A Benchmarking Campaign for the Multimodal Detection of Violent Scenes in Movies

Claire-Hélène Demarty[1], Cédric Penet[1], Guillaume Gravier[2], and Mohammad Soleymani[3]

[1] Technicolor, 1 ave de Belle-Fontaine, 35576 Cesson-Sévigné, France
[2] CNRS/IRISA, Campus de Universitaire de Beaulieu, 263 Avenue du Général Leclerc, 35042 Rennes, France
[3] Department of Computing, Imperial College London, SW7 2AZ, London, United Kingdom

Abstract. We present an international benchmark on the detection of violent scenes in movies, implemented as a part of the multimedia benchmarking initiative MediaEval 2011. The task consists in detecting portions of movies where physical violence is present from the automatic analysis of the video, sound and subtitle tracks. A dataset of 15 Hollywood movies was carefully annotated and divided into a development set and a test set containing 3 movies. Annotation strategies and resolution of borderline cases are discussed at length in the paper. Results from 29 runs submitted by the 6 participating sites are analyzed. The first year's results are promising, but considering the use case, there is still a large room for improvement. The detailed analysis of the 2011 benchmark brings valuable insight for the implementation of future evaluation on violent scenes detection in movies.

1 Introduction

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. This challenge derives from a use case at 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. Given this, 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.

In the literature, violent scenes detection in movies has received very little attention so far. Monomodal static approaches were initially proposed [1,2].

[1] http://www.multimediaeval.org/
[2] http://www.technicolor.com

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 416–425, 2012.

Multimodality has been recently considered in [3,4]. However, the main drawbacks of all these methods lie in the lack of a standard definition of violence and of standard databases. For example, a dataset of 20 minutes is used in [5], a 200-clip collection of scenes from action movies is considered in [6]. In [2,7], four movies are considered for training and testing. Hence the need for a dedicated benchmark for violent scenes detection is beneficial to provide a consistent and substantial dataset, together with a common definition of violence and with evaluation protocols and metrics.

The choice of the targeted content, i.e., Hollywood movies, 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 and special montage effects, which may alter the events to detect. The affect task of MediaEval 2011 therefore constitutes a first attempt to address all these needs.

The paper provides an overview of the 2011 task. Its main contributions are: first, the provision of a definition of violence in movies, second, the description of a comprehensive dataset of 15 Hollywood movies together with their annotations, and valuable insights on the elaboration of the dataset and annotation strategies. Last, this paper reports on the collective effort of the organizers and participants to detect the violent segments in movies. Section 2 details the chosen definition of violence, the task definition and the dataset and evaluation protocols and metrics. In Section 3, results of the benchmark are reported with a short comparative description of the systems. The paper concludes in section 4, with a summary of the lessons learned and directions for a future benchmark.

2 Task Description

The 2011 Affect Task required participants to deploy multimodal approaches to automatically detect portions of movies depicting violence. This calls for a clear definition of violence that serves as a basis for annotating data for the benchmark.

2.1 Towards a Definition of Violence

The notion of violence remains highly subjective as it depends on viewers. The World Health Organization (WHO) defines violence as [8]: "*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 [9]. Each category is divided according to characteristics related to the setting and nature of violence, e.g., physical,

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

sexual, psychological, and deprivation or neglect. In the context of movies and television, Kriegel [10] 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, 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 anesthesia, fit into the definition and were therefore deemed violent. Another borderline case keenly discussed was 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).

2.2 Data Description

In line with the considered use case, the dataset consisted of 15 Hollywood movies from a comprehensive range of genres, from extremely violent to movies without violence. From these 15 movies, the 12 following ones were designated as development data: *Armageddon, Billy Elliot, Eragon, Harry Potter and the*

Order of the Phoenix[4]*, I am Legend, Leon, Midnight Express, Pirates of the Caribbean and the Curse of the Black Pearl*[5]*, Reservoir Dogs, Saving Private Ryan, The Sixth Sense, the Wicker Man*. The three following movies were used as test set: *Kill Bill 1, The Bourne Identity* and *the Wizard of Oz*.

Table 1. Movie dataset (Dev. set: first 12 movies; test set: last 3 movies).

Movie	**Duration**	**Shot length**	**Violence Duration (%)**	**Violent Shots (%)**
Armageddon	8680.16	3562	14.03	14.6
Billy Elliot	6349.44	1236	5.14	4.21
Eragon	5985.44	1663	11.02	16.6
Harry Potter 5	7953.52	1891	10.46	13.43
I am Legend	5779.92	1547	12.75	20.43
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**	**9.52**	**12.7**
Kill Bill	5626.6	1597	17.4	24.8
The Bourne Identity	5877.6	1995	7.5	9.3
The Wizard of Oz	5415.7	908	5.5	5.0
Total	**16919.9**	**4500**	**10.2**	**14.0**

Statistics on violent scenes in each movie are provided in Table 1. The development dataset represents a total of 21,608 shots—as given by automatic shot segmentation developed internally at Technicolor—for a total duration of 83,800 seconds. Violent content corresponds to 9.5% of the total duration and 12.7% of the shots, pointing out the fact that violent segments are not scarce in the database. We tried to respect the genre repartition (from extremely violent to non violent) both in the development and evaluation sets. This appears in the provided statistics, as some movies such as *Billy Elliot* or *The Wizard of Oz* contain a smaller 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.

[4] Harry Potter 5
[5] Pirates Carib. 1

The violent scenes dataset[6] was created by seven human assessors. In addition to segments containing physical violence (with the above definition), annotations also include high-level concepts for the visual modality. For violent segments, the annotation 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 annotator. 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.

[6] The annotations, shot detections and key frames for this task were made available by Technicolor. The dataset can be obtained after signing the User Agreement form, available on the website (`http://www.multimediaeval.org`).

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

2.3 Evaluation Rules

Due to copyright issues, the video content was not distributed and participants were required to buy the DVDs. We provided the online store's URLs for the DVDs which were used for annotations. This was done to ensure that every participant can access the exact, same version of the movies. 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 MediaEval 2011, 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, optionally with a confidence score—the higher the score, the more likely the violence. The segment detection run involved detection of the violent segment boundaries, regardless of the shot segmentation provided.

For official ranking of the systems, system comparison was based on a detection cost function weighting false alarms (FA) and missed detections (MI), according to

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

where the costs $C_{fa} = 1$ and $C_{miss} = 10$ are 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 respectively the FA (false positive) and MI (false negative) rates given the system's output and the reference annotation. In the shot classification, the FA and MI rates 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 the following. To avoid only evaluating systems at given operating points and enable full comparison of systems, we also used detection error trade-off (DET) curves whenever possible, plotting P_{fa} as a function of P_{miss} given a segmentation and a confidence score for each segment. 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.

3 Results

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 (see Table 2). From the number of groups interested and the number of participants, the Affect task has proved to be of interest for the research community. This was confirmed by the active mailing list, which also denoted more a collaborative spirit between the teams than a competitive one, as promoted by MediaEval campaigns. All participants submitted runs for the required shot classification task and none to the optional segment detection. Results are summarized in Table 2.

Table 2. Participation and results (DYN: University of Toulon; NII: National Institute of Informatics; TUB: Technical University of Berlin; UGE: University of Geneva; LIG: Laboratoire d'Informatique de Grenoble; TI: joint participation Technicolor-INRIA. The MediaEval cost value corresponds to the best run per participant, according to this metric. MAP: mean average precision. (*) task organizers.

Part.	**Country**	**# Runs**	**Med. Cost**	**MAP**
DYN	France	2	6.46	0.08
NII	Japan	6	1.00	0.18
TUB	Germany	3	1.26	0.12
UGE*	Switzerland	5	2.00	0.17
LIG	France	1	7.93	0.04
TI*	France	12	0.76	0.25

The 29 submissions mostly correspond to 6 different systems which can be grouped in three main categories. Two participants (NII [11] and LIG [12]) treated the problem of violent scenes detection as a concept detection problem, applying generic systems developed for TRECVid evaluations to violent scenes detection, potentially with specific tuning. Both sites used classic video only features, computed on the keyframes provided, based on color, textures, edges, either local (interest points) or global, and classic classifiers. One participant (DYNI [13]) 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, decision was taken by comparison with a threshold. The last group of participants (TUB [14] , UGE [15] and TI [16]) built dedicated supervised classification systems for the task of violent scenes 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 keyframes (TUB). Both early (TUB, UGE, TI) or late (TI) fusions were used, together with a temporal integration of the decisions at the output of the classifiers (UGE, TI).

Based on the results achieved by the different systems, 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), as used by LIG, NII and TUB, did not add a lot of information to the systems. On the contrary,

taking advantage of different modalities seems to improve performance. This is confirmed by comparing the two runs from DYNI, and the TI runs among which monomodal or multimodal configurations were submitted. Although results do not prove their action 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.

4 Lessons Learned

One goal of this task proposal was to provide a shared framework for violence detection systems for videos. Having the same database and annotations, and the same definition of the violent events is already a significant step towards building of such a framework. However, many lessons were learned from the implementation of the task as a pilot in MediaEval 2011.

Globally, it should first be noted from Table 2 that the overall performances of the proposed systems are not good enough to satisfy the requirements of a real-life commercial product. This means that the problem of automatically detecting violent scenes is still far from being solved and needs further attention.

Figure 1 shows the evolution of the detection error curves. These curves were build using the scores provided by all but one participant. On each of them, only the best run per participant, according to the MediaEval cost, was kept. Instead of giving an evaluation of the systems at different operating points, Figure 1 proposes a more complete comparison. Additionally, it should be noted that the ordering of the systems differs according to the chosen metric.

From Table 2, the metric also appears to be excessively biased towards MI errors or, equivalently, towards high recall at the expense of precision. A ratio of 10 between the FA and MI rates turned out to be so high that it leads to classifying all the scenes into the violent class for some systems (NII). Indeed, classifying all shots as violent results in a MediaEval cost value of 1 which is in most cases lower than what automatic systems obtained. This conclusion calls for a review of the metric in the future, towards a less biased criterion but still reflecting the Technicolor use case.

The definition of violence in the last campaign was chosen to satisfy a need of objectivity in the events, but does not cover all the violent scenes in the context of the Technicolor use case. With such a definition, some actions, e.g., one hurting himself accidentally against a chair, belong to the class of events to detect, despite their minor violent content. Conversely shocking scenes of dead or severely injured people will not be counted as violent. This emphasizes the need for further improvement of the current definition in the future campaigns. Another drawback of this definition is that it complicates the choice of relevant features for the task. For example, the presence of blood which could be a decisive feature is no longer enough to recognize a scene as violent.

A relatively large and standard dataset of movies for violence detection has been developed. The developed dataset is roughly four times larger than the

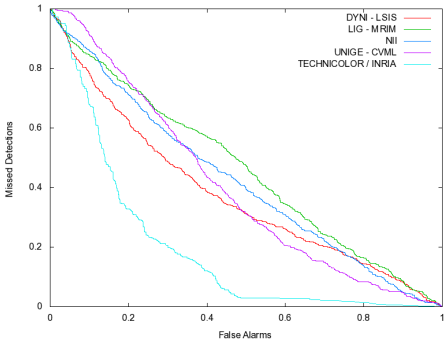

Fig. 1. False alarms vs. missed detections for each best run per participant. The best runs were selected according to the MediaEval cost values.

largest reported dataset in the literature. But even 15 movies is still not large enough to cover the variety of violent scenes in movies. First, violent events remain relatively rare (10% of the dataset). Second, because of the large variation between existing violent actions or events, in a dataset with 15 movies, there are only few similar violent excerpts. Therefore, the development of a larger dataset is certainly beneficial.

5 Conclusions

The Affect Task on Violent Scenes Detection in the context of the MediaEval 2011 benchmarking initiative has been presented. As a pilot task, this first year reached its objective: a common definition of the events to detect, together with a standard dataset and its associated ground truth were proposed, leading to a first solid basis for further research on this topic. Detailing this set-up to the research community was the main contribution of this paper. The task also successfully attracted participants, showing that this task and its open issues are interesting for multimedia indexing and discovery research community. For MediaEval 2012, the task will be further developed, with improvements in the task definition, dataset and the chosen evaluation metrics.

Acknowledgement. We would like to thank all the task participants and other contributors whose combined efforts contributed to make this first year run successful, especially our annotators who did a tedious but nevertheless necessary and valuable job. We also greatly appreciate our participants for giving us consent to describe their systems and results in this paper.

References

1. Pikrakis, A., Giannakopoulos, T., Theodoridis, S.: Gunshot detection in audio streams from movies by means of dynamic programming and Bayesian networks. In: Int. Conf. on Accoustic, Speech and Signal Processing, pp. 21–24 (2008)
2. 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)
3. Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., Theodoridis, S.: Audio-Visual Fusion for Detecting Violent Scenes in Videos. In: Konstantopoulos, S., Perantonis, S., Karkaletsis, V., Spyropoulos, C.D., Vouros, G. (eds.) SETN 2010. LNCS, vol. 6040, pp. 91–100. Springer, Heidelberg (2010)
4. Gong, Y., Wang, W., Jiang, S., Huang, Q., Gao, W.: Detecting Violent Scenes in Movies by Auditory and Visual Cues. In: Huang, Y.-M.R., Xu, C., Cheng, K.-S., Yang, J.-F.K., Swamy, M.N.S., Li, S., Ding, J.-W. (eds.) PCM 2008. LNCS, vol. 5353, pp. 317–326. Springer, Heidelberg (2008)
5. Giannakopoulos, T., Kosmopoulos, D.I., Aristidou, A., Theodoridis, S.: Violence Content Classification Using Audio Features. In: Antoniou, G., Potamias, G., Spyropoulos, C., Plexousakis, D. (eds.) SETN 2006. LNCS (LNAI), vol. 3955, pp. 502–507. Springer, Heidelberg (2006)
6. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence Detection in Video Using Computer Vision Techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds.) CAIP 2011, Part II. LNCS, vol. 6855, pp. 332–339. Springer, Heidelberg (2011)
7. Chen, L.H., Hsu, H.W., Wang, L.Y., Su, C.W.: Violence detection in movies. In: 2011 Eighth International Conference on Computer Graphics, Imaging and Visualization (CGIV), pp. 119–124 (2011)
8. Violence: a public health priority. Technical report, World Health Organization, Geneva, Switzerland (1996) WHO/EHA/SPI.POA.2
9. Krug, E.G., Mercy, J.A., Dahlberg, L.L., Zwi, A.B.: The world report on violence and health. The Lancet 360, 1083–1088 (2002)
10. 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. Technical report, Ministère de la Culture et de la Communication, Paris, France (2003)
11. Lam, V., Le, D.D., Satoh, S., Duong, D.A.: Nii, japan at mediaeval 2011 violent scenes detection task. In: Multimedia Benchmark Workshop, MediaEval 2011 (2011)
12. Safadi, B., Quenot, G.: Lig at mediaeval 2011 affect task: use of a generic method. In: Multimedia Benchmark Workshop, MediaEval 2011 (2011)
13. 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: Multimedia Benchmark Workshop, MediaEval 2011 (2011)
14. Acar, E., Spiegel, S., Albayrak, S.: Mediaeval 2011 affect task: Violent scene detection combining audio and visual features with svm. In: Multimedia Benchmark Workshop, MediaEval 2011 (2011)
15. Gninkoun, G., Soleymani, M.: Automatic violence scenes detection: A multi-modal approach. In: Multimedia Benchmark Workshop, MediaEval 2011 (2011)
16. Penet, C., Demarty, C.H., Gravier, G., Gros, P.: Technicolor and inria/irisa at mediaeval 2011: learning temporal modality integration with bayesian networks. In: Multimedia Benchmark Workshop, MediaEval 2011. CEUR Workshop Proceedings, vol. 807. CEUR-WS.org (2011)