the sole use of audio-visual features or the combination of all for a precision of 25 % and above.

Arbitrary segment-level results. At the segment detection level, the use of the fusion of the mid-level concepts achieves average precision and recall values of 42.21 % and 40.38 %, respectively, while the F-measure is 41.3. This yields a miss

Table 3: ARF team violence shot-level detection results at MediaEval 2012.

run	modality	precision	recall	F_1-score
ARF-(c)	concepts	46.14%	54.40%	49.94%
ARF-(a)	audio	46.97%	45.59%	46.27%
ARF-(av)	audio-visual	32.81%	67.69%	44.58%
ARF-(avc)	audio-visual	31.24%	66.15%	42.44%
ARF-(v)	visual	25.04%	61.95%	35.67%

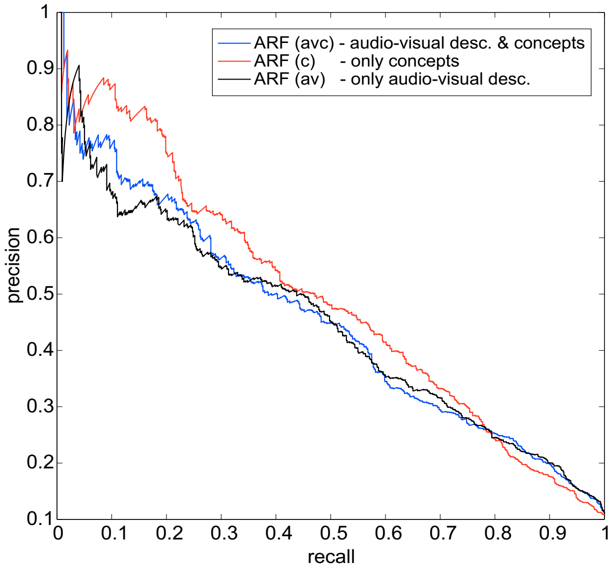


Fig. 4: ARF system precision-recall curves [20].

rate (at time level) of 50.69 % and a very low false alarm rate of only 6 %. These results are promising considering the difficulty of precisely detecting the exact time interval of violent scenes, but also the subjectivity of the human assessment (reflected in the ground truth).

5.2 Direct modeling of multimodal features

We describe here the approach adopted in the joint submission of Technicolor and IRISA in 2012, which directly models a set of multimodal features to infer violence

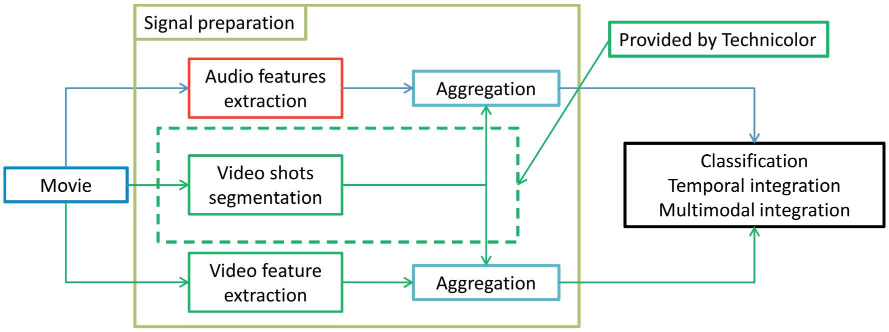


Fig. 5: Description of the Technicolor/IRISA system at MediaEval 2012.

at the shot level. Relying on Bayesian networks and, more specifically, on structure learning in Bayesian networks [17], we investigate multimodal integration via early and late fusion strategies, along with temporal integration.

5.2.1 Description of the system

Figure 5 provides a schematic overview of the various steps implemented in Technicolor's system. Violence detection is performed at the shot level via direct modling of audio and visual features aggregated over shots. Classification is then performed either based on the entire set of multimodal features or independently for each modality. In this last case, late fusion is used to combine modalities. In both cases, temporal information can be used at two distinct levels: in the model with contextual features or as a post-processing step to smooth decisions taken on a per shot basis.

Multimodal features. For each shot, different low level features are extracted from both the audio and the video signals of the movies:

- audio features: the audio features, extracted using 40ms frames with 20ms overlap, are: the energy (E), the frequency centroid (C), the asymmetry (A), the flatness (F), the 90 % frequency roll-off (R) and the zero-crossing rate (Z) of the signal. These features are normalized to zero mean and unit variance, and averaged over the duration of a shot, in order to obtain a single value per shot for each feature. The audio feature vector dimension is $D = 6$;
- video features: the video features extracted per shot are: the shot length (SL), the mean proportion of blood color pixels (B), the mean activity (AC), the number of flashes (FL), the mean proportion of fire color pixels (FI), a measurement of color coherence (CC), the average luminance (AVL), and three color harmony features, the majority harmony template (Tp), the majority harmony template mean angle (Al) and the majority harmony template mean energy (Em) [4]. The feature vector dimension is $D = 10$.

Features are quantized in 21 bins on a per movie basis, except for the majority template whose values are already quantized over 9 bins.

Bayesian network classification. Bayesian networks are used as a classification technique. The idea behind Bayesian networks is to build a probabilistic network on top of the input features with a node in the network for classification of violence. The network represents conditional dependencies and independencies between the features, and it is possible to learn the structure of the graph using structure learning algorithms. The output of the classifier is, for each shot, the estimated posterior probabilities for each class, viz., violence and non-violence.

We compared a so-called naive structure, which basically links all the features to the class variable, with structures learned using either forest-augmented networks (FAN) [27] or K2 [8]. The FAN structure consists in building a tree on top of the naive structure based on some criterion related to classification accuracy. On the contrary, the K2 algorithm does not impose the naive structure but rather attempts a better description of the data based on a Bayesian information criterion, thus not necessarily targeting better classification.

Temporal integration. Two strategies for integrating temporal information were tested. The first one is a contextual representation of the shots at the input of the classifier, where classification of a shot relies on the features for this shot augmented with the features from the neighboring shots. If we denote F_i the features for shot i, the contextual representation of shot i is given by:

$$F_i^{\star} := \{F_{i-n}, F_{i-n+1}, \dots, F_{i-1}, F_i, F_{i+1}, \dots, F_{i+n-1}, F_{i+n}\} \tag{2}$$

where the context size was set to $n = 5$ (empirically determined).

In addition to contextual representation, we also used temporal filtering to smooth the shot by shot independent classification, considering two types of filters:

- a majority vote over a sample window of size $k = 5$, after thresholding the probabilities.
- an average of the probabilities over a sliding window of size $k = 5$, before thresholding the probabilities.

Contrary to averaging, majority vote does not directly provide a confidence score in the decision taken. We implemented the following heuristics in this case. For a given shot, if the vote results in violence, the confidence score is set to $\min\{P(S_v)\}$, where $P(S_v)$ is the set of probabilities of the shots that were considered as violent within the window. If the vote results in a non violent decision, the confidence score is set to $\max\{P(S_{nv})\}$, where $P(S_{nv})$ is the set of probabilities of the shots that were considered as non violent within the window.

Multimodal integration. As for multimodal integration, early fusion and late fusion are compared. Early fusion consists in the concatenation of the audio and the video attributes in a common feature vector. The violence classifier is then learned using this feature vector. Late fusion consists in fusing the outputs of both a video classifier and an audio classifier. In order to fuse the output of the i^{th} shot, the following rule is used:

$$P^{s_i}_{fused}(P^{s_i}_{v_a}, P^{s_i}_{v_v}) = \begin{cases} max\{P^{s_i}_{v_a}, P^{s_i}_{v_v}\} & \text{if both decisions are violent} \\ min\{P^{s_i}_{v_a}, P^{s_i}_{v_v}\} & \text{if both decisions are non violent} \\ P^{s_i}_{v_a} \cdot P^{s_i}_{v_v} & \text{otherwise} \end{cases} \tag{3}$$

where $P^{s_i}_{v_a}$ (resp. $P^{s_i}_{v_v}$) is the probability that shot i is violent as given by the audio (respectively video) classifier. This simple rule of thumb yields a high score when both classifiers agree on violence, and a low score when they agree on non violent.

5.2.2 Results

We first compare the different strategies implemented using cross validation over the 15 development movies, leaving one movie out for test on each fold. We then report results for the best configuration on the official 2012 evaluation.

The MAP@100 values obtained in cross-validation for the audio only, the video only and the early fusion experiments are presented in Table 4. For the late fusion experiments, all classifier combinations, i.e., the naive structure, the FAN or the K2 networks, with or without context, with or without temporal filtering, have been tested. The seven best combinations are presented in Table 5.

Table 4: MAP@100 values obtained via cross-validation. Results are reported for the audio and the video modalities, and for early fusion. For each modality, column 1 corresponds to no temporal filter, column 2 to a sliding window averaging, and column 3 to a majority vote.

Network structure	Context	Audio			Video			Early fusion		
		1	2	3	1	2	3	1	2	3
Naive	No	36,3	**39,4**	38,4	25,4	**30,0**	27,9	36,0	**40,3**	37,5
	Yes	36,9	36,2	**37,3**	31,1	30,8	**31,3**	38,5	37,1	**38,5**
FAN	No	26,9	**30,9**	29,3	22,4	**26,9**	25,0	29,0	34,7	**34,8**
	Yes	20,1	20,6	**21,4**	25,5	**27,4**	26,9	25,6	**26,2**	26,1
K2	No	36,3	**39,1**	37,8	26,0	**30,7**	29,0	37,4	**40,9**	39,2
	Yes	36,1	**39,0**	37,0	27,0	27,5	**27,9**	32,3	32,3	**33,2**

It is interesting to note that, while the FAN networks are supposed to perform well in classification, they are outclassed by the K2 and the naive structures in these experiments. As for the other two types of structure, they both seem to provide equivalent results, which shows that structure learning is not always beneficial. One must also note that, if the influence of context is not always clear for the modalities presented in Table 4, temporal filters systematically improve the results, thus showing the importance of the temporal aspect of the signal. However, it is not possible to say which filter provides the best performances. Finally, the importance of multimodal integration is clearly shown as the best results were obtained via both early

Table 5: Results obtained for the seven best late fusion parameter combinations. S_a: Audio structure, C_a: Audio context, S_v: Video structure, C_v: Video context, T_c: Temporal filter applied to the classifiers, T_{lf}: Temporal filter applied after late fusion.

S_a	C_a	S_v	C_v	T_c	T_{lf}	MAP@100
K2	No	Naive	Yes	1	2	43,18
K2	Yes	Naive	Yes	3	2	42,59
K2	Yes	Naive	Yes	1	2	42,55
K2	Yes	Naive	Yes	2	2	42,53
Naive	No	Naive	Yes	3	2	42,45
K2	No	Naive	Yes	3	3	42,36
Naive	No	Naive	Yes	3	3	42,32

and late fusions. The importance of temporal integration is further reinforced by the results obtained via late fusion: among the best combinations, the contextual naive structure is always used for the video modality, and a temporal filter is always used after the fusion step. Moreover, it seems that late fusion performs better than early fusion.

The system chosen and submitted to the 2012 campaign is the best system obtained via late fusion. This system uses a non contextual K2 network for the audio modality, a contextual naive network for the video modality, and a sliding window probability averaging filter after the fusion. It is applied to the test movies and the obtained results are presented in Table 6.

Table 6: Results obtained on the test movies. Column P corresponds to the Precision, R to the Recall, F1 of the F1-measure, and MC to the MediaEval Cost. The values in the MAP@100 column presented for each movie actually correspond to the average precision over the first hundred top ranked samples (AP@100), the MAP@100 being the value in the Total row.

Movie	P	R	F1	MAP@100	MC
Dead Poet Society	5,06	64,71	9,38	60,56	4,09
Fight Club	25,14	58,06	35,09	53,15	3,70
Independence Day	26,22	75,20	38,89	71,76	1,35
Total	**21,72**	**67,27**	**32,83**	**61,82**	**3,57**

The first thing to note is that results are much better than in the cross-validation experiments ($\simeq +18\,\%$). Taking a closer look at the individual results for each movie, it appears that the lowest results are obtained for the movie *“Fight Club”*, while for the other systems presented in the 2012 campaign, the lowest results were usually obtained for *“Dead Poet Society”*. This is encouraging as, contrary to the other systems, this system was able to cope with such a non violent movie. The “low” results obtained for *“Fight Club”* can be explained by the very particular type of violence present in this movie, which might be under-represented in the training database. Similarly, the good results obtained for *“Independence day”* can be explained by its similarity with the movie *"Armageddon"* present in the training set.

These results clearly emphasize again the importance of multimodal integration, through late fusion of classifiers. Finally, the overall result of 61.82 for the MAP@100 is already convincing for the evolution of the task towards real-life commercial systems.

6 Conclusions

Running the Violent Scene Detection task in the framework of the MediaEval benchmark initiative for two years have resulted in two major results: a comprehensive data set to study violence detection in videos, with a focus on Hollywood movies; state-of-the-art multimodal methods which establish a baseline for future research to compare with. Results in the evaluation, demonstrated by the two systems described in this paper, clearly emphasize the crucial role of multimodal integration, either for mid-level concept detection or for direct detection of violence. The two models compared here, namely Bayesian networks and neural networks, have proven beneficial to learn relations between audio and video features for the task of violence detection.

Many questions are still to be addressed, among which we believe two to be crucial. First, Bayesian networks with structure learning, as well as neural networks, implicitly learn the relations between features for better classification. Still, it was observed that late fusion performs similarly. There is therefore a need for better models of the multimodal relations. Second, mid-level concept detection has proven beneficial, reducing the semantic gap between features and classes of interest. There is however still a huge gap between features and concepts such as gunshots, screams or explosions, as demonstrated by various experiments [20, 33]. An interesting idea for the future is that of inferring concepts in a data-driven manner, letting the data define concepts whose semantic interpretation is to be found post-hoc. Again, Bayesian networks and neural networks might be exploited to this end, with hidden nodes whose meaning have to be inferred.

Acknowledgements

This work was partially supported by the Quaero Program. We would also like to acknowledge the MediaEval Multimedia Benchmark for providing the framework to evaluate the task of violent scene detection. We also greatly appreciate our participants for giving us their consent to describe their systems and results in this paper. More information about the MediaEval campaign is available at: http://www.multimediaeval.org/. The working note proceedings of the MediaEval

2011 and 2012 which included the participants' contributions can be found online at http://ceur-ws.org/Vol-807 and http://ceur-ws.org/Vol-927, respectively.

References

1. Violence: a public health priority. Tech. rep., World Health Organization, Geneva, Switzerland (1996). WHO/EHA/SPI.POA.2
2. Acar, E., Albayrak, S.: Dai lab at mediaeval 2012 affect task: The detection of violent scenes using affective features. In: MediaEval 2012, Multimedia Benchmark Workshop (2012)
3. Acar, E., Spiegel, S., Albayrak, S.: Mediaeval 2011 affect task: Violent scene detection combining audio and visual features with svm. In: MediaEval 2011, Multimedia Benchmark Workshop (2011)
4. Baveye, Y., Urban, F., Chamaret, C., Demoulin, V., Hellier, P.: Saliency-guided consistent color harmonization. In: Computational Color Imaging, *Lecture Notes in Computer Science*, vol. 7786, pp. 105–118. Springer Berlin Heidelberg (2013)
5. Chen, L.H., Hsu, H.W., Wang, L.Y., Su, C.W.: Violence detection in movies. In: Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on, pp. 119–124. IEEE (2011)
6. Chen, L.H., Su, C.W., Weng, C.F., Liao, H.Y.M.: Action Scene Detection With Support Vector Machines. Journal of Multimedia **4**, 248–253 (2009). DOI 10.4304/jmm.4.4.248-253
7. Cheng, W.H., Chu, W.T., Wu, J.L.: Semantic context detection based on hierarchical audio models. In: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval, pp. 109–115. ACM (2003)
8. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Machine Learning **9**, 309–347 (1992). URL http://dx.doi.org/10.1007/BF00994110
9. Datta, A., Shah, M., Da Vitoria Lobo, N.: Person-on-person violence detection in video data. In: Pattern Recognition, 2002. Proceedings. 16th International Conference on, vol. 1, pp. 433–438. IEEE (2002)
10. Demarty, C.H., Penet, C., Gravier, G., Soleymani, M.: A benchmarking campaign for the multimodal detection of violent scenes in movies. In: Computer Vision–ECCV 2012. Workshops and Demonstrations, pp. 416–425. Springer (2012)
11. Derbas, N., Thollard, F., Safadi, B., Quénot, G.: Lig at mediaeval 2012 affect task: use of a generic method. In: MediaEval 2012, Multimedia Benchmark Workshop (2012)
12. Eyben, F., Weninger, F., Lehment, N., Rigoll, G., Schuller, B.: Violent scenes detection with large, brute-forced acoustic and visual feature sets. In: MediaEval 2012, Multimedia Benchmark Workshop (2012)
13. Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., Theodoridis, S.: Audio-visual fusion for detecting violent scenes in videos. In: S. Konstantopoulos et al. (ed.) Artificial Intelligence: Theories, Models and Applications, *LNCS*, vol. 6040, pp. 91–100. Springer (2010)
14. Glotin, H., Razik, J., Paris, S., Prevot, J.M.: Real-time entropic unsupervised violent scenes detection in hollywood movies - dyni @ mediaeval affect task 2011. In: MediaEval 2011, Multimedia Benchmark Workshop (2011)
15. Gninkoun, G., Soleymani, M.: Automatic violence scenes detection: A multi-modal approach. In: MediaEval 2011, Multimedia Benchmark Workshop (2011)
16. Gong, Y., Wang, W., Jiang, S., Huang, Q., Gao, W.: Detecting violent scenes in movies by auditory and visual cues. In: Y.M. Huang et al. (ed.) Advances in Multimedia Information Processing - PCM 2008, *LNCS*, vol. 5353, pp. 317–326. Springer (2008)
17. Gravier, G., Demarty, C.H., Baghdadi, S., Gros, P.: Classification-oriented structure learning in bayesian networks for multimodal event detection in videos. Multimedia Tools and Applications pp. 1–17 (2012). DOI 10.1007/s11042-012-1169-y. URL http://dx.doi.org/10.1007/s11042-012-1169-y

18. Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. arXiv.org, http://arxiv.org/abs/1207.0580 (2012)
19. Ionescu, B., Buzuloiu, V., Lambert, P., Coquin, D.: Improved cut detection for the segmentation of animation movies. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (2006)
20. Ionescu, B., Schlüter, J., Mironică, I., Schedl, M.: A naive mid-level concept-based fusion approach to violence detection in hollywood movies. In: Proceedings of the 3rd ACM conference on International conference on multimedia retrieval, pp. 215–222. ACM (2013)
21. Jiang, Y.G., Dai, Q., Tan, C.C., Xue, X., Ngo, C.W.: The shanghai-hongkong team at mediaeval2012: Violent scene detection using trajectory-based features. In: MediaEval 2012, Multimedia Benchmark Workshop (2012)
22. Kriegel, B.: La violence à la télévision. rapport de la mission d'évaluation, d'analyse et de propositions relative aux représentations violentes à la télévision. Tech. rep., Ministère de la Culture et de la Communication, Paris, France (2003)
23. Krug, E.G., Mercy, J.A., Dahlberg, L.L., Zwi, A.B.: The world report on violence and health. The Lancet **360**(9339), 1083–1088 (2002). DOI 10.1016/S0140-6736(02)11133-0. URL http://www.sciencedirect.com/science/article/pii/S0140673602111330
24. Lam, V., Le, D.D., Le, S.P., Satoh, S., Duong, D.A.: Nii, japan at mediaeval 2012 violent scenes detection affect task. In: MediaEval 2011, Multimedia Benchmark Workshop (2012)
25. Lam, V., Le, D.D., Satoh, S., Duong, D.A.: Nii, japan at mediaeval 2011 violent scenes detection task. In: MediaEval 2011, Multimedia Benchmark Workshop (2011)
26. Lin, J., Wang, W.: Weakly-supervised violence detection in movies with audio and video based co-training. In: Advances in Multimedia Information Processing-PCM 2009, pp. 930–935. Springer (2009)
27. Lucas, P.: Restricted Bayesian Network Structure Learning. In: Advances in Bayesian Networks, Studies in Fuzziness and Soft Computing, pp. 217–232 (2002)
28. Ludwig, O., Delgado, D., Goncalves, V., Nunes, U.: Trainable classifier-fusion schemes: An application to pedestrian detection. In: IEEE Int. Conf. On Intelligent Transportation Systems, pp. 432–437 (2009)
29. Martin, V., Glotin, H., Paris, S., Halkias, X., Prevot, J.M.: Violence detection in video by large scale multi-scale local binary pattern dynamics. In: MediaEval 2012, Multimedia Benchmark Workshop (2012)
30. Nam, J., Alghoniemy, M., Tewfik, A.H.: Audio-visual content-based violent scene characterization. In: Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, vol. 1, pp. 353–357. IEEE (1998)
31. Nievas, E.B., Suarez, O.D., García, G.B., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Computer Analysis of Images and Patterns, pp. 332–339. Springer (2011)
32. Penet, C., Demarty, C.H., Gravier, G., Gros, P.: Technicolor and inria/irisa at mediaeval 2011: learning temporal modality integration with bayesian networks. In: MediaEval 2011, Multimedia Benchmark Workshop, *CEUR Workshop Proceedings*, vol. 807. CEUR-WS.org (2011)
33. Penet, C., Demarty, C.H., Gravier, G., Gros, P.: Audio event detection in movies using multiple audio words and contextual Bayesian networks. In: Workshop on Content-Based Multimedia Indexing (2013)
34. Penet, C., Demarty, C.H., Soleymani, M., Gravier, G., Gros, P.: Technicolor/inria/imperial college london at the mediaeval 2012 violent scene detection task. In: MediaEval 2012, Multimedia Benchmark Workshop (2012)
35. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning Representations by Back-Propagating Errors. Nature **323**, 533536 (1986)
36. Safadi, B., Quéenot, G.: Lig at mediaeval 2011 affect task: use of a generic method. In: MediaEval 2011, Multimedia Benchmark Workshop (2011)
37. Schlüter, J., Ionescu, B., Mironică, I., Schedl, M.: Arf @ mediaeval 2012: An uninformed approach to violence detection in hollywood movies. In: MediaEval 2012, Multimedia Benchmark Workshop (2012)

38. de Souza, F.D.M., Cha?vez, G.C., do Valle, E., de A Araujo, A.: Violence detection in video using spatio-temporal features. In: Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on, pp. 224–230. IEEE (2010)
39. de Weijer, J.V., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Trans. on Image Processing **18**(7), 1512–1523 (2009)
40. Zajdel, W., Krijnders, J.D., Andringa, T., Gavrila, D.M.: Cassandra: audio-video sensor fusion for aggression detection. In: Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on, pp. 200–205. IEEE (2007)