Multimed Tools Appl
DOI 10.1007/s11042-014-1984-4

VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation

Claire-Hélène Demarty · Cédric Penet · Mohammad Soleymani · Guillaume Gravier

Abstract Content-based analysis to find where violence appears in multimedia content has several applications, from parental control and children protection to surveillance. This paper presents the design and annotation of the Violent Scene Detection dataset, a corpus targeting the detection of physical violence in Hollywood movies. We discuss definitions of physical violence and provide a simple and objective definition which was used to annotate a set of 18 movies, thus resulting in the largest freely-available dataset for such a task. We discuss borderline cases and compare with annotations based on a subjective definition which requires multiple annotators. We provide a detailed analysis of the corpus, in particular regarding the relationship between violence and a set of key audio and visual concepts which were also annotated. The VSD dataset results from two years of benchmarking in the framework of the MediaEval initiative. We provide results from the 2011 and 2012 benchmarks as a validation of the dataset and as a state-of-the-art baseline. The VSD dataset is freely available at the address: http://www.technicolor.com/en/innovation/research-innovation/scientific-data-sharing/violent-scenes-dataset.

Keywords Content-based analysis · Multimedia evaluation · Violent scene detection · Corpus design · Physical violence definition · Semantic audio concepts · Semantic video concepts

C. -H. Demarty (✉) · C. Penet (✉)
Technicolor, 975 avenue des Champs Blancs, ZAC des Champs Blancs,
35576 Cesson Sévigné, France
e-mail: claire-helene.demarty@technicolor.com
e-mail: penetcedric@gmail.com

M. Soleymani
Imperial College London, Huxley building, 180 Queen's gate, London SW7 2AZ, UK
e-mail: m.soleymani@imperial.ac.uk

G. Gravier
CNRS - Irisa, Inria Rennes Campus de Beaulieu, 35042 Rennes Cedex, France
e-mail: guig@irisa.fr

Published online: 15 May 2014

1 Introduction

With the rapid increase in the amount of video material available online, the risk of children being exposed to inappropriate content is increasing. In particular, videos are easily accessible to the general public, e.g., via video on demand (VoD) portals, replay sites from most TV channels or dedicated sharing sites such as Blip or even YouTube, with both professional or semi-professional edited content and user-generated content. We focus here on the former, spanning a variety of content including movies, series, reality shows and documentaries. In particular, movies and series might include material inappropriate for young children, among which violent scenes which can potentially offend children's sensitivity.

To prevent exposure of young children to inappropriate content, ratings are usually adopted in most countries where a global rating is given to a movie or a series by a specific agency, e.g., the Motion Picture Association of America (MPAA) in the USA, the Centre National du Cinéma et de l'image animée (CNC) in France, or the Motion Picture Code of Ethics Committee (Eiga Rinri Kanri Iinkai) in Japan. Based on a set of criteria, the rating is provided globally for a given material, usually setting a minimum age for viewers. As an example, ratings in Japan are chosen according to four categories: all ages admitted (G), some material may be inappropriate for children under the age of 12 (PG-12), forbidden under 15 (R15+) or under 18 (R18+). Global ratings however suffer limitations. Some countries, e.g., the People's Republic of China, do not have a movie rating system as of today. More importantly, ratings are highly dependent on time and culture. Rating rules have considerably changed over time—one can easily find a movie rated by the MPAA as "no one 17 and under admitted" (NC-17) a few years back, now rated as "parental guidance for children under 13" (PG-13)—and are also different from one country to another, with however sex and violence as a constant preoccupation throughout all cultures. Finally, global ratings are poorly adapted to today Internet-based content diffusion, where the number of video distributors has exploded along with the volume of material available, where national barriers almost no longer exist and where not all material goes through rating. These limits also clearly emphasize that the notion of offending material is first and foremost a personal matter, depending on one's culture and sensitivity.

As an answer to the impossibility of rating all material available on the Internet and to the inadequacy of a global rating system to facilitate personalized decisions, content-based video analysis can be used to offer features enabling fast and accurate personal decisions, helping users to decide whether a particular video is acceptable or not, e.g., for their children to watch. With this objective in mind, we focus here on detecting violent scenes in professionally edited content, primarily targeting movies, as a tool for video distribution sites. A typical scenario relying on the automatic detection of violent scenes consists in offering a preview of the most violent segments for people to make their own decisions regarding the appropriateness of the video.

Violent scene detection in video is not a novel subject per se and was already discussed in the literature, especially for video surveillance. But movies significantly differ from video surveillance material. To cite a few differences, movies are highly edited, including special visual and audio effects that are not present in video surveillance, and depict a large spectrum of violence which is seldom seen in real life. Viewpoints are also significantly different in movies, with a single viewpoint at any time which varies frequently, as opposed to potentially multiple viewpoints from fixed or slowly moving cameras in surveillance videos. Developing multimodal approaches to content-based violence detection in movies thus requires specific techniques which have so far received limited attention [12, 13, 26, 27, 29, 30, 35, 38]. In particular, the lack of publicly available datasets and benchmarks has severely limited advances in the field. Apart from the collection and annotation of adequate material, a key issue is that of establishing a clear definition of violence on which human annotators can rely to create a ground truth reference with reduced ambiguity. Because of the multiple facets of violence, no common and generic enough definition for violent events was ever proposed, even when restricting ourselves to physical violence. Examples taken in previous work for the definition of violence (in movies) are: "*a series of human actions accompanied with bleeding*" [3]; "*scenes containing fights, regardless of context and number of people involved*" [6, 31]; "*behavior by persons against persons that intentionally threatens, attempts, or actually inflicts physical harm*" [14]; "*fast paced scenes which contain explosions, gunshots and person-on-person fighting*" [16]. Some definitions even target action scenes, disregarding the distinction between action and violence [4, 39]. As a consequence, available studies rely on different definitions, possibly partial, and hence on different datasets, thus making comparative evaluation impossible. In this paper, we describe the violent scene detection (VSD) corpus[1], a publicly available dataset specifically designed for the development of content-based detection techniques targeting physical violence in movies. We rely on a definition of physical violence that was designed to be generic enough so as to cover a wide variety of events, yet limiting subjectivity in judgments made by annotators. We describe in detail the dataset, which includes the annotation of semantic concepts related to violence, and provide an in-depth analysis of the annotations. The VSD dataset was initially introduced in [7–9] in the framework of the MediaEval benchmark initiative, which serves as a validation framework for the dataset and establishes a state of the art baseline for the task. This paper extends previous publications related to the MediaEval benchmark, providing an in-depth analysis of the annotations, comparing two annotation guidelines and including results from the 2012 evaluation as a report of progress. We believe that a unified and detailed study of the design, annotation, analysis and evaluation of the corpus will contribute to providing widely accepted annotation rules for violence detection, in establishing VSD as a standard resource and in fostering new research directions in the field.

The paper is organized as follows. Section 2 provides a short description of related datasets and benchmarks. Section 3 establishes the definition of violence and gives a detailed overview of the VSD dataset. An analysis of the dataset in the light of its two-year existence and use is given in Section 4. Finally, evaluation of the dataset is reported in Section 5 through an overview of the results of the 2011 and 2012 MediaEval benchmarks.

2 Related datasets and benchmarks

Among the many existing video datasets, none is particularly suited for the task of detecting violent scenes in professionally edited content such as movies. Most datasets, including VSD, are designed for the recognition of targeted concepts, actions or events. On the one hand, some well-known existing datasets include concepts or actions which might be related to violence but do not exhaustively cover these concepts and fail to address violence explicitly. On the other hand, datasets that were made for violence detection are either unadapted for edited content or too small, not considering public availability. We briefly review here the major relevant existing corpora.

[1] http://www.technicolor.com/en/innovation/research-innovation/scientific-data-sharing/violent-scenes-dataset

The TRECVid dataset is probably one of the largest available for multimodal event detection (MED) in videos, with about 1,500 hours for training and 3,700 for testing in 2012 [32]. However, the MED corpus contains user-generated content which significantly differs from edited movies and target the detection of specific events, e.g., *changing a vehicle tire* or *working on a sewing project*, which are irrelevant for violence detection. More interestingly, the TRECVid semantic indexing task provides material for the detection of concepts, some of which related to violence (e.g., Explosion_Fire, Car_Crash). While TRECVid concept detectors can be adapted to movies, training material in TRECVid differs from the type of content targeted here, with short videos (between 10 s and 4 min) in the TRECVid-IACC data, thus calling for more specific data for violence detection in movies. In fact, very few datasets dedicated to event recognition are extracted from edited movies, with the exception of [28] which provides a database dedicated to the recognition of specific actions, e.g., hugging, shaking hands, in movies, containing circa 20.1 hours of video from 69 different movies. While violent actions are not considered, this nevertheless gives an idea of what is freely available nowadays concerning movies.

Violence detection per se has naturally been considered in the domain of video surveillance, e.g., [5, 25, 41], however with content that significantly differs from our target.

In the framework of broadcast videos, detection of fights is developed in [31], with a dataset of 1,000 short sports video clips from the National Hockey League. A larger variety of 400 broadcast videos is used in [6] for classification between fight and non fight, with 200 videos for each category, taken from sport games, music clips or daily life news. Interestingly, videos containing no fights were selected from situations which can be easily mismatched with violent fights: traffic, matches, people hugging, running, etc.

Movie-like material is specifically targeted in a few studies which are closely related to the VSD dataset, relying on datasets that are not publicly available. In [24], a database of 11.4 hours of violent videos coming from 100 violent movies is used for the detection of 5 audio events, namely fights, explosions, gunplay, screams and car-racing. The dataset is divided into a training subset and a test subset of respectively 3.3 hours and 8 hours. Non violent samples were extracted from the TRECVid MED 2010 database, which contains only user generated video clips. In [26, 27], five action movies are used, proposing a test set of about 4.5 hours. In [14], a dataset of 50 video clips, extracted from 10 movies for a total duration of 2.5 hours, 19.4 % of which were annotated as violent, are used. Complete and rather violent movies are used in [16], representing a little less than 8 hours of video.

This quick overview of existing datasets for violence detection shows that there is currently no reference dataset that is large enough, targeting a large range of violent actions and suited for edited content such as movies. Moreover, most datasets are not publicly available. The VSD dataset intends to establish itself as such a reference, proposing a definition for violence which encompasses a large number of situations.

3 Violent scene detection dataset

As a workaround to the main limitations of existing datasets with respect to the targeted use case, we developed the publicly available VSD dataset which is at the core of the implementation of a yearly evaluation within the MediaEval initiative. The violent scene detection task has gained interest over the years and is now a de facto standard. We provide here details on the dataset creation and annotation, while an in-depth analysis of the annotations is reported in the next section.

3.1 Violence definition

In a report from the 1996 Global Consultation on Violence and Health organized by the World Health Organization (WHO), violence is defined as [40]: "*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*". This global definition can be further divided into three main types, namely, self-inflicted, interpersonal, and collective, depending on the number of persons engaged in violence. Finally, depending on the setting and nature of violence — e.g., verbal, physical, sexual, psychological, and deprivation or neglect — each category can be further divided [19]. In the context of movies and television, a report from the French Ministry of Culture and Communication defines violence on TV as follows [18] : "*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 crash with blood visible or earthquake leading to some people's death. We therefore propose an extended definition of violence that includes accidental violence resulting in physical damages.

To deal with the diversity of violence and provide a fairly objective definition so as to ease the annotation process and maintain the dataset consistency, VSD annotations are limited to physical violence according to the following definition, referred to as the 'objective' definition in the sequel: *"physical violence or accident resulting in human injury or pain"* [7]. This definition implies that segments annotated as violent must contain both the violent action and the result of the action. Before converging towards this definition and proceeding with annotation, alternative propositions were confronted with the data, some of which are discussed in Section 4.3. We found that the objective definition was the simplest to apprehend and implement for annotators and left a minimum number of unclear cases, which will be further discussed in Section 4.2.

3.2 Dataset

The choice of movies for the VSD dataset is a crucial issue. Choosing entire movies rather than excerpts or trailers appeared as a natural option given the use case targeted, the scarcity of violence within complete movies being one of the challenges for the violent scene detection task. Regarding the movies to include in the dataset, priority was given to the diversity of the movies genres and of the violence types. The VSD dataset eventually consists of 18 movies, from extremely violent ones (e.g., *Kill Bill* or *Fight Club*) to movies with virtually no violent content (e.g., *the Wizard of Oz*). Movies were also chosen to provide a large palette of violence types and of film editing techniques. For example, war movies, such as *Saving private Ryan*, contain specific gunfights and battle scenes involving lots of people, with a loud and dense audio stream containing numerous special effects. Action movies, such as *the Bourne Identity*, usually contain scenes of fights involving only a few participants, possibly hand to hand. This in turn differs from disaster movies, such as *Armageddon*, showing the destruction of entire cities and huge explosions. Finally, the choice of movies also results from a compromise between two opposite criteria: having a fairly large amount of violent segments for training purposes on the one hand and, on the other hand, having reasonably intermediate violent movies for which personalized parental guidance via previewing the violent scenes makes sense. A few completely non violent movies were also

Table 1 Composition of the VSD dataset, where the last three movies correspond to the MediaEval 2012 test set

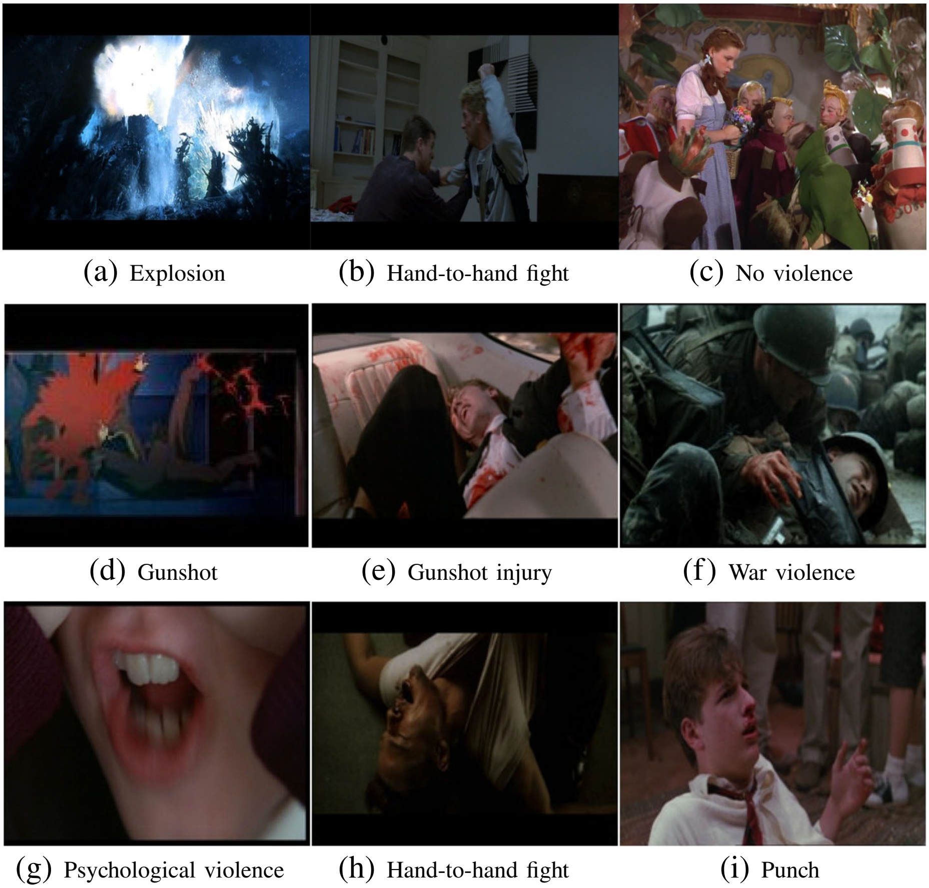
Movie title	Duration	#shots	Length
Armageddon	8,680.1	3,562	2.43
Billy Elliot	6,349.4	1,236	5.13
Eragon	5,985.4	1,663	3.59
Harry Potter 5	7,953.5	1,891	4.20
I am Legend	5,779.9	1,547	3.73
Leon	6,344.5	1,547	4.10
Midnight Express	6,961.0	1,677	4.15
Pirates Carib. 1	8,239.4	2,534	3.25
Reservoir Dogs	5,712.9	856	6.67
Saving Private Ryan	9,750.9	2,494	3.9
The Sixth Sense	6,178.0	963	6.41
The Wicker Man	5,870.4	1,638	3.58
The Bourne Identity	6,816.0	1,995	3.41
Kill Bill	6,370.4	1,597	3.98
The Wizard of Oz	5,859.2	908	6.45
Average	**6,856.7**	**1,740**	**3.93**
Total	**10,2851.1**	**26,108**	–
Dead Poets Society	7,413.2	1,583	4.68
Fight Club	8,005.7	2,335	3.42
Independance Day	8,834.3	2,652	3.33
Average	**8,084.4**	**2,190**	**3.69**
Total	**24,253.2**	**6,570**	–

Average and Total results are indicated in bold for both parts of the dataset

added to the dataset to study the behavior of algorithms on such content which we believe may be different from the non violent part of the rest of the movies. The downside of covering a large spectrum of genres, violence types and editing styles is that only a few representatives of each category are included in the dataset. However, this is a reasonable price to pay for diversity, made possible by the relatively large size of the VSD corpus which is, to the best of our knowledge, the biggest dataset available on entire movies.

The complete list of selected movies is given in Table 1, where 15 movies are identified as the training subset and 3 movies constitute the MediaEval 2012 test set. We tried as much as possible to respect the variety of genres in both subsets. The total duration of the training set is 28h34 and that of the test set is 6h44. Figure 1 shows some sample images taken from the VSD dataset with the corresponding type of violence. Annotations are made publicly available on-line.[2] However, for evident copyright issues, the video content cannot be distributed freely on-line. Instead we distribute the exact references of the DVDs that were

[2] http://www.technicolor.com/en/innovation/research-innovation/scientific-data-sharing/violent-scenes-dataset



(a) Explosion (b) Hand-to-hand fight (c) No violence

(d) Gunshot (e) Gunshot injury (f) War violence

(g) Psychological violence (h) Hand-to-hand fight (i) Punch

Fig. 1 Example images from the VSD dataset

used for annotation and provide strict guidelines to rip the DVD content into an MPEG file so as to maintain synchrony between video frames and annotations. This point also guided the choice of using freely available software such as VirtualDub[3] for the annotation process so that users of the VSD data set could revisit annotations easily without synchronization issues.

Results from automatic shot segmentation are provided for all movies, with a total of 26,108 shots for the training set and of 6,570 shots for the test set, as reported in Table 1. Unsurprisingly, we can observe that the average shot length depends on the movie type and is somewhat correlated with the presence of violence. Movies with little or no violence, or, at least, movies with little action, such as *Billy Elliot* or *The Sixth Sense*, have longer than average shots. It is interesting to note however that the movie *The Sixth Sense* contains violent scenes, but of a psychological rather than physical nature. This shows that action scenes and violent scenes may be related but for some specific types of physical violence only.

[3] http://www.virtualdub.org

3.3 Annotations

The VSD dataset comes with rich annotations going beyond the annotation of violent segments. In addition, a number of key concepts likely to be related to violence were annotated on the training data. Seven visual concepts are considered, namely, the presence of blood, fights, presence of fire, presence of guns, presence of cold weapons, car chases and gory scenes. In addition, three concepts related to the audio modality were are also provided, namely, gunshots, explosions and screams. The choice of such concepts comes from preliminary observations which link the presence of such semantic concepts to violence. A more detailed analysis of the correlation between concepts and violence is given in Section 4.

Violence and key concepts annotation was carried out on the MPEG files extracted from the DVDs as specified in the corpus description and involved a total of 9 different annotators, among which two 'master' annotators were designated to ensure consistency between movies. As annotations are highly time consuming, each concept (i.e., violence, audio concepts and visual concepts) was annotated by a single annotator and was reviewed by at least one 'master' annotator. In case of disagreement between the annotator and the master annotator, final decisions were taken by voting after further discussion between the annotator and the two master annotators. While disagreement were regular (about 10 % of the cases), discussions enabled resolving the conflict in almost all cases. Violence and audio concepts were extracted using the English audio tracks. Violent segments and visual concepts are identified by their starting and ending frames in the MPEG file, with 25 frames per second, while audio concepts are identified by start and end times in seconds, with respect to the beginning of the MPEG file. All labels are in English. Concept labels might be complemented with additional tags to provide further information.

We discuss in turn all the annotated events, detailing the guidelines followed during the annotation process and the difficulties encountered.

3.3.1 Violent segments

Violent segments were annotated following the objective definition of Section 3.1. The boundaries of a segment are given by the start and end frames of the segment, where the start frame indicates the beginning of the action, the ending frame being the last frame in which the result of the action (injury, pain, etc.) is seen. Non continuous violent segments can be found occasionally, with an interruption between the action and its result. If the interruption is less than 2 s (duration chosen empirically), it is ignored and we consider a unique segment which contains the separation. When a longer separation occurs, action and results are considered as separated events and thus are not annotated. This choice is clearly questionable and will be further discussed later, but directly derives from the definition taken which we tried to keep as simple and objective as possible.

In scenes in which two persons or more are fighting hand to hand, only the parts where a punch or a kick actually hits the target were annotated. In other words, intent to strike does not fit the objective definition that we followed for annotation (no resulting pain or injury). The starting point corresponds to the moment when the kick actually starts its way towards the target, while ending point corresponds to when the person hit shows no more pain. For gunfights, the starting point is the moment where the shooter begins to point his gun, or if not visible the moment when the gunshot is heard or seen, and the end point is when the person shot at clearly does not move anymore or disappears from the scene.

3.3.2 Blood

Appearance of blood on screen is annotated, whatever the amount of blood visible. Dried blood was also annotated, as well as some reddish injuries, even if the actual presence of blood in such injuries was unclear. Additional tags—among 1 %, 5 %, 10 %, 25 % and 50 %—indicate the proportion of the screen covered in blood. Tag 1 % means that the surface of blood pixels is roughly between 1 % and 5 % of the screen, etc. One of the difficulties has consisted in appreciating the presence of blood when it is hardly visible (in some dark scenes for example) and estimating the amount of blood turned out very difficult for some annotators.

3.3.3 Fights

Different types of fights were annotated, resulting in four different tags for each segment: 1vs1 (only two people fighting), small (a small group of people, i.e., roughly less than 10), large (a large group of people, i.e., more than 10), distant attack (no physical fight, but corresponds to somebody being shot at or attacked from a distance). Fights between humans and animals were also annotated. Starting and ending frame numbers correspond to the first (resp. last) frame in which the fight begins (resp. ends), e.g., the first frame in which the first blow actually starts to the last frame in which we see the people fighting or the result of the fight (last frame in which people are moving).

3.3.4 Fire

The presence of fire was annotated in all cases, whatever the type and size of the fire. This includes big fires, explosions as well as fire coming out of a gun shooting, candles or cigarette lighters, or even a cigarette or sparks. When the fire does not have the natural color of a fire (i.e., yellowish or orange tone), as in a fire resulting from a magic spell, an additional tag specifies its color, the tag 'multicolor' being used when too many extra colors are visible.

3.3.5 Guns

The presence of any type of guns or assimilated firearms on screen, either fully or partially, even if on a few pixels, is indicated. Guns with bayonets were annotated as guns, even if only the bayonet is seen. No additional tag was added for this concept.

3.3.6 Cold weapons

Similarly to firearms, the presence of any cold weapon (e.g., knives, swords, arrows, halberds, etc.) was annotated. In particular, note that all types of knives were considered, including kitchen knives. Potentially dangerous tools that may be used as weapons, such as scissors and axes, were also annotated.

3.3.7 Car chases

All car chases were annotated at the frame level (i.e., first frame exactly corresponds to the first frame in which the car chase appears and the last frame corresponding to the precise end). Car chases with only one car involved (e.g., car chasing animals) were also considered and annotated.

3.3.8 Gory scenes

The gory scene is a highly subjective concept that we nonetheless decided to annotate. Gory scenes were annotated primarily to indicate graphic images of bloodletting and/or tissue damage. This definition includes horror or war representations. Segments showing undoubtedly disgusting mutants or creatures were also annotated. Additional tags describing the event/scene were added in all cases. The list of such tags is not limited, annotators doing their best at succinctly describing the scene.

3.3.9 Gunshots

Gunshots in the audio tracks were annotated, considering a single segment whenever possible. The Tag 'multiple_actions' was used when several gunshots happen together such that it is impossible to split them into separated segments. Cannon firing were either annotated as gunshots, e.g., in *Pirates of the Caribbean*, or with a tag 'cannon_fire', e.g., in *Saving Private Ryan*, wherever possible. An additional tag 'multiple_actions_cannon_fire' was also used when appropriate. These two last tags are used when cannon explosions can be heard but no gunshot, whereas the tags 'gunshot' and 'multiple_actions' may indicate that cannon fires were possibly heard in addition to gunshots.

3.3.10 Explosions

Explosions were annotated in the audio track following the same principle as for gunshots, with the tags 'explosion' or 'multiple_actions'. Any kind of explosion was annotated, including magic explosions. As a borderline case, when a cannon fires, the cannon fire sound was annotated as a gunshot, but explosions resulting from shells or cannonballs in cannon fire were considered as explosions.

3.3.11 Screams

Anything from non verbal screams to what was called 'effort noise' was annotated as screams, as long the noise came from a human or a humanoid (e.g., mutant in *I Am Legend*). Verbal screams, i.e., screams in which one can recognize words, were considered as speech and therefore not annotated. For economical reasons, scream annotations are provided for nine movies only.

As in the case of gunshots and explosions, two tags were considered for screams, namely 'scream' and 'multiple_actions'. Effort noises were annotated using tags 'scream_effort', or 'multiple_actions_scream_effort'. Animal screams were not annotated.

4 Analysis of the dataset

The reasonable size of the VSD dataset offers the possibility of a meaningful qualitative and quantitative analysis regarding violence annotation and key audio and visual concepts. We first report the statistics and study the correlation between violence and key concepts. We further discuss borderline cases in the annotation of violence, leading to a comparison with subjective annotations of violence.

4.1 Analysis of the violence and key concepts annotations

4.1.1 Concept occurrence frequencies

The rarity of the events considered is one of the key characteristics of the VSD dataset, where unbalanced data with few positive examples is bound to yield low precision values at detection time. The repartition of the violence or concept occurrences varies significantly from one movie to another as well as from one concept to another, as illustrated in Table 2 for violence and in Table 3 for the audio and visual concepts. The proportion of violent segments goes from 0.76 % of the movie duration in *Dead Poets Society* to 10.17 % in *Kill Bill*, with an average over the movies of approximately 5 %. Audio and visual concepts also appear with scarce occurrences. Depending on the concept, the proportion goes from 0.2 % to 11.9 %, with a large majority around 4 %. Even when limiting the computation of the statistics to movies which actually contain the concept considered (Average-2 values in Table 3), proportions range between 1.3 % and 14.8 %, with an average of about 4 %. This is comparable with the TRECVid MED 2011 data in which the event proportions were in the range [2.5 %, 5.8 %].

Table 2 Comparison between the annotations of violent scenes at the shot level and at the segment level

	Shot level			Segment level		
Movie	#shots	dur.	prop.	#segs	dur.	prop.
Armageddon	392	882.5	10.2	55	509.6	5.9
Billy Elliot	52	326.2	5.1	35	106.4	1.7
Eragon	276	659.7	11.0	116	361.5	6.0
Harry Potter 5	240	774.2	9.7	96	312.9	3.9
I am Legend	306	720.1	12.4	103	442.5	7.6
Leon	112	273.1	4.3	30	170.4	2.7
Midnight Express	187	506.8	7.3	74	316.4	4.5
Pirates Carib. 1	316	931.4	11.3	150	472.3	5.7
Reservoir Dogs	106	659.9	11.5	29	282.0	4.9
Saving Private Ryan	469	1,259.4	12.9	115	829.4	8.5
The Sixth Sense	27	82.9	1.3	6	63.5	1.0
The Wicker Man	110	491.0	8.4	32	172.4	2.9
The Bourne Identity	184	519.0	7.6	98	231.9	3.4
Kill Bill	383	1113.3	17.5	142	647.9	10.2
The Wizard of Oz	46	323.4	5.5	23	107.7	1.8
Total	**3,206**	**9,522.80**	**9.2**	**1,104**	**5,027.00**	**4.9**
		2h38			**1h23**	
Dead Poets Society	34	111.8	1.5	15	55.3	0.8
Fight Club	310	1082.2	13.5	116	608.8	7.6
Independance Day	371	876.8	9.9	82	565.2	6.4
Total	**715**	**2,070.76**	**8.5**	**213**	**1,229.40**	**5.1**
		34 min			**20 min**	

Total durations are indicated in bold

Table 3 Audio and visual concept annotations on the training set: Percentage of the movie duration in which the concept appears for resp. blood (Bl), fights (Fg), fire (Fr), guns (Gu), cold weapons (We), car chases (Ca), gory scenes (Go), gunshots (Sh), explosions (Ex) and screams (Sc). Average-1 is the average over all movies while Average-2 is computed only over movies containing the event

Movie	Bl	Fg	Fr	Gu	We	Ca	Go	Sh	Ex	Sc
Armageddon	0.8	2.9	9.3	3.8	0.04	0.1	0	0.4	5.7	–
Billy Elliot	0.2	1.9	1	–	0.2	–	–	–	–	4.8
Eragon	5	10.5	21.1	–	13.4	–	2.2	–	0.4	6.1
Harry Potter 5	4.3	4.8	14.2	–	2.3	–	–	–	1.7	2.8
I am Legend	8.7	5.5	1.9	12.9	3.3	1.1	9.4	0.7	0.4	8
Leon	12	3.4	0.7	20.1	1.6	–	0.01	1.3	0.2	1
Midnight Express	1.9	5.1	3.6	6.7	0.4	–	0.1	0.2	–	–
Pirates Carib. 1	0.6	9.3	17.9	20	25.5	–	4.8	1.8	0.7	-
Reservoir Dogs	36.7	4	0.2	19	1.8	–	21.5	0.7	–	4.3
Saving Private Ryan	21.4	10.8	11.6	53.9	18.6	–	8	25.6	12.6	-
The Sixth Sense	0.9	0.1	1.8	0.8	4.3	–	0.1	0.03	–	1.3
The Wicker Man	0.6	0.4	4.7	6.2	1.8	–	0.02	0.2	0.2	–
The Bourne Identity	3.2	2.6	0.4	6.1	2.2	2.9	–	0.4	0.08	2.1
Kill Bill	30.4	11.5	2.2	2.6	31.5	–	6	0.3	0.03	–
The Wizard of Oz	–	1.2	6.7	7.4	33.3	–	–	–	1	4.9
Average-1	**8.4**	**5.2**	**7**	**11.9**	**9.6**	**0.2**	**3.4**	**2.8**	**2**	**3.8**
Average-2	**8.9**	**5.2**	**7**	**14.8**	**9.6**	**1.3**	**5.2**	**3.8**	**2.6**	**3.8**

Average percentages are indicated in bold

4.1.2 Violence vs. audio and visual concepts

Figure 2 depicts the occurrences of violent segments and of each of the audio and visual concepts for the movie *I Am Legend*, as an illustration of the relationship between concepts and violence. Even though correlation differs from one movie to another, Fig. 2 nevertheless shows that the audio and visual concepts annotated are to some extent informative with respect to violence. Screams, fights, fire and explosions appear to be highly correlated with violent segments. Blood and gory scenes are also related to violence, but to a lesser extent: While violence often comes with blood, the presence of blood does not necessarily imply violence as defined in this paper (see, e.g., *Reservoir Dogs* where blood is very present but violence rarely shown). The same happens for gory scenes, specifically in *I Am Legend* where gory mutants (annotated as gory scenes) often fight with humans. In this specific movie, other concepts, on the contrary, do not happen to be correlated with violence, such as the presence of cold arms and firearms. This is because the hero constantly holds a rifle. On the contrary, the presence of firearms will be highly correlated with violent events in other movies, e.g., *Leon* or *The Bourne Identity* where weapons are mostly shown during fights.

A more detailed analysis of the entire dataset is summarized in Table 4 where we report for each concept two quantities: the proportion of time an occurrence of the concept overlaps with violence (first line) and, conversely, the proportion of violence time overlapping with the concept (second line). Concepts such as *fights* and *screams* exhibit significant overlap in both cases, meaning that these concepts are highly correlated with violence, whatever the type of violence and the movie genre. Concepts such as *gory scenes*, *gunshots* and *explosions* show high proportions when compared to the total duration of the concepts, but low proportion compared to the violence duration. Hence, having a gory scene, a gunshot or an explosion in a movie quite certainly corresponds to a violent event, but violent events with gory scenes and explosions do not correspond to the majority of violent events. Proportions for *blood* (roughly 15 %) validate the conclusion drawn from Fig. 2: Blood is correlated to a certain extent with violent events, but one may also have some violent events without blood,

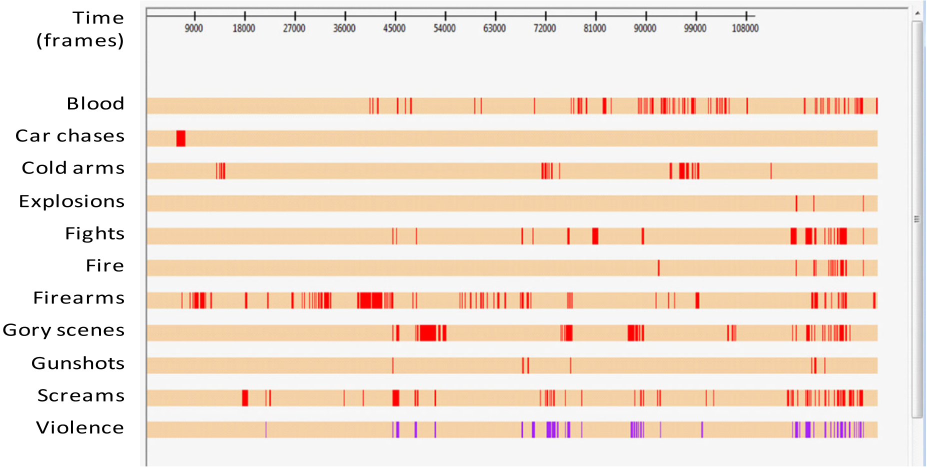


Fig. 2 Correlation between violence (*purple timeline*) and the audio and video concepts (*red timelines*). [movie: *I Am Legend*]

Table 4 Correlation between concepts and violence. For each movie, we report the recall (in %, rounded to the nearest integer) of concepts with respect to violent segments (first row) and the recall of violence with respect to concepts (second row), for resp. blood (Bl), fights (Fg), fire (Fr), guns (Gu), cold weapons (We), car chases (Ca), gory scenes (Go), gunshots (Sh), explosions (Ex) and screams (Sc). Averages are computed only on movies containing the concept considered

Movie	Bl	Fg	Fr	Gu	We	Ca	Go	Sh	Ex	Sc
Armageddon	8	34	26	7	0	0	–	4	38	–
	1	17	42	5	0	0	–	0	37	–
Billy Elliot	22	40	0	–	0	–	–	–	–	13
	3	46	0	–	0	–	–	–	–	39
Eragon	5	39	12	–	13	–	36	–	40	38
	4	68	42	–	30	–	13	–	3	40
Harry Potter 5	15	44	5	–	0	–	–	–	21	37
	17	53	19	–	0	–	–	–	10	28
I am Legend	12	50	46	6	12	0	32	30	77	47
	14	37	12	11	5	0	39	3	5	51
Leon	8	53	42	6	0	–	100	34	89	16
	35	68	12	46	0	–	1	17	7	7
Midnight Express	27	51	1	7	11	–	81	0	–	–
	12	58	1	10	1	–	2	0	–	–
Pirates Carib. 1	10	40	11	6	8	–	23	26	39	–
	1	65	35	22	35	–	19	9	5	–
Reservoir Dogs	8	28	23	9	15	–	14	83	–	27
	58	23	1	34	5	–	63	13	–	26
Saving Private Ryan	13	51	18	9	8	–	44	17	13	–
	33	65	24	60	19	–	42	51	20	-
The Sixth Sense	1	38	0	6	0	–	44	53	–	25
	1	4	0	5	0	–	0	2	–	33
The Wicker Man	0	42	13	2	0	–	0	0	87	–
	0	6	21	4	0	–	0	0	6	-
The Bourne Identity	22	77	10	14	5	12	–	29	0	24
	21	59	1	25	3	10	–	9	0	17
Kill Bill	22	68	2	4	18	–	46	25	0	–
	66	78	0	1	55	–	27	1	0	–
The Wizard of Oz	–	43	10	0	0	–	–	–	1	12
	–	29	37	2	0	–	–	–	0	34
Average	**12**	**47**	**15**	**6**	**6**	**4**	**42**	**27**	**37**	**27**
	18	**45**	**16**	**15**	**13**	**1**	**14**	**7**	**6**	**30**

Average recalls are indicated in bold

and some frames showing blood which do not belong to violent events. From this table, one may also infer that *guns* and *cold weapons* are not systematically correlated with violence in the sense that seeing some guns or weapons does not necessarily imply that a violent event is happening. On the contrary, roughly 15 % of violent events involve weapons. Finally, no real conclusion can be drawn for car chases due to the small amount of examples in the database.

4.2 Borderline cases

Even with the objective violence definition of Section 3.1 which leaves limited interpretation possibilities, borderline or surprising cases still remain. We discuss here in turn several such cases that we encountered when annotating the dataset and justify the choices that were made.

When strictly sticking to *"physical violence or accident resulting in human injury or pain"*, events depicting an intent to kill, in which one sees somebody trying unsuccessfully to harm or kill somebody else with clear intentions, are considered as non violent because of the absence of actual pain or injuries. Such events are therefore not annotated as violent in the VSD dataset even though this choice is controversial with respect to the use case considered.

Detecting violent segments in videos is multimodal by essence due to the nature of the material processed. However, violent events are not always present on all modalities. All events, whether visible in a single modality or on several modalities, were annotated. Hence, 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 deemed violent.

Actions resulting in pain with no intent to be violent or with the aim of helping rather than harming—e.g., segments showing surgery without anesthetics in *Saving private Ryan*—fit the objective definition and were therefore annotated.

The reference to "human" in the definition of violence is bound to interpretation when one tries to infer the limit of being a human. Are humanoids still human beings? What about mutants? We decided to consider all human-shaped characters as humans and to include violent actions resulting in pain or injuries on such characters in the annotations. This includes for example injuries applied to skeletons or parts of skeletons (e.g., moving skeleton hand in *Pirates of the Caribbeans*) or manikins (e.g., hero shooting at a manikin in *I am legend*).

A borderline case keenly discussed among annotators is that of the destruction of a whole city or the explosion of a moving tank. Technically speaking, there is no proof of death or injury in such a scene, 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 deemed violent.

4.3 Comparison with subjective violence annotation

While the interest of the objective definition is to provide a simple and tractable perimeter clearly designating which events to annotate, the objective definition leads to decisions that are not well adapted to the use case of personalized parental guidance via previewing. Some movies show the result of a violent action but not the violent act itself, e.g., shots in which one can see a dead body with a lot of injuries and blood. As already mentioned, these scenes do not match our objective definition of violence and are therefore not annotated as violent. Clearly, previewing such scenes would help users in making their decision as whether the scenes are suitable for their children to view or not. On the opposite, a character slapping another one in the face is considered as a violent action according to the task definition, although one would certainly consider that such scenes are not hurtful for children to see.

We investigate here on a limited scale how subjective annotations of violence better targeting the parental guidance scenario would behave and compare with the objective annotations of violence taken so far. A subset of the VSD dataset was annotated with the following guideline given to annotators: mark as violent all events *"one would not let an 8 years old child see in a movie because they contain physical violence"*. Note that once again, this new definition is only targeting physical violence. To study inter-annotator variations inherent to such loose guidelines, we resorted to multiple annotators, involving three distinct annotators with different background and sensitivity: The first two annotators are women with children of the requested age or older; The third one is a male student with no children but with an 8 year old sister. Note however that the three annotators had been previously involved in the annotation of the VSD dataset according to the objective definition of Section 3.1.

Figure 3 shows a timeline comparing annotations with, respectively, the objective guidelines and the subjective ones for a single annotator. The example is given on *The Sixth Sense* which was chosen as it contains quite a few examples of violent events according to the objective definition and several examples of scenes in which one sees the result of a violent action without seeing the action. The differences between the objective and the subjective definitions are clearly visible on such scenes, as pointed out in the figure.

We also studied agreement between annotators to get a feeling of whether such a definition could be considered for future annotations or not. Regardless of annotation, this comparison also provides some insight on the use case, hinting at what one would expect to preview. Inter annotator agreement was evaluated based on a single movie, *Billy Elliot*. The three annotations are shown in Fig. 4, along with the annotation resulting from the objective definition of violence. Surprisingly, they are relatively coherent. Apart from a few scenes, the main differences are in the duration of the violent scenes which differ from one annotator to another, and which are globally far much longer for the subjective segments than for the objectives ones. Interestingly, the fact that the third annotator is a young adult without children does not seem to have impacted his annotation towards limiting himself to highly violent segments. On the contrary, a more detailed study of his annotated segments tends to show that he annotated even more segments than the two other annotators. For instance, segments corresponding to the objective definition (e.g., just after 4,500 seconds and last segment on the first annotation line in Fig. 4) were not indicated by the two first annotators. They were however retained by the third annotator, although the two segments correspond to children teasing each other without real violence (*kicking with pillows, with a ruler*).

Obtaining annotations better suited to the parental selection use case considered from a subjective and relatively fuzzy definition such as the one we are exploring in the preliminary

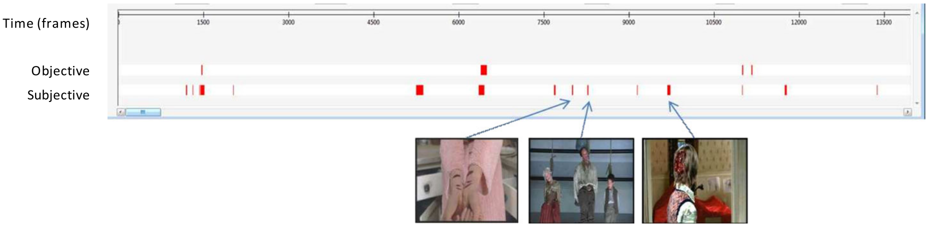


Fig. 3 Comparison between the objective (*first line*) and subjective (*second line*) definitions of physical violence. Segments are added with the subjective definition which do not appear in the objective one (examples are indicated with keyframes on the timeline). [movie: *The Sixth Sense*]

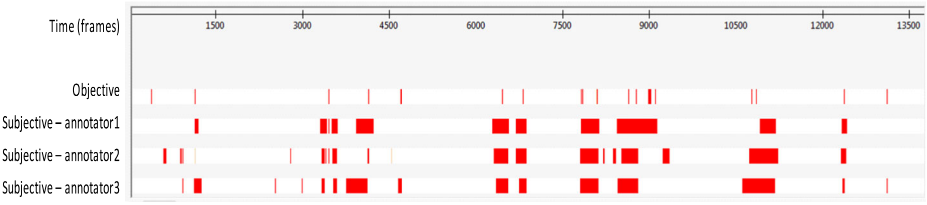


Fig. 4 Comparison of annotations from different annotators for the subjective definition of violence. The first line gives annotations from the objective definition. Subsequent lines are from different annotators with the subjective definition: First and second annotators are women with children (8 years old or older), third annotator is a student with no child but a little sister of the right age. [movie: *Billy Elliot*]

experiments reported here therefore seems feasible with multiple annotators. Based on these very first experiments, factors such as sex, age and parenthood (with/without children) of annotators seem surprisingly to have limited impact on the annotations. However, these conclusions are to be consolidated on a larger scale, studying the impact of other factors such as cultural origin and educational background.

5 The MediaEval affect task

The VSD dataset results from 2 years running the MediaEval Affect task for the detection of violent scenes in movies, as part of the MediaEval benchmark initiative. While the dataset is a result of the evaluation effort, evaluation in turns serves as a validation for the dataset and provides state-of-the-art baseline performance to whoever is interested in using the dataset. We describe the Affect task and provide an overview of the results from the 2011 and 2012 evaluations.

5.1 Task description

MediaEval is a yearly ongoing effort dedicated to evaluating new algorithms for multimedia access and retrieval, emphasizing multimodality in a broad sense: audio, visual content, tags, users, etc. Social and human aspects of multimedia are also relevant features of MediaEval with tasks rooted in industrial use cases targeting users. MediaEval emphasizes collaborative work between participants towards learning together and building even better systems.

In this context, the Affect task for Violent Scene Detection in Movies has been running since 2011, targeting researchers in the areas of event detection, multimedia affect and multimedia content analysis. With the Technicolor use case in mind, as described in the introduction, participants are requested to automatically detect portions of movies depicting violence as defined in Section 3.1, and are strongly encouraged to deploy multimodal approaches. All material available on the DVDs can be used, including subtitles and multiple audio tracks. Use of any other data (e.g., synopsis from the Internet) is prohibited.

The Affect task implements two run conditions. The primary run requires participants to classify each shot as violent or not, using the shot segmentation given with the dataset. An optional secondary run requires participants to provide violent segments with their boundaries, independently of the shot segmentation provided. In all cases, a confidence value indicating the confidence of the decision was requested for each shot/segment.

5.2 Evaluation metrics

Different official metrics were proposed in 2011 and 2012. In 2011, system comparison was based on a cost function (referred to as the MediaEval cost in all that follows), weighting false alarms (FA) and missed detections (MI), according to

$$\text{MediaEvalCost} = C_{\text{fa}} P_{\text{fa}} + C_{\text{miss}} P_{\text{miss}}, \tag{1}$$

where the costs $C_{\text{fa}} = 1$ and $C_{\text{miss}} = 10$ were arbitrarily defined to reflect both the prior probability of the situation and the cost of making an error. P_{fa} and P_{miss} are respectively the FA and MI 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.

Although the MediaEval cost function was designed to reflect the use case in which the cost of missing one violent segment should be higher than the cost of having some false alarms, it appeared strongly biased towards low miss detection rate. In 2011, results showed that it was difficult for most of the submissions to reach cost values lower than 1, which corresponds to a naive system classifying all shots as violent. In 2012, the computation of the mean average precision value over the first 100 top-ranked violent shots (MAP@100) was chosen as the official measure. This measure is also better adapted to a search-related use case in which a user will be proposed a limited number of violent scenes for visualization. For the task, the MAP@100 is computed by taking the mean of the average precision scores over the first 100 top-ranked violent shots for all test movies.

For a more complete description of the task and evaluation metrics, one can refer to [7–9].

5.3 Results

Figure 5 shows the evolution of the participation to the Affect task for its two-year life, indicating the number of participants in each phase of the task selection process—preliminary interest survey, registrations, submissions and workshop attendance—and according to different factors—previous experience in violence detection, countries, teams other than organizers. Some additional details concerning the participants are proposed in Table 5. In 2011, the Affect task ran as a pilot with 6 registrations, among which 2 were from the organizing teams. This pilot evaluation nevertheless demonstrated the interest of the research community for the topic. The variety of participants' origins (four different countries) is also worth considering. The 6 registered teams sent 29 submissions in total. The interest was confirmed in 2012, with 11 teams registered for participation. Among them, three teams only did not cross the final line, leading to a grand total of 8 teams sending 36 submissions.

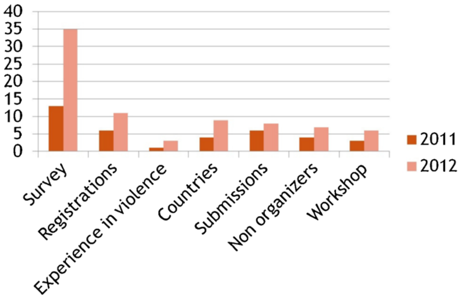


Fig. 5 Evolution of the participation in the task between 2011 and 2012

Table 5 Description of the participants to the 2011 and 2012 benchmarks

Team	Joint	Labs	#subm.
2011 benchmark			
DYNI (FR)	no	LSIS, Univ. of Toulon	2
LIG (FR)	no	Laboratoire d'Informatique de Grenoble	1
NII (JP)	no	Video Proc. Lab, National Institute of Informatics	6
TEC* (FR)	yes	Technicolor	12
		+ Irisa/Inria	
TUB (DE)	no	DAI Lab, Technische Univ. Berlin	3
UNIGE* (CH)	no	CVML Lab, Univ. of Geneva	5
Total			**29**
2012 benchmark			
ARF (RO+AT)	yes	Univ. Polytehnica of Bucharest	2
		+ Dep. of Computational Perception, Johannes Kepler Univ.	
		+ Austrian Research Inst. for Artificial Intelligence	
DYNI (FR)	no	LSIS, Univ. of Toulon	5
LIG (FR)	no	Laboratoire d'Informatique de Grenoble	4
NII (JP)	no	Video Proc. Lab, National Institute of Informatics	5
SHK (CN)	yes	School of Computer Science, Fudan Univ.	5
		+ City Univ. of HongKong	
TEC* (FR+UK)	yes	Technicolor	5
		+ Irisa/Inria	5
		+ Imperial College of London	
TUB (DE)	no	DAI Lab, Technische Univ. Berlin	5
TUM (DE)	no	IHMCT, Technische Univ. München	5
Total			**36**

Total submissions per year are indicated in bold

Moreover, three teams already had an experience in detecting violent scenes on their own, whereas this was only the case for one team in 2011. Once again, the wide geographic coverage area should be highlighted: the 11 teams, some of which being a consortium of labs, were coming from 9 different countries all over the world, leading to the conclusion that the issue is of interest all over the world.

Official results are reported in Table 6. For the sake of comparison between the 2011 and the 2012 results, MAP@100 values are also reported for 2011. Note however that MAP@100 was not the official evaluation metric in 2011, with systems biased towards very low miss detection rates, and therefore not optimized for high values of MAP@100. MAP@20 values are also reported in Table 6 as a complementary metric. In 2012, the best MAP@100 value was over 65 % and two additional teams reached values over 60 %. In comparison to the 2011 results, for which the best MAP@100 value is just over 40 %, improvements have been achieved and the current results show that detection of violence in movies is no utopia but remains an open research axis. Results from the 2013 evaluation, not reported in the paper, indicate continuing progress as the task is better understood and as the amount of consistent training data increases.

Table 6 Official results of the 2011 and 2012 Affect task evaluation at MediaEval

Team	MAP@20	MAP@100	Medcost
2011 benchmark			
DYNI	13.81 *(31.22)*	18.33 *(19.07)*	6.46 *(7.57)*
LIG	23.87 *(23.87)*	18.01 *(18.01)*	7.93 *(7.93)*
NII	40.73 *(33.14)*	24.78 *(27.71)*	1 *(1)*
TEC*	33.33 *(44.94)*	21.89 *(40.58)*	0.76 (0.89)
TUB	4.69 *(4.69)*	14.29 *(14.29)*	1.26 *(1.26)*
UNIGE*	29.28 *(29.28)*	24.57 *(24.57)*	2.00 *(2.83)*
2012 benchmark			
ARF	70.08	65.05	3.56
DYNI	0	12.44	7.96
LIG	28.64	31.37	4.16
NII	40.07	30.82	1.28
SHK	73.6	62.38	5.52
TEC*	66.89	61.82	3.56
TUB	35.92	18.53	4.2
TUM	50.42	48.43	7.83

In 2011, we report in plain figures results from the best run according to the MediaEval cost (Medcost) and indicate in parentheses results corresponding to the best run according to the mean average precisions at the first 20 (MAP@20) and 100 top-ranked violent shots (MAP@100)

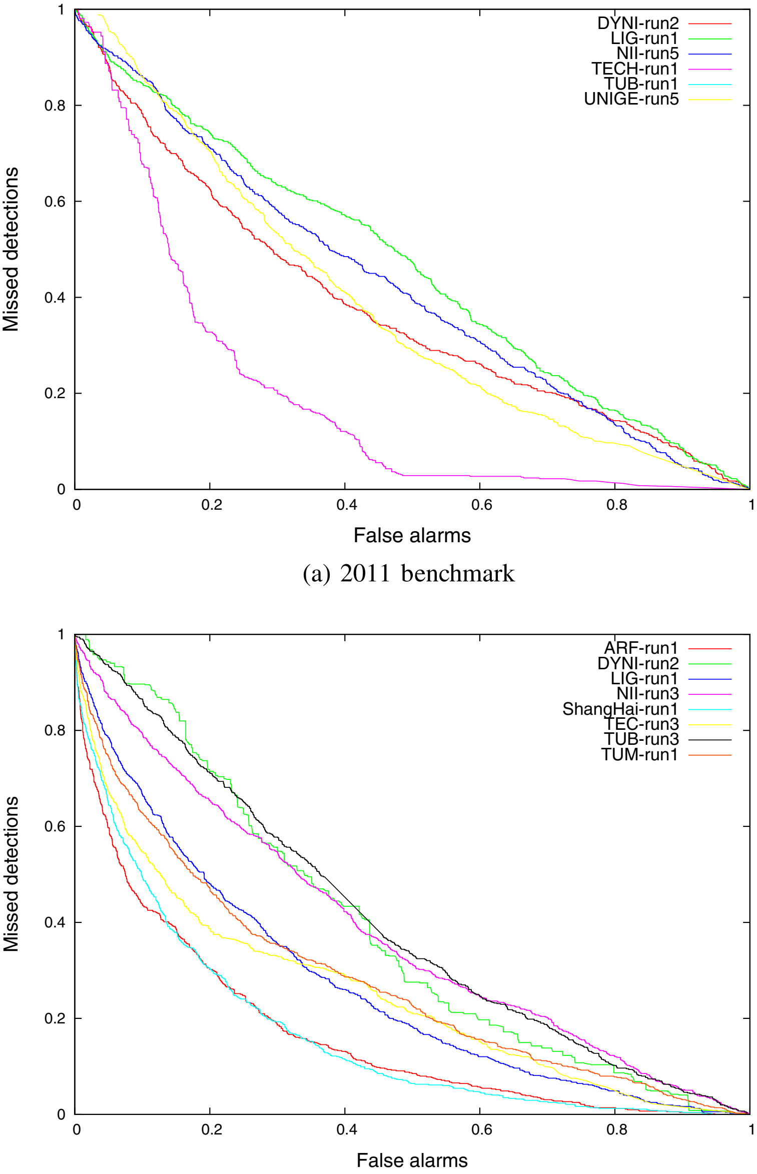
Team names indicated with a star correspond to the task organizers

Detection error trade-off curves, obtained from the confidence values provided by participants, are given in Fig. 6a and b. 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 about 25 %, in 2012, at least two participants have similar results and three more additional teams have decent results.

Interestingly, results are in the same order of magnitude as those obtained in TRECVid MED 2011: Pmiss = 20 % and Pfa = 1 % for the easiest event to detect and Pmiss = 35 % and Pfa = 3 % for the most difficult event. For the Affect task, missed detection rates are equivalent and false alarm rates are higher, but violence remains a concept that is highly subjective with respect to TRECVid concepts, and therefore potentially more difficult to detect.

5.4 Discussion

A detailed description of the systems and the analysis of the results is out of the scope of this paper and can be found in [22, 23]. However, a few general conclusions can be drawn from two years of evaluation. Participants mainly used classic features and classifiers to violence detection. Various submissions [1, 2, 10, 11, 15, 17, 33, 34, 37] tend to demonstrate that multimodality helps. This last point is however more a global feeling than a scientifically proven conclusion as all systems differed by more than only their monomodal or multimodal features. Several submissions made use of intermediate concept detection with key concepts



(a) 2011 benchmark

(b) 2012 benchmark

Fig. 6 Detection error trade-off curves for all participants in 2011 and 2012

related to violence such as those annotated in the VSD dataset [10, 17, 20, 37]. Using a two-step classification system (i.e., infer concepts from data and violence from concepts) seems to be rewarding, with the best 2012 performance obtained by a such system.

It was pointed out during the two evaluations that automatic shot segmentation biases the results of the shot classification run. Statistics in Table 2 show that shot segmentation may significantly affect results of the violence detection task. As a reminder, all shots intersecting a reference violent segment were annotated as violent. With such a definition, the total duration of violent shots is much higher than that of violent segments, by a factor of up to 3. This is of considerable size, especially since the added duration actually corresponds to non violence as a complete shot tagged as violent might depict violence only for a small fraction of its duration. Globally, for the learning data set, violence covers 4.88 % of the training data duration. When considering shots, the proportion goes up to 9.25 %. The difference is smaller on the test data but still remains sizable, with around 40 % of the violent shot duration corresponding to non violent content. This fact certainly penalized the systems during the 2011 and 2012 benchmarks which relied only on keyframes automatically extracted from the shots for classification [21, 36].

In addition, though automatic shot segmentation gives overall good results, segmentation errors inevitably remain. We have observed over-segmentation when objects in the foreground have important movements, due to the lack of motion estimation in the shot boundary detector. This results in very short segments of only a few frames which in turn lower the reliability of features such as shot duration for classification.

In spite of all these drawbacks, few participants attempted the optional violent segment detection run which does not necessarily make use of shot segmentation. We believe that several reasons explain this state of fact, among which the young age of the task which leads participants to prefer the simpler classification task to find relevant features and classifiers and the difficulty of segmenting movie data properly, in particular with respect to violence.

6 Conclusion and perspectives

The Violent Scene Detection dataset has been designed for research in the broad field of event detection in videos, focusing on physical violence in Hollywood movies. Design issues were for a great part governed by a parental control use case where an easy access to the most violent scenes of a movie is offered to parents for them to decide whether the movie is suitable for their children or not. However, we believe that the definition of physical violence adopted here is generic enough to include a wide diversity of violent event types, while staying as objective as possible. We are therefore convinced that the resulting dataset is useful in a number of scenarios and will foster research activities in computer vision and multimedia content analysis. Moreover, to allow common and consistent evaluations in the field, the VSD dataset is made freely available. We have presented in this paper a thorough analysis of the dataset that we hope will serve for future corpus development, for a better understanding of the relation between audio and visual concepts on the one hand and perception of violence on the other hand, and, more generally, for advances in the field of violence detection in multimedia material.

Acknowledgments 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.

References

1. Acar E, Albayrak S (2012) Dai lab at mediaeval 2012 affect task: the detection of violent scenes using affective features. In: MediaEval 2012, multimedia benchmark workshop
2. Acar E, Spiegel S, Albayrak S (2011) Mediaeval 2011 affect task: violent scene detection combining audio and visual features with svm. In: MediaEval 2011, multimedia benchmark workshop

3. Chen L-H, Hsu H-W, Wang L-Y, Su C-W (2011) Violence detection in movies. In: 2011 8th international conference on computer graphics, imaging and visualization (CGIV), pp 119–124
4. Chen L-H, Su C-W, Weng C-F, Liao H-YM (2009) Action scene detection with support vector machines. J Multimed 4:248–253
5. Chen Y, Zhang L, Lin B, Xu Y, Ren X (2011) Fighting detection based on optical flow context histogram. In: Second international conference on innovations in Bio-inspired computing and applications (IBICA), 2011, pp 95–98
6. de Souza FDM, Chavez GC, do Valle Jr EA, de Araujo AA (2010) Violence detection in video using spatio-temporal features. In: Proceedings of the 2010 23rd SIBGRAPI conference on graphics, patterns and images. IEEE Computer Society, Washington, DC, pp 224–230
7. Demarty C-H, Penet C, Gravier G, Soleymani M (2011) The mediaeval 2011 affect task: violent scenes detection in hollywood movies. In: MediaEval 2011, multimedia benchmark workshop, CEUR workshop proceedings, vol 807. CEUR-WS.org
8. Demarty C-H, Penet C, Gravier G, Soleymani M (2011) The MediaEval 2012 affect task: violent scenes detection. In: MediaEval 2012 workshop, vol 927, Pisa, Italy, 4–5 October 2012. ceur-ws.org.
9. Demarty C-H, Penet C, Gravier G, Soleymani M (2012) A benchmarking campaign for the multi-modal detection of violent scenes in movies. In: Springer, editor, ECCV 2012 workshop on IFCVCR, pp 416–425
10. Derbas N, Thollard F, Safadi B, Quénot G (2012) Lig at mediaeval 2012 affect task: use of a generic method. In: MediaEval 2012, multimedia benchmark workshop
11. Eyben F, Weninger F, Lehment N, Rigoll G, Schuller B (2012) Violent scenes detection with large, brute-forced acoustic and visual feature sets. In: MediaEval 2012 multimedia benchmark workshop
12. Giannakopoulos T, Kosmopoulos DI, Aristidou A, Theodoridis S (2006) Violence content classification using audio features. In: Proceedings of the 4th helenic conference on artificial intelligence, pp 502–507
13. Giannakopoulos T, Kosmopoulos DI, Aristidou A, Theodoridis S (2007) A multi-class audio classification method with respect to violent content in movies using Bayesian networks. In: Proceedings of the 9th IEEE workshop on multimedia signal processing, pp 90–93
14. Giannakopoulos T, Makris A, Kosmopoulos D, Perantonis S, Theodoridis S (2010) Audio-visual fusion for detecting violent scenes in videos. In: Konstantopoulos S et al (eds) Artificial intelligence: theories, models and applications, LNCS, vol 6040. Springer, pp 91–100
15. Gninkoun G, Soleymani M (2011) Automatic violence scenes detection: a multi-modal approach. In: MediaEval 2011, multimedia benchmark workshop
16. Gong Y, Wang W, Jiang S, Huang Q, Gao W (2008) Detecting violent scenes in movies by auditory and visual cues. In: Huang Y-M et al (eds) Advances in multimedia information processing - PCM 2008, LNCS, vol 5353. Springer, pp 317–326
17. Jiang Y-G, Dai Q, Tan CC, Xue X, Ngo C-W (2012) The shanghai-hongkong team at mediaeval 2012: violent scene detection using trajectory-based features. In: MediaEval 2012, multimedia benchmark workshop
18. Kriegel B (2003) 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
19. Krug EG, Mercy JA, Dahlberg LL, Zwi AB (2002) The world report on violence and health. Lancet 360(9339):1083–1088
20. Lam V, Le D-D, Le S-P, Satoh S, Duong DA (2012) Nii, Japan at Mediaeval 2012 violent scenes detection affect task. In: MediaEval 2011, multimedia benchmark workshop
21. Lam V, Le D-D, Satoh S, Duong DA (2011) Nii, Japan at Mediaeval 2011 violent scenes detection task. In: MediaEval 2011, multimedia benchmark workshop
22. Larson M, Rae A, Demarty C-H, Koer C, Metze F, Troncy R, Mezaris V, Jones GJF (eds) (2011) Working notes proceedings of the MediaEval 2011 workshop, Pisa, Italy, 1–2 September 2011, CEUR workshop proceedings, vol 807. CEUR-WS.org
23. Larson M, Schmiedeke S, Kelm P, Rae A, Mezaris V, Piatrik T, Soleymani M, Metze F, Jones GJF (eds) (2012) Working notes proceedings of the MediaEval 2012 workshop, Pisa, Italy, 4–5 October 2012, CEUR workshop proceedings, vol 927. CEUR-WS.org
24. Li L (2012) A novel violent videos classification scheme based on the bag of audio words features. In: 2012 9th international conference on information technology: new generations (ITNG), pp 7–13
25. Lin W, Sun M-T, Poovendran R, Zhang Z (2010) Group event detection with a varying number of group members for video surveillance. IEEE Trans Circ Syst Video Technol 20(8):1057–1067
26. Lin J, Sun Y, Wang W (2010) Violence detection in movies with auditory and visual cues. In: Proceedings of the international conference on computational intelligence and security, pp 561 –565

27. Lin J, Wang W (2009) Weakly-supervised violence detection in movies with audio and video based co-training. In: Proceedings of the 10th pacific-rim conference on multimedia, pp 930–935
28. Marszałek M, Laptev I, Schmid C (2009) Actions in context. In: IEEE conference on computer vision & pattern recognition
29. Moncrieff S, Dorai C, Venkatesh S (2001) Affect computing in film through sound energy dynamics. In: Proceedings of the ACM international conference on multimedia, pp 525–527
30. Moncrieff S, Dorai C, Venkatesh S (2001) Detecting indexical signs in film audio for scene interpretation. In: Proceedings of the IEEE internation conference on multimedia and expo. pp 989–992
31. Nievas EB, Suarez OD, García G B, Sukthankar R (2011) Violence detection in video using computer vision techniques. In: Proceedings of the 14th international conference on computer analysis of images and patterns - vol Part II, CAIP'11. Springer, Berlin, pp 332–339
32. Over P, Awad G, Fiscus J, Antonishek B, Michel M, Smeaton FA, Kraaij W, Quénot G (2011) An overview of the goals, tasks, data, evaluation mechanisms and metrics. In: TRECVID 2011 - TREC video retrieval evaluation online, Gaithersburg, MD, USA
33. Penet C, Demarty C-H, Gravier G, Gros P (2011) 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
34. Penet C, Demarty C-H, Soleymani M, Gravier G, Gros P (2012) Technicolor/Inria/Imperial College London at the Mediaeval 2012 violent scene detection task. In: MediaEval 2012, multimedia benchmark workshop
35. Perperis T, Giannakopoulos T, Makris A, Kosmopoulos DI, Tsekeridou S, Perantonis SJ, Theodoridis S (2011) Multimodal and ontology-based fusion approaches of audio and visual processing for violence detection in movies. J Expert Syst Appl 38(11):14102–14116
36. Safadi B, Quéenot G (2011) Lig at Mediaeval 2011 affect task: use of a generic method. In: MediaEval 2011, multimedia benchmark workshop
37. Schlüter J, Ionescu B, Mironică I, Schedl M (2012) Arf @ mediaeval 2012: an uninformed approach to violence detection in hollywood movies. In: MediaEval 2012, multimedia benchmark workshop
38. Vasconcelos N, Lippman A (1997) Towards semantically meaningful feature spaces for the characterization of video content. In: Proceedings of the IEEE international conference on image processing, vol 1, pp 25–28
39. Wang S, Jiang S, Huang Q, Gao W (2008) Shot classification for action movies based on motion characteristics. In: Proceedings of the IEEE international conference on image processing, pp 2508–2511
40. WHO (1996) Violence: a public health priority. Technical Report WHO/EHA/SPI.POA.2, World Health Organization, Geneva, Switzerland
41. Zajdel W, Krijnders JD, Andringa T, Gavrila DM (2007) Cassandra: audio-video sensor fusion for aggression detection. In: IEEE conference on advanced video and signal based surveillance, 2007. AVSS 2007. IEEE, pp 200-205

Claire-Hélène Demarty graduated from Telecom ParisTech in 1994 and received a Ph.D. degree in Computer Science, Mathematical Morphology, from Mines ParisTech in 2000. She joined the Technicolor Research & Innovation Center in 2004 as a senior researcher and became senior scientist in 2010. Located in Rennes, France, she is working on multimedia indexing technologies. Prior to Technicolor, she worked at LTUTechnologies (2000–2002) and at INRIA Rennes - IRISA (2003–2004), a French public research center, as a researcher in image analysis and video indexing technologies. She is author or co-author of more than 15 papers and is holding several patents. Her research interests are in multimedia content analysis, in particular multimodal event detection, and multimodal statistical modeling.

Cédric Penet is a Ph. D. student in Technicolor R&D France in Rennes, France. He works on violence detection in movies and audio events detection. Prior to working for Technicolor, he obtained an engineering Master's degree from Institut National des Sciences Appliquées (INSA) de Rennes, France.

Mohammad Soleymani is a Marie Curie Fellow at the intelligent Behaviour Understanding Group (iBUG) at Imperial College London, where he conducts research on sensor-based and implicit emotional tagging. Soleymani received his PhD in computer science from the University of Geneva, Switzerland in 2011. He has worked extensively on assessing emotional reactions in response to video content and developing multimedia techniques to predict these reactions. He has served as a special session chair, program committee member and reviewer for multiple conferences and workshops including ACM ICMR, ACM MM, ACM ICMI, IEEE SMC, and IEEE ICME.

Guillaume Gravier After industry work on speech synthesis at ELAN Informatique, Guillaume Gravier obtained a Ph. D. in Signal and Image Processing at the Ecole National Superieure des Telecommunications (ENST Paris) in 2000. After a one year post-doctoral stay at Irisa, he joined the Audio Visual Speech Technology group at IBM T. J. Watson research center from 2001 to 2002. Since 2002, he is a research scientist at the Centre National pour la Recherche Scientifique (CNRS), working at the Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA). His research interests are in multimedia content analysis, in particular speech recognition, spoken language processing and multimodal statistical modeling.