Benchmarking Violent Scenes Detection in Movies

Claire-Hélène Demarty*, Bogdan Ionescu†, Yu-Gang Jiang‡, Vu Lam Quang§, Markus Schedl¶ and Cédric Penet*
* Technicolor, 35576 Cesson Sévigné Cedex, Email: *claire-helene.demarty@technicolor.com, penetcedric@gmail.com*
† LAPI, University Politehnica of Bucharest, 061071 Romania, Email: *bionescu@imag.pub.ro*
‡ Fudan University, Shanghai, China, Email: *ygj@fudan.edu.cn*
§ MMLAB, University of Information Technology, Vietnam, Email: *vulq@uit.edu.vn*
¶ Johannes Kepler University, A-4040 Linz, Austria, Email: *markus.schedl@jku.at*

***Abstract*—This paper addresses the issue of detecting violent scenes in Hollywood movies. In this context, we describe the MediaEval 2013 Violent Scene Detection task which proposes a consistent evaluation framework to the research community. 9 participating teams proposed systems for evaluation in 2013, which denotes an increasing interest for the task. In this paper, the 2013 dataset, the annotations process and the task's rules are detailed. The submitted systems are thoroughfully analysed and compared through several metrics to draw conclusions on the most promising techniques among which multimodal systems and mid-level concept detection. Some further late fusions of the systems are investigated and show promising performances.**

I. Introduction

Detecting violent scenes in movies is an important requirement in various use cases related to video on demand and child protection against offensive content. Apart from the inherent scientific challenge, solving this task requires an adequate formalisation of this highly subjective concept, i.e., violence.

The World Health Organization [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". From the literarily perspective of the Cambridge dictionary[1], violence represents "actions or words which are intended to hurt people". A formalization adapted to the movie context can be the one proposed by the French Ministry of Culture and Communication [2] where TV violence is defined 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 for instance accidents, which also result in potentially shocking gory and graphic scenes. A more adapted formalization should be therefore investigated in the aforementioned context.

Due to the complexity of the research problem, starting with the formulation of violence to the inference of highly semantic concepts out of low-level information, the problem of violence detection in videos has been marginally studied in the literature. This problem requires carrying out several tasks, ranging from *detection of affective content* which refers to the characterisation of emotions that are expected to arise in the user while watching a video [4]; *action recognition* which focuses on detecting *human violence in real-world scenarios* such as for instance fight detection [5]; to a broader category of methods that focus on a more general framework, such as detecting *video segments with violent content* that may be considered disturbing for different categories of viewers [6].

Another important aspect is the validation of the techniques. Before 2011, there was a lack of a standard consistent and substantial evaluation framework (both from the dataset and annotations point of view). This limited significantly the reproducibility of results in the community and consequently the advances in this specific field. Each of the proposed methods tended to be tested on closed data, usually very restraint and annotated for very particular types of violence. For instance, in [7] instances of aggressive human behavior in public environments were detected on 13 clips featuring various scenarios, such as "aggression towards a vending machine" or "supporters harassing a passenger"; or [5] which tests fight detection using 1,000 clips containing different sport actions from ice hockey videos.

In this paper we address the problem of violence detection in typical Hollywood productions and introduce the 2013 edition of the evaluation framework, the Affect Task: Violent Scenes Detection (VSD) [8], which has been run annualy since 2011 during the MediaEval Benchmarking Initiative for Multimedia Evaluation [3]. The main focus of this work is to provide a comparative analysis of the state-of-the-art systems submitted to the task. This analysis is helpful in that it evidences strong and weak points of current violence detection techniques and can be used to guide further work in the area. The VSD task takes root in a use case at *Technicolor*[2], which is to help users choose movies that are suitable for their children, in terms of violent content. To this end, the user is provided with a summary of the most violent scenes in a given movie.

Compared to the 2011 and 2012 editions [20], a novelty of the task consists of addressing two distinct use case scenarios for the definition of violence with the aim to understand how different scenarios influence the systems' performance. Furthermore, a very consistent benchmarking dataset was made publicly available providing full annotations for no less than 25 Hollywood movies. Finally, the 2013 edition allowed participants to submit systems that make use of external data (e.g., from Internet) which allows for testing open systems.

This work was supported, in part, by the following projects: Quaero Program (http://www.quaero.org/), China's National 973 Program #2010CB327900, China's NSF Projects #61201387 and #61228205 and the VNU-HCM Project #B2013-26-01. We acknowledge the MediaEval Benchmarking Initiative for Multimedia Evaluation (http://www.multimediaeval.org/) for supporting the Violent Scenes Detection Task.

[1] http://dictionary.cambridge.org/

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

Each of these aspects are presented in the sequel. Section II provides a detailed description of task and ground truth annotation. Section III overviews the key contributions of each submitted system. Experimental results are presented in Section IV: we elaborate on the used evaluation metrics and give a comparative analysis of the submitted algorithms. Section V concludes and presents the outlook of the results.

II. Task and dataset description

In this section we focus on presenting the benchmarking framework as well as the provided data and annotations.

A. Description of the dataset

The 2013 edition of the dataset was built on top of the 2012 edition. All the movies used in 2012 were adopted as training data (*development dataset*) and a set of seven additional movies were provided for testing (*test dataset*; see also Table I). The 2013 data set was made publicly available[3].

The main novelty of the 2013 benchmarking consists of adopting two different definitions of violence according to two different use case scenarios. The first use case scenario is a following of what was proposed in the previous years, where targeted violent segments are those showing "*physical violence or accident resulting in injury or pain*" (denoted *objective definition*). Although it was designed to be as objective as possible, this definition has proven to lead to inconsistencies/ambiguities between the annotated segments and the original Technicolor use case, e.g., not really violent segments such as "somebody hurting himself while shaving" were considered as violent whereas segments depicting dead people but without showing the cause of death were discarded from the annotations. To experiment with another perspective of violence, the second use case features a more subjective definition, namely violent segments are "*those which one would not let an 8 years old child see because they contain physical violence*" (*subjective definition*). Data was annotated for both use case scenarios.

For the objective definition, the annotations were carried out by three human assessors using the following protocol. Firstly, two annotators labelled all the videos separately. During the annotation process, no discussions were held between them in order for the process to be totally independent. Secondly, a third master annotator merged all their annotations and reviewed the movies once again to minimize the chance of missing any violent segments. Doubtful annotations were solved via panel discussions. Each annotated violent segment contains a single action of violence whenever possible. However, if there are multiple actions in a continuous segment, the segment was annotated as a whole. The annotation granularity was decided to be at frame level. Each violent segment is accompanied by a short textual label describing its contents.

For the subjective definition, the annotations were carried out by seven human assessors (5 regular annotators and 2 master annotators). Given the specificity of this scenario it is worth mentioning the profile of the annotators: regular annotators were graduate students (single with no children) and master annotators were lecturers (married with children). In this case the following protocol was used. Firstly, two regular annotators labelled all the movies separately. Secondly, the third regular annotator merged, reviewed and also revisited the movies to retrieve any possible missing violent segments. Once again, no discussions were held between annotators. Finally, a fourth master annotator reviewed the data from a parent perspective and refined the results. All the uncertain (borderline) cases were solved via panel discussions, involving different people from different countries and culture, to avoid cultural bias in the annotations. A textual description was added to each segment. In contrast with the objective definition where violent segments focused on violent actions and their results, the subjective violent segments focus on the overall context of violent scenes. As a result, subjective segments tend to be slightly longer than the objective ones.

Table I: The 2013 movie dataset. The columns indicate the duration in seconds, the number of shots, and their proportions, for both the objective and the subjective violences.

2013 development dataset			**objective**		**subjective**	
movie	**dur.(s)**	**#shots**	**dur.(%)**	**#shots(%)**	**dur.(%)**	**#shots(%)**
Armageddon	8,680.16	3,562	10.16	11.0	9.33	13.08
Billy Elliot	6,349.44	1,236	5.14	4.21	4.77	5.33
Eragon	5,985.44	1,663	11.02	16.6	16.04	27.23
Harry Potter 5	7,953.52	1,891	9.73	12.69	8.93	17.39
I am Legend	5,779.92	1,547	12.45	19.78	17.71	32.12
Kill Bill	6,370.4	1,597	17.47	23.98	27.82	40.70
Leon	6,344.56	1,547	4.3	7.24	18.51	28.24
Midnight Express	6,961.04	1,677	7.28	11.15	8.17	14.25
Pirates Carib. 1	8,239.4	2,534	11.3	12.47	19.89	26.44
Reservoir Dogs	5,712.96	856	11.55	12.38	34.37	35.51
Saving Private Ryan	9,751.0	2,494	12.92	18.81	34.54	47.91
The Bourne Identity	6,816.0	1,995	7.61	9.22	9.50	12.88
The Sixth Sense	6,178.04	963	1.34	2.80	2.49	5.50
The Wicker Man	5,870.44	1,638	8.36	6.72	11.74	11.78
The Wizard of Oz	5,859.2	908	5.5	5.06	1.14	2.42
Dead Poets Society	7,413.2	1,583	1.5	2.14	0.72	1.45
Fight Club	8,004.5	2,335	13.51	13.27	19.24	22.09
Independence Day	8,833.9	2,652	9.92	13.98	14.62	24.13
Total	**127,103.1**	**32,678**	**9.12**	**12.0**	**14.74**	**21.45**
	35h18					

2013 test dataset			**objective**		**subjective**	
movie	**dur.(s)**	**#shots**	**dur.(%)**	**#shots(%)**	**dur.(%)**	**#shots(%)**
Fantastic Four 1	6,093.96	2,002	16.08	21.23	22.54	35.81
Fargo	5,646.4	1,061	8.78	13.38	15.73	24.12
Forrest Gump	8,176.72	1,418	11.39	12.55	8.83	16.78
Legally Blond	5,523.44	1,340	1.02	0.97	0	0
Pulp Fiction	8,887.0	1,686	8.56	8.13	25.97	29.41
The God Father	10,194.7	1,893	4.51	6.08	6.32	10.45
The Pianist	8,567.04	1,845	8.29	9.21	16.89	20.10
Total	**53,089.3**	**11,245**	**8.28**	**10.49**	**13.91**	**20.24**
	14h44					

As for the previous editions of the benchmark, in addition to general violent segments annotation, a set of 10 high-level violence related concepts were annotated, i.e., presence of blood, fights, presence of fire, presence of guns, presence of cold weapons, car chases, gory scenes, gunshots, explosions and screams (for more details see [19][20]).

Table I gives some statistics on the data. The development set contains 32,678 shots (as obtained with automatic segmentation) from 18 movies for a total duration of 35h18min. According to the objective definition, violent shots cover 12% of the shots and 9.12% of the total duration, whereas for the subjective definition, violent shots represent 21.45% of the shots and 14.74% of the duration. These figures highlight the fact that globally the subjective definition proposes segments of longer durations, and therefore covers a bigger proportion

[3] To download the data, please follow the instructions at: http://research.technicolor.com/rennes/vsd/.

of the database. This comes from the differences both in the definitions and the annotations, e.g., the subjective annotations take into account the global violence context whereas for the objective annotations violence is annotated locally with frame-level precision. The dataset consists of a large variety of Hollywood movies, which range from highly violent ones (covering a wide variety of violence types, e.g., war, disasters, etc) to a few with almost no violence at all. For instance, *Dead Poets Society* contains very little violence both for the objective and the subjective definitions (respectively 2% and 1.45% of the shots). On the other hand, the most violent movie changes from one definition to another: *Kill Bill* contains the largest proportion of objective violent shots (24% of the total shots), while in contrast *Saving Private Ryan* is the most violent movie for the subjective definition (48% of the shots) containing a lot of scenes with dead people that were included in the subjective definition but not in the objective one.

The 2013 test set is the largest in the history of this benchmarking with 7 movies (containing non violent to highly violent movies; total duration of 14h44min and 11,245 shots). Looking at numbers in Table I, once again one may notice that subjective annotations globally reach higher violence proportions than the objective ones. For a given movie, their proportions also varies, highlighting the differences in the two definitions.

B. Description of the benchmarking

The proposed benchmarking framework was validated during the MediaEval 2013 Violent Scenes Detection Task [8]. It asked participants to automatically detect violent portions of Hollywood movies by the use of multimodal features. As explained in Section II-A, two definitions of violence were considered in 2013 leading to two different sub-staks.

For each substak, participants were allowed to submit the following types of runs (up to 5 runs): shot-based classification without use of any external data other than the content of the DVDs (shot segmentation is provided by organizers), shot-based classification with use of external data, segment level classification without external data (participants are required to provide segment boundaries independently of the shot segmentation) and segment level classification with external data. In each case, each shot or segment has to be provided with a confidence score. For both subtasks, the required run is the run at shot level without use of external data.

In 2013, the proposed benchmarking has seen a substantial increase both in the number of persons looked highly interested in the task (54 vs. 35 in 2012 and 12 in 2011 - numbers recorded during a community survey that runs each year prior to the task and involving more than 150 respondents) and in the registrations number (18 teams vs. 10 in 2012 and 5 in 2011). These 18 teams, which could be broken up into 4 organising teams and 14 additional teams, were representing 22 research groups (including 3 joint submissions), coming from 16 countries all over the world. In total, 59 runs have been evaluated, divided between the objective (36 runs) and the subjective (23 runs) subtasks.

III. System descriptions

In total, from the 18 teams registered to the 2013 MediaEval Violent Scenes Detection Task, 9 crossed the finish line and submitted results for both objective and subjective tasks. Same systems were used to address both subtasks, the only difference being in the training data (this makes subtask results comparable). In the following we overview the key contributions of each submitted system.

-**FAR** (*machine learning, cascade classifiers — multimodal, mid-level concepts*) [10]: uses a machine learning scheme to predict violence at frame level. Video content is described in terms of visual (color histograms, Histograms of Oriented Gradients (HoG) and visual activity), auditory information (e.g., Linear Predictive Coeffcients, Mel-Frequency Cepstral Coeffcients (MFCC) and their statistics in time windows) and mid-level concepts. Classification is carried out with Multi-Layer Perceptrons (MLP) trained by backpropagating cross-entropy error and random dropouts to reduce overfitting. Mid-level descriptors are determined as the outputs of the MLPs for the provided violence related concepts (e.g., blood, firearms). Audio-visual fusion is achieved using early fusion and audio-visual-concept integration is achieved with late fusion;

-**FUDAN** (*machine learning, temporal score smoothing — multimodal, mid-level concepts*) [9]: uses Support Vector Machines (SVMs) to classify video shots as violent - non-violent. Visual information is represented with trajectory-based features that include a Bag-of-Visual-Words representation of HoG, Histograms of Optical Flow (HoF), Motion Boundary Histograms (MBH) and trajectory shape information, as well as the description of motion relationships between trajectory pairs; together with Space-Time Interest Points and a descriptor called part-level attribute that is derived from object detection (e.g., outputs the likelihood that a frame contains particular objects/scenes). Audio content is represented with MFCCs. Multimodal integration is achieved using late fusion with score-level averaging. Final prediction scores are smoothed by taking the average value over a three-shot window;

-**LIG** (*machine learning, temporal re-ranking — multimodal*) [13]: uses a late fusion approach of SVMs (for a better handling of class imbalance) and k nearest neighbors (kNN) classifiers. Outputs of individual classifiers are merged by taking the linear combination of the prediction scores whose weights are optimized on the development dataset. Video information is represented with visual (color, texture, feature points — Bag-of-Visual-Words of SIFT and HoF), audio (Bag-of-Audio-Words of MFCCs) and audio-visual descriptors (combination of MFCC and HoF). The system uses also a temporal re-ranking scheme where shot level violence prediction scores are re-evaluated according to some global (video level) or local (neighborhood level) score estimations;

-**MTM** (*machine learning, statistical analysis — multimodal*) [15]: explores the spatial correlation between acoustic (MFCCs and their first and second derivatives) and visual features (optical flow features: average velocity and acceleration magnitude). Audio-visual information is converted to canonical base vector representations using Canonical Correlation Analysis, which maximizes the correlation between two multivariate random vectors. Features are combined using early fusion. Final shot classification is achieved with a Bayesian Network;

-**NII-UIT** (*machine learning — multimodal, mid-level concepts*) [16]: achieves shot level violence prediction using SVMs. The video features include MPEG-7-like color and

Table II: Overall MAP@100 and MAP for best team runs (according to the official metrics).

team	*objective*						*subjective*					
	shots			segments			shots			segments		
	runid	MAP@100	MAP	runid	MAP@100	MAP	runid	MAP@100	MAP	runid	MAP@100	MAP
FUDAN [9]	run5(avc-l)	**0.5531**	**0.5114**	-	-	-	run5(avc-l)	0.6816	0.5862	-	-	-
LIG [13]	run2(av-l)	0.5208	0.5051	-	-	-	run1(av-l)	**0.6904**	0.6731	-	-	-
FAR [10]	run1(a)	0.4958	0.4764	run5(avc-l)	0.3504	**0.3452**	-	-	-	-	-	-
TUDCL [14]	run2(av-l)	0.4695	0.3866	run1(av-l)	**0.4202**	0.3434	-	-	-	-	-	-
NII-UIT [16]	run1(avc-l)	0.4361	0.2339	-	-	-	run1(avc-l)	0.5959	0.3793	-	-	-
TECH-INRIA [17]	run1(c_a)	0.3382	0.2882	run3(c_{av}-l)	0.1248	0.1464	run1(c_a)	0.5359	0.4456	run1(c_a)	**0.4479**	**0.353**
VIREO [12]	run4(avc-l)	0.3157	0.3157	-	-	-	run4(avc-l)	0.6896	**0.6752**	-	-	-
MTM [15]	run1(av-e)	0.0738	0.1258	-	-	-	-	-	-	-	-	-
VISILAB [11]	-	-	-	run2(v)	0.1498	0.1388	-	-	-	-	-	-

texture descriptors, SIFT-based representations and Bag-of-Visual-Words, motion information (MBH with Fisher kernel representations), audio descriptors (MFCC also with Fisher representations) and mid-level descriptors (predictions of the provided violence related concepts, e.g., fire, gore, etc that are trained on the 2012 dataset). Features are combined using a late fusion integration;

-**TECH-INRIA** (*machine learning, cascade classifiers — multimodal, mid-level concepts*) [17]: builds around the idea of using mid-level concept detectors as input to a global violence detector [18]. Video information is represented at segment level with audio concept detectors (MFCCs, energy and flatness coefficients fed into Bayesian networks to predict audio concepts, e.g., explosions, screams, etc) and video concepts/feature detectors (e.g., shot statistics, color, blood color proportion, flash and fire detection, motion detection). Final prediction of violence is achieved with naïve contextual Bayesian networks. Experiments are conducted with both early fusion (single classifier) and late fusion (first audio and visual classifications and results fed to a final Bayesian classifier);

-**TUDCL** (*multiple kernel learning — multimodal*) [14]: uses a Multiple Kernel Learning (MKL) framework to determine optimal SVM kernel weighting (different kernels are assigned for each feature space). Video information is represented with visual temporal descriptors (e.g., motion trajectories, MBH, color histograms around trajectories) and Bag-of-Visual-Words and audio descriptors (MFCCs with Bag-of-Audio-Words representations). Final prediction of violence is achieved in two steps: preliminary segment predictions are determined as the sum of the individual classifiers' outputs, while final prediction involves a moving average smoothing of the previous scores;

-**VIREO** (*machine learning, concept refinement using web data, cascade SVM classifiers — multimodal, mid-level concepts*) [12]: uses dense trajectories (through HoG, HoF, MBH and shape features), SIFT feature points and Bag-of-Visual-Words, audio descriptors (e.g., octave band signal intensity, MFCCs) and mid-level concepts. In particular, for mid-level concept description, provided concepts are used to infer 42 additional ones using ConceptNet (e.g., punishment, victim, rape, etc). New concepts are trained with Youtube data. Prediction scores for the concepts are used as descriptors. Concepts are refined by filtering out redundant information via the analysis of the co-occurrence information in a concept graph model ontology. Multimodal integration is achieved via late fusion;

-**VISILAB** (*machine learning — visual*) [11]: approaches the issue of violent shot prediction by adapting a visual-based fight detector. Video information is represented with descriptors adapted to this particular application and use extreme acceleration patterns that are estimated via the Radon transform of the power spectrum of consecutive frames. Classification is performed with either SVMs or kNN.

IV. Experimental results

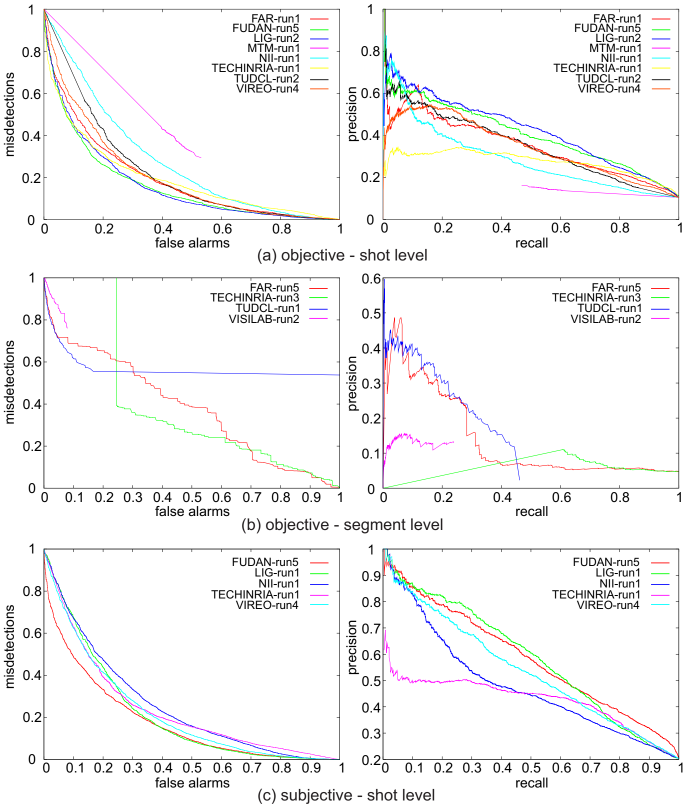
This section presents the results achieved during the 2013 evaluation campaign. 36 runs were submitted for the objective use case, among which 30 were targetting a shot level prediction and 6 a segment level prediction, where segments could be of arbitrary length. For the subjective scenario, 23 runs were received: 21 runs for shot level prediction while only 2 for the segment level prediction. During the competition, participants designed and trained their methods on the development dataset, while the actual benchmarking was conducted on the test dataset (see datasets details in Section II-A and Table I).

To assess performance, similar to the last years' benchmarkings, several metrics were computed, from false alarm and miss detection rates, AED-precision/recall, MediaEval cost (a function weighting false alarms and missed detections) to Detection Error Trade-off curves and Mean Average Precision (more details are presented in [8]). However, in 2013, the official metric was selected to be the standard Mean Average Precision (MAP) which is defined as the average value of the Average Precision achieved at movie level. In particular, in 2013, systems were optimized for a cutoff point of 100 top ranked violent segments (MAP@100).

A. Evaluation in terms of MAP

Table II reports the MAP@100 and MAP metrics for the best team runs for both objective and subjective use cases as well as for shot and segment level evaluation (notations: a - audio, v - visual, c - mid-level concepts, l - late fusion and e - early fusion; highest values are represented in bold). What is interesting to notice is that regardless of the use case scenario and the granularity of the prediction, highest performance is achieved when including mid-level information with multimodal late fusion approaches (see avc-l runs): objective shot level prediction — MAP=0.5114, FUDAN run5 [9], objective segment level — MAP=0.3452, FAR run5 [10] and subjective shot level — MAP=0.6752, VIREO run4 [12].

At modality level, visual information alone seems to provide too little discriminative power for this high level task, e.g., VISILAB run2 [11] is able to achieve a MAP of only 0.1388 which is less than half the performance of the best system. Another interesting result is that in particular, using audio



(a) objective - shot level

(b) objective - segment level

(c) subjective - shot level

Figure 1: Misdetections/false alarms and precision/recall curves for best participant runs (see also Table II).

modality alone, it allows to achieve very good performance, e.g., for objective use case FAR run1 [10] leads to a MAP of 0.4764 while TECH-INRIA run1 [17] reaches a MAP of 0.4456 for the subjective scenario. This may be due to the fact that most of the violent scenes in movies tend to come with specific audio signatures.

In what concerns the granularity of the predictions, shot-based estimation is more accurate than the prediction at arbitrary length segments, e.g., TUDCL run2 [14] leads to MAP=0.3866 for shot level while the same run achieves MAP=0.3434 for segments (the difference is greater for the best performing runs). Tagging directly some predefined shots is indeed a classification task, and therefore easier than the task at segment level, where a step of boundaries segmentation is involved. Furthermore, systems proposing oversegmented events at segment level will be penalized during MAP computation, as potentially a higher number of false alarms (one per segment) may be ranked in the first 100 returned results.

The predictions of the subjective use case scenario lead to significantly higher results than for the objective one, e.g., highest MAP at shot level is 0.5114 (FUDAN run5 [9]) for the objective scenario while the same run achieves up to 0.5862 for the subjective one. A possible explanation comes from the fact that the subjective annotations lead to longer and more unitary shots than for the objective one, where the focus was on identifying each particular individual scene (see total results in Table I). Moreover, in the objective use case, systems may have difficulties in classifying as violent events such as *somebody simply pushing another person*, which nevertheless fit with the objective definition, especially as the chosen features are better adapted to typical violent scenes, involving explosions, blood, etc, than to those specific cases.

Table III: AP@100 at movie level for best team runs.

	objective		*subjective*	
movie	shots FUDAN[9] run5	segments TUDCL[14] run1	shots LIG[13] run2	segments TECH-INRIA[17] run1
Fantastic Four 1	0.6372	0.5739	**0.9602**	**0.7035**
Fargo	**0.8445**	**0.7081**	0.9391	0.6534
Forrest Gump	0.6501	0.4535	0.7587	0.4175
Legally Blond	0.0181	0.0133	0	0
Pulp Fiction	0.5114	0.2474	0.7104	0.4365
The God Father 1	0.7808	0.6078	0.635	0.3775
The Pianist	0.4295	0.3370	0.8296	0.5465

Finally, in what concerns the cutoff point, reporting MAP@100 leads to slightly better results than the overall MAP prediction. This is useful in case the violence prediction system is considered from the perspective of retrieval where violence segments are searched within the movies. In this case, highest performing system is the one retrieving the largest number of best results at the first top ranks.

B. *Evaluation of false and missed detections*

As shown in Figure 1, the overall performances in terms of false and missed detections are similar for the best participants and reach 20% false alarms for 20% missed detections for objective definition at shot level, and 25% false alarms for 25% missed detections for subjective definition at shot level. For runs at segment level and for the objective scenario, performances vary from one system to another and achieve at best 40% false alarms for 25% missed detections. Recall/precision curves show that, regardless of the scenario, all systems reach high recall values (which corresponds to the targeted operating point, where one does not want to miss any violent scene) at the expense of very low precision values (between 0.1 and 0.2). The relative rarety of the events to detect in the dataset (8.28% of the duration for objective and 13.91% for subjective) partly explains these values. Last, it should be noted that the ranking of the best performing systems slightly changes while considering recall/precision or false and missed detections curves, compared to the official ranking based on MAP@100.

C. *Evaluation at movie level*

Table III presents the best runs in terms of Average Precision (AP) at 100 segments for each test movie. As expected, due to the high variability of movie content, results are very different. For instance, the movie *Legally Blond* which independently of the scenario has no or very few violence annotations leads to very low or null AP values. Again, results for the subjective use case tend to be much more accurate than for the strict objective definition.

D. *Late fusion of systems*

As a final experiment we aimed at constructing a super violence prediction system that exploits the advantages of each individual system. We investigated several fusion schemes by either keeping the intersection of all input systems' violent shots (*inter*), or by taking their union (*union*). In both cases, two resulting confidence scores were computed, firstly by averaging all confidence values from all input systems (*ave*), secondly by keeping the maximal value (*max*). To investigate

Table IV: Fusion of the best systems' results (MAP@100).

type	*best runs*		*fusion*	
obj. shot	FUDAN	**0.5531**	FUDAN+LIG — ave/union	0.4055
	LIG	0.5208	FUDAN+LIG — ave/alltrue	0.4039
	FAR	0.4958	FUDAN+FAR — ave/union	0.4803
			FUDAN+FAR — ave/alltrue	0.4918
			LIG+FAR — ave/union	**0.5558**
			LIG+FAR — ave/alltrue	**0.5480**
			FUDAN+LIG+FAR — ave/union	0.5383
			FUDAN+LIG+FAR — ave/alltrue	0.5343
obj. segm.	TUDCL	**0.4202**	TUDCL+FAR — max/inter	**0.4409**
	FAR	0.3504	TUDCL+FAR — max/alltrue	**0.7186**
subj. shot	LIG	**0.6904**	LIG+FUDAN — ave/alltrue	0.5816
	VIREO	0.6896	LIG+FUDAN — max/union	0.6037
	FUDAN	0.6816	LIG+VIREO — ave/union	**0.7164**
			LIG+VIREO — ave/alltrue	**0.7139**
			FUDAN+VIREO — max/union	0.6468
			FUDAN+VIREO — max/alltrue	0.6468
			FUDAN+LIG+VIREO — max/union	0.6004
			FUDAN+LIG+VIREO — max/alltrue	0.6004

the accuracy of the confidence values only, some additional results were tested by classifying all shots as violent (*alltrue*).

For each use case scenario we use different combinations of the two or three best performing systems. Table IV summarizes the performances of the best fusion schemes for each selection of input systems. In each use case scenario, depending on the systems used as input, the fusion improves the MAP@100 at least slightly. Except for the objective segmentation, there is no significant change in performance when including the final decisions. This highlights the fact that the initial chosen thresholds in this case may not be optimized for the task.

At shot level, for both scenarios, the fusion scheme *ave/union* gives the best improvement, leading to the conclusion that the input systems do not have a lot of false alarms or have complementary correct detections. One may note that in both cases the best fused system only takes a selection of the best individual systems as input. For the subjective scenario, a finer analysis of the confidence scores returned by FUDAN, LIG and VIREO shows that for some movies, FUDAN and LIG systems return very similar confidence values, giving some first insight on why their fusion might not be effective.

Judging from the results, one may conclude that for this particular task, having a late fusion hybrid cascade classifier is a more effective solution than using individual classifiers and early fusion. However, not any system can be exploited in this scheme. Preliminary results fusing all approaches led to lower results than the best performing systems. The fusion was able to provide improvement only by fusing two or three best systems as presented in Table IV.

V. Conclusions and outlook

We presented the Violent Scenes Detection task that is held in conjunction with the MediaEval Benchmarking Initiative for Multimedia Evaluation and gave a comparative overview of the systems proposed in 2013. Judging from the results, we believe that the proposed task stands as a consistent and standardized benchmarking framework for violence detection in movies. Its publicly available annotated dataset provides a relevant testbed for the evaluation of a broad category of multimodal approaches. Several perspectives may be drawn for future extensions of this framework: we should head towards a qualitative evaluation, in addition to the quantitative metrics currently used. In 2013, although strongly motivated to submit a video summary of their detected scenes, only one team did so. If provided, these summaries could be used to conduct user surveys that will give further insight on the results; we should investigate strategies to expand the task to other types of video material, e.g., user-generated content, and see how the proposed systems generalize to different types of content; we will continue promoting multimodal approaches to the task by encouraging participants to use metadata (e.g., from the Internet) as a complement to the classic audiovisual features.

References

[1] "Violence: a public health priority", World Health Organization, 1996.

[2] B. Kriegel, "La violence à la télévision", Rapport de Ministère de la Culture et de la Communication, Paris, France, 2003.

[3] MediaEval 2013 Workshop, Eds. M. Larson, X. Anguera, T. Reuter, G.J.F. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, M. Soleymani, Barcelona, Spain, CEUR-WS.org, ISSN 1613-0073, Vol. 1043, http://ceur-ws.org/Vol-1043/, 2013.

[4] A. Hanjalic, L. Xu, "Affective Video Content Representation and Modeling", IEEE Trans. on Multimedia, pp. 143-154, 2005.

[5] E. Bermejo, O. Deniz, G. Bueno, R. Sukthankar, "Violence Detection in Video using Computer Vision Techniques", Int. Conf. on Computer Analysis of Images and Patterns, LNCS 6855, pp. 332-339, 2011.

[6] C. Penet, C.-H. Demarty, G. Gravier, P. Gros, "Multimodal Information Fusion and Temporal Integration for Violence Detection in Movies", IEEE ICASSP, Kyoto, 2012.

[7] W. Zajdel, J. D. Krijnders, T. Andringa, D. M. Gavrila, "CASSANDRA: audio-video sensor fusion for aggression detection", IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, pp. 200-205, 2007.

[8] C.-H. Demarty, C. Penet, M. Schedl, B. Ionescu, V.L. Quang, Y.-G. Jiang, "The MediaEval 2013 Affect Task: Violent Scenes Detection", Working Notes Proc. [3], 2013.

[9] Q. Dai, J. Tu, Z. Shi, Y.-G. Jiang, X. Xue, "Fudan at MediaEval 2013: Violent Scenes Detection using Motion Features and Part-Level Attributes", Working Notes Proc. [3], 2013.

[10] M. Sjöberg, J. Schlüter, B. Ionescu, M. Schedl, "FAR at MediaEval 2013 Violent Scenes Detection: Concept-based Violent Scenes Detection in Movies", Working Notes Proc. [3], 2013.

[11] I. Serrano, O. Déniz, G. Bueno, "VISILAB at MediaEval 2013: Fight Detection", Working Notes Proc. [3], 2013.

[12] C.C. Tan, C.-W. Ngo, "The Vireo Team at MediaEval 2013: Violent Scenes Detection by Mid-level Concepts Learnt from Youtube", Working Notes Proc. [3], 2013.

[13] N. Derbas, B. Safadi, G. Quénot, "LIG at MediaEval 2013 Affect Task: Use of a Generic Method and Joint Audio-Visual Words", Working Notes Proc. [3], 2013.

[14] S. Goto, T. Aoki, "TUDCL at MediaEval 2013 Violent Scenes Detection: Training with Multimodal Features by MKL", Working Notes Proc. [3], 2013.

[15] B.D.N. Teixeira, "MTM at MediaEval 2013 Violent Scenes Detection: Through Acoustic-visual Transform", Working Notes Proc. [3], 2013.

[16] V. Lam, D.-D. Le, S. Phan, S. Satoh, D.A. Duong, "NII-UIT at MediaEval 2013 Violent Scenes Detection Affect Task", Working Notes Proc. [3], 2013.

[17] C. Penet, C.-H. Demarty, G. Gravier, P. Gros, "Technicolor/INRIA Team at the MediaEval 2013 Violent Scenes Detection Task", Working Notes Proc. [3], 2013.

[18] B. Ionescu, J. Schlüter, I. Mironică, M. Schedl, "A Naive Mid-level Concept-based Fusion Approach to Violence Detection in Hollywood Movies", ACM Int. Conf. on Multimedia Retrieval, USA, 2013.

[19] C.-H. Demarty, C. Penet, M. Soleymani, G. Gravier, "VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation", MTAP, 2013 (accepted).

[20] C.-H. Demarty, C. Penet, B. Ionescu, G. Gravier, M. Soleymani, "Multimodal violence detection in Hollywood movies: State-of-the-art and Benchmarking", in book "Fusion in Computer Vision - Understanding Complex Visual Content", Springer, to appear 2014.