VSD2014: A Dataset for Violent Scenes Detection in Hollywood Movies and Web Videos

Markus Schedl*, Mats Sjöberg†, Ionuţ Mironică‡, Bogdan Ionescu‡, Vu Lam Quang§, Yu-Gang Jiang¶, Claire-Hélène Demarty‖

*Johannes Kepler University, Linz, Austria, Email: *markus.schedl@jku.at*

†Helsinki Institute for Information Technology, University of Helsinki, Finland, Email: *mats.sjoberg@helsinki.fi*

‡University Politehnica of Bucharest, Romania, Email: *imironica@imag.pub.ro, bionescu@imag.pub.ro*

§University of Science, VNU-HCMC, Vietnam, Email: *lamquangvu@gmail.com*

¶Fudan University, China, Email: *yugang.jiang@gmail.com*

‖Technicolor, Rennes, France, Email: *claire-helene.demarty@technicolor.com*

***Abstract*—In this paper, we introduce a violent scenes and violence-related concept detection dataset named *VSD2014*. It contains annotations as well as auditory and visual features of Hollywood movies and user-generated footage shared on the web. The dataset is the result of a joint annotation endeavor of different research institutions and responds to the real-world use case of parental guidance in selecting appropriate content for children. The dataset has been validated during the *Violent Scenes Detection* (VSD) task at the *MediaEval* benchmarking initiative for multimedia evaluation.**

I. Introduction

For parents of young children, deciding which movies are suitable to watch with regard to their level of violence can be a big challenge. Even though official violence ratings may exist, the reaction to different types of violence may be very individual. Here, computer systems that can automatically detect violent scenes in video material using multimedia analysis techniques can serve an important use case. Such a system may be used to distill the most violent scenes in a movie, which can then be easily reviewed by the parents for making the final decision.

The explosive growth of user-generated videos shared on web media platforms such as *YouTube*[1] presents a fresh challenge for automatic violence detection systems, since such videos are often characterized by bad video and audio quality as well as short duration. In the presented dataset, both kinds of material are considered.

To support the emerging research on the task of violence detection, this paper introduces an annotated dataset, named *VSD2014*, which is intended for benchmarking violence detection in Hollywood movies and short user-generated videos. Parts of the dataset are used in the *MediaEval Violent Scenes Detection* task [1], which has been run annually since 2011. The *VSD2014* set is a considerable extension of similar datasets we presented previously [16], [5], [6], in several regards. First, it considers for all movies and user-generated videos a violence definition closer to the targeted real-world scenario by focusing on physical violence one would not let an 8-year-old child watch ("subjective definition"). This definition was used in the annotation process. In contrast, the previously employed definition considers physical violence or accident resulting in injury or pain ("objective definition"). Second, the dataset at hand relates to a substantial set of 31 Hollywood movies. Third, *VSD2014* includes 86 web video clips and their metadata retrieved from *YouTube* to serve for testing the generalization capabilities of approaches to different types of footage. Forth, it includes state-of-the-art audio-visual content descriptors.

The remainder of the paper is organized as follows. Section II overviews the related work and positions our contribution accordingly. Subsequently, we introduce the dataset, report statistical details, outline the methodology used to create the dataset, and explain its format in Section III. Section IV discusses the validation of the dataset during the 2014 *MediaEval*[2] *Violence Scenes Detection* (VSD) task[3] [1] and proposes several baselines for benchmarking. We round off the paper by a summary and possibilities for future extension of the dataset in Section V.

II. Related Work

Due to the complexity of the research problem, which first necessitates defining the concept of *violence*, before elaborating on methods to infer semantic concepts out of low-level information, violence detection in videos has been marginally studied in the literature until recently [2], [3], [4]. In particular, related work reveals a wide variety in the *definition of violence* and the used *datasets*.

The high variability of violent events in videos results in a high diversity of interpretations for violence detection. For instance, the authors of [7] target "a series of human actions accompanied with bleeding"; in [8], [9], authors search for "scenes containing fights, regardless of context and number of people involved"; in [10], the focus is on "behavior by persons against persons that intentionally threatens, attempts, or actually inflicts physical harm"; in [11], authors are interested in "fast paced scenes which contain explosions, gunshots and person-on-person fighting".

At the same time, there is a huge diversity of data, both in size and content, which are used to validate existing methods. These data are usually closed and adapted to a specific context

[1] http://www.youtube.com

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

[3] http://www.multimediaeval.org/mediaeval2014/violence2014

of a certain method. For instance, authors of [12] use for validation 15 short sequences (around 12 seconds each) with 40 violent scenes performed by 8 different persons. In [13], 13 clips (up to 150 seconds) are recorded at a train station featuring 2–4 professional actors who are engaged in a variety of activities, ranging from walking, shouting, running to pushing, hitting a vending machine, and clashing. In [9], evaluation is carried out on 1,000 short sport clips (2 seconds each) containing different fight/non-fight actions from ice hockey videos. In [11], the authors use 4 Hollywood movies, from the science-fiction, war, crime, and thriller genre. In [10], 50 video segments ripped from 10 different movies (totaling 150 minutes) are used.

The lack of a common definition and the resulting absence of a substantial reference dataset make it very difficult to compare methods, which are frequently developed for a very specific type of violence. This is precisely the issue that we attempt to correct with this release. We propose a well-formulated violence detection use case and a corresponding annotated dataset, *VSD2014*, for benchmarking violent scenes detection in movies and videos. We provide annotations of violent scenes and of violence-related concepts for a collection of (i) Hollywood movies and (ii) user-generated videos shared on the web. In addition to the annotations, pre-computed audio and visual features and various metadata are provided to facilitate the contribution of different research communities, such as signal processing or machine learning, and to encourage multimodal approaches. This dataset is also intended to support related areas such as event detection, multimedia affect detection, and multimedia content analysis.

III. Dataset Description

The *VSD2014* dataset is split into three different subsets, called *Hollywood: Development*, *Hollywood: Test*, and *YouTube: Generalization*.[4] Table I gives an overview of the three subsets and provides basic statistics, including duration, fraction of violent scenes (as percentage on a per-frame-basis), and average length of a violent scene. Note that space limitations prevent us from providing detailed statistics on the *YouTube: Generalization* subset, instead we have given the average and standard deviation calculated across all videos.

The dataset is publicly available for download as one single compressed file[5] of 11.5 GB or as 10 smaller files for the ease of downloading.[6] The content of the *VSD2014* dataset can be categorized into three types: movies/videos (and metadata), features, and annotations. In the following, we describe in detail the content, the data acquisition process, and the data format.

A. Movies/Videos

For the Hollywood part of the dataset, we selected various popular movies, ranging from very violent ones (e.g., *Saving Private Ryan* with 34% violent frames) to movies with (almost) no violence (e.g., *Dead Poets Society* with <1% of violent frames and *Legally Blond* with no violence at all). The chosen movies range in their release year between the 1930s and the 2000s, with a strong focus on the 1990s and 2000s. While the dataset does not include the actual movies dues to copyright reasons, it provides annotations for all 31 movies, 24 in the *Hollywood: Development* and 7 in the *Hollywood: Test* set.

TABLE I. Statistics of the movies and videos in the *VSD2014* subsets. Columns indicate movie name, total playtime, fraction of violence, and average duration of a violent scene in seconds.

Name	Duration	V (%)	Avg. V
Hollywood: Development			
Armageddon	8,680.16	7.78	25.01
Billy Elliot	6,349.44	2.46	8.68
Dead Poets Society	7,413.20	0.58	14.44
Eragon	5,985.44	13.26	39.69
Fantastic Four 1	6,093.96	20.53	62.57
Fargo	5,646.40	15.04	65.32
Fight Club	8,004.50	15.83	32.51
Forrest Gump	8,176.72	8.29	75.33
Harry Potter 5	7,953.52	5.44	17.30
I am Legend	5,779.92	15.64	75.36
Independence Day	8,833.90	13.13	68.23
Legally Blond	5,523.44	0.00	0.00
Leon	6,344.56	16.36	41.52
Midnight Express	6,961.04	7.12	24.80
Pirates of the Caribbean	8,239.40	18.15	49.85
Pulp Fiction	8,887.00	25.05	202.43
Reservoir Dogs	5,712.96	30.41	115.82
Saving Private Ryan	9,751.00	33.95	367.92
The Bourne Identity	6,816.00	7.18	27.21
The God Father	10,194.70	5.73	44.99
The Pianist	8,567.04	15.44	69.64
The Sixth Sense	6,178.04	2.00	12.40
The Wicker Man	5,870.44	6.44	31.55
The Wizard of Oz	5,859.20	1.02	8.56
Total	**180,192.40 (50h02)**	**12.35**	
Hollywood: Test			
8 Mile	6,355.60	4.70	37.40
Braveheart	10,223.92	21.45	51.01
Desperado	6,012.96	31.94	113.00
Ghost in the Shell	4966.00	9.85	44.47
Jumanji	5993.96	6.75	28.90
Terminator 2	8831.40	24.89	53.62
V for Vendetta	7625.88	14.27	25.91
Total	**50,009.72 (13h53)**	**17.18**	
YouTube: Generalization			
Average (std.dev.)	109.76 (68.05)	31.69 (36.28)	26.62 (50.41)
Total	**9,439.39 (2h37)**	**31.69**	

For the web videos (*YouTube: Generalization* set), we considered video clips shared on *YouTube* under the Creative Commons Attribution 3.0 Unported license in order to be able to redistribute the actual video material as part of the *VSD2014* dataset. To retrieve relevant videos, we identified queries targeted at violent material, such as "brutal accident" or "killing video games". The results to each query underwent a preliminary informal investigation based on which we selected an approximately uniform number of violent and non-violent videos. This process yielded a total of 86 clips, which are included as MP4 files in the dataset. The clips are between 6 seconds and 6 minutes in length. We additionally retrieved metadata offered through the *YouTube* API:[7] video identifier, publishing date, updating date, category, title, author, uploader identifier, aspect ratio, duration, user rating (minimum, maximum, average), number of raters, and number of likes and dislikes. This metadata is provided as XML files.

[4] In order to avoid confusion, we decided to keep the names of the subsets as they had been chosen for the *MediaEval Violent Scenes Detection* task (cf. Section IV).

[5] http://www.cp.jku.at/datasets/VSD2014/VSD2014.zip

[6] http://www.cp.jku.at/datasets/VSD2014/VSD2014.zip.[001-009]

[7] https://developers.google.com/youtube

B. Features

We provide a set of common audio and visual descriptors which may serve as a starting point for a violence detection algorithm. These are targeted at interested researchers who are new to the field or who are interested in aspects other than feature extraction, e.g., in classification.

From the *audio*, we provide on a per-video-frame-basis amplitude envelop (AE), root-mean-square energy (RMS), zero-crossing rate (ZCR), band energy ratio (BER), spectral centroid (SC), frequency bandwidth (BW), spectral flux (SF), and Mel-frequency cepstral coefficients (MFCC). As the audio exhibits a sampling rate of $44{,}100$ Hz and the videos are encoded with 25 fps, we consider windows of $1{,}764$ audio samples in length. We compute 22 MFCC for each window, while all other features are 1-dimensional.

For what concerns *visual* features, we include color naming histograms (CNH), color moments (CM), local binary patterns (LBP), and histograms of oriented gradients (HOG). The CNH features are 99-dimensional, the CM and HOG features 81-dimensional, and the LBP 144-dimensional. The CNH are computed on 3-by-3 image regions and map colors to 11 universal color names: black, blue, brown, gray, green, orange, pink, purple, red, white, and yellow. The global CM in the hue-saturation-value (HSV) color space (9 values) contain the first three central moments of an image color distribution: mean, standard deviation, and skewness, which are computed on 3-by-3 image regions [14]. Also, the global LBP (16 values) and the global HOG are computed using a 3-by-3 spatial division. The global HOG contain the average of the HOG features (9 values), that exploit the local object appearance and shape within an image via the distribution of edge orientations. The LBP features represent a powerful tool for texture classification. Furthermore, it has been shown that combining LBP with HOG descriptors is advantageous for certain tasks [15].

C. Annotations

The *VSD2014* dataset contains binary annotations of all violent scenes, where a scene is identified by its start and end frames. For 17 of the Hollywood movies (printed in italics in Table I), we additionally provide violence-related concept annotations, where the concept is indicated either visually or aurally.

The annotations for Hollywood movies and *YouTube* videos have been created by several human assessors in a hierarchical, bottom-up manner. For the annotators, we define violent scenes as *scenes one would not let an 8-year-old child watch because they contain physical violence*.[8]

The annotation protocol consists of the following procedure: In a first step, all videos are annotated separately by two groups of annotators from two different countries (China and Vietnam). Each group consists of regular annotators and master annotators.[9] The former are graduate students (typically single with no children), while the latter are senior researchers (typically married with children). All movies are first labeled by the regular annotators so that each movie receives 2 different sets of annotations. These annotations are subsequently reviewed and merged by the master annotators of each group, in order to ensure a certain level of consistency. The annotations from the two groups are then merged and reviewed by a different set of master annotators. Borderline cases are resolved via panel discussions among this larger set of master annotators, who originate from six different countries in Asia and Europe to ensure a wide cultural diversity, and in turn reduce cultural bias in the annotations. We hence believe that the annotations can be regarded as widely agreed upon. Even though we adopted this rigid annotation protocol, we faced several borderline cases during the annotation process. Some examples and the corresponding annotations agreed on are given in Table II.

TABLE II. SOME BORDERLINE CASES THAT EMERGED DURING THE ANNOTATION PROCESS AND THE CORRESPONDING FINAL CHOICE.

Description of the Scene	Violence
Scenes showing the result of a violent action, without seeing the action itself.	✓
Actions revealing an intent to kill, but which fail.	✓
Violent behavior against dead people.	✓
Scenes from cartoons that show violent actions.	✓
Medical treatment as a result of violence against a person.	✓
Medical treatment without obvious connection to a violent act.	✗
Pushing between players of a soccer game.	✗
Violent actions against supernatural humans, who do not suffer thanks to their powers.	✗
A person fighting alone or against an invisible person.	✗
Car crashes with no people injured.	✗

All violent segments are annotated at video frame level, i.e., a violent segment is defined by its starting and ending frame numbers. Each annotated violent segment contains only one action, whenever this is possible. In cases where different actions are overlapping, the segments are merged. This is indicated in the annotation files by adding the tag "multiple action scene".

In addition to binary annotations of segments containing physical violence, annotations also include high-level concepts for 17 movies in the *Hollywood: Development* set. In particular, 7 visual concepts and 3 audio concepts are annotated, employing a similar annotation protocol as used for violent/non-violent annotations. The concepts are: presence of blood, fights, presence of fire, presence of guns, presence of cold arms, car chases, and gory scenes, for the visual modality; presence of gunshots, explosions, and screams for the audio modality. Furthermore, these high-level concepts may be refined by an additional description indicating a certain intensity (e.g., blood:low) or detail (e.g., fight:one-versus-one or gore:decapitation). An exhaustive list of these additional descriptions is provided as an external resource.[10] Please note that the annotations for the auditory concepts are provided by start and end times in seconds, while visual concepts are annotated on the frame level. The reason for this is that the annotation process for video is based on per-frame analysis of the respective image, while audio annotations are performed on a temporal scale.

[8] We previously experimented with other, more objective, definitions, but classifiers trained on those definitions typically perform inferior to algorithms targeting the subjective definition provided by the dataset at hand [16].

[9] The first group encompasses 2 regular annotators and 1 master annotator, while the second group consists of 5 regular annotators and 3 master annotators.

[10] http://www.cp.jku.at/datasets/VSD2014/description.html

D. Data Format

The directory structure of the dataset is the following:

```
VSD2014.zip
├── Hollywood-dev
│   ├── annotations
│   │   ├── [movie]_violence.txt
│   │   └── [movie]_[concept].txt
│   └── features
│       ├── [movie]_auditory.mat
│       └── [movie]_visual.mat
├── Hollywood-test
│   ├── annotations
│   │   └── [movie]_violence.txt
│   └── features
│       ├── [movie]_auditory.mat
│       └── [movie]_visual.mat
├── YouTube-gen
│   ├── annotations
│   │   └── [video-id]_violence.txt
│   ├── features
│   │   ├── [video-id]_auditory.mat
│   │   └── [video-id]_visual.mat
│   ├── metadata
│   │   └── [video-id].xml
│   └── videos
│       └── [video-id].mp4
└── import_VSD2014.py
```

At the root level of the compressed file `VSD2014.zip`, you can find folders corresponding to the three subsets *Hollywood: Development*, *Hollywood: Test*, and *YouTube: Generalization*. Each of them contains several subfolders holding the data that is available for each subset. More precisely, violence annotations are stored as text files in folder `annotations`. The binary violence annotations are stored in files `[movie]_violence.txt`; the detailed concept annotations in files named `[movie]_[concept].txt`. Both come in space-separated text format. Each line in the former thus complies to the structure:

`start-frame end-frame`.

Each line in the latter either complies to:

`start-frame end-frame [concept-detail]`

or

`start-second end-second [concept-detail]`,

respectively, for visual and audio annotations.

Audio and visual features are provided in *Matlab* version 7.3 MAT files, which correspond to HDF5 format, and are located in folder `features`.[11] For easy use in Python, we further provide a script `import_VSD2014.py` that shows how to import the MAT files. The files named `[movie]_auditory.mat` contain the variables AE, RMS, ZCR, BER, SC, BW, SF, and MFCC; the files named `[movie]_visual.mat` hold the variables CNH, CM, LBP, and HOG. These abbreviations correspond to the audio and video features introduced in Section III-B.[12]

The *YouTube* videos were converted to MP4 format using 25 fps and are provided in the folder `videos`. The corresponding XML files containing the metadata extracted from *YouTube* can be found in the folder `metadata`.

IV. The Affect Task at MediaEval

In the following, we illustrate an application of the proposed dataset to the real-world use case of guiding parents in deciding which movies or videos are suited for their children with respect to violence. Parents may choose to select or reject movies after previewing the most violent parts of the movies. This use case was defined based on real commercial demands by the company *Technicolor*.[13] Targeting this use case, we have been running the violent scenes detection task as part of the *MediaEval* evaluation campaign since 2011. The amount of movies and annotations has increased from year to year and in 2014 it reached the figures reported for the *VSD2014* dataset in this paper.

A. Task description

The goal of the *Violent Scenes Detection* task in *MediaEval* 2014 was to automatically detect violent segments in movies and indicate the start and end frames of each violent scene, for each video. From this information, it is easy to create a summary video of the most violent scenes, which serves for parental guidance.

Task participants were provided the three subsets included in *VSD2014*, excluding the audio and visual features. The *Hollywood: Development* set came with complete annotations and was intended for training purposes. The *Hollywood: Test* set was distributed slightly later, withholding its annotations. This set was used to evaluate the performance of the submitted algorithms. In addition, in 2014 we ran for the first time a generalization task, using the set *YouTube: Generalization*, in order to investigate how well the proposed algorithms generalize to video material other than Hollywood movies. The data used in the generalization task also addresses the emerging interest in user-generated videos shared on the web, which are often characterized by short duration and inferior quality. For this task, we distributed the actual video clips and the corresponding metadata given by *YouTube*.

B. Summary of Approaches

Eight participating teams submitted a total of 67 experimental runs (37 for the Hollywood movies and 30 for the *YouTube* videos). Each submitted run had to indicate start and end frames of all detected violent scenes and a score indicating the confidence in the prediction. The teams were allowed to submit up to 5 runs for each task: the main task involving the Hollywood movies and the generalization task on the *YouTube* videos.

The proposed approaches were typically multimodal. Except for one team, all participants employed algorithms that

[11] We decided to use Matlab/HDF5 format because it is very compact. Other formats are available upon request to the authors.

[12] Numerical issues caused slightly different numbers of audio frames and video frames for some movies. However these offsets are at most a few frames and might just marginally impair the performance of algorithms.

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

TABLE III. PERFORMANCE FIGURES FOR THE MAIN TASK (HOLLYWOOD MOVIES). PRECISION, RECALL, MAP@100, AND MAP2014 VALUES ARE SHOWN FOR THE BEST RUN OF EACH PARTICIPATING TEAM.

Team	Prec.	Rec.	MAP@100	MAP2014
FUDAN [24]	41.1%	72.1%	72.7%	63.0%
NII-UIT [25]	17.1%	100.0%	77.3%	55.9%
FAR [26]	28.0%	71.3%	57.0%	45.1%
MIC-TJU [27]	17.0%	98.4%	63.6%	44.6%
RECOD [28]	33.0%	69.7%	49.3%	37.6%
VIVOLAB [29]	38.1%	58.4%	38.2%	17.8%
TUB-IRML [30]	31.7%	17.3%	40.9%	17.2%
MTMDCC [31]	15.8%	24.6%	16.5%	2.6%

TABLE IV. PERFORMANCE FIGURES FOR THE GENERALIZATION TASK (*YouTube* MOVIES). PRECISION, RECALL, MAP@100, AND MAP2014 VALUES ARE SHOWN FOR THE BEST SUBMITTED ALGORITHM OF EACH PARTICIPATING TEAM.

Team	Prec.	Rec.	MAP@100	MAP2014
FAR [26]	49.7%	85.8%	86.0%	66.4%
RECOD [28]	48.1%	88.4%	86.8%	61.8%
FUDAN [24]	59.0%	43.4%	71.9%	60.4%
MIC-TJU [27]	44.4%	97.3%	55.5%	56.6%
TUB-IRML [30]	63.3%	25.2%	58.2%	51.7%
VIVOLAB [29]	51.3%	33.6%	56.5%	43.0%

made use of at least visual and auditory information. One team additionally incorporated textual features extracted from the movie subtitles. The provided violence-related concept annotations were used by two teams.

The most common features for the audio modality were MFCC, sometimes complemented with other low-level features, such as zero-crossing rate or spectral centroid. In contrast, a wider range of video features was used. Improved dense trajectories [17] was a particularly popular feature, implemented using, e.g., histograms of oriented gradients (HOG), histograms of optical flow (HOF), or motion boundary histograms (MBH). Two teams additionally included static image features, such as SIFT [18], colorfulness, saturation, brightness, and hue. Visual features were frequently encoded using Fisher vectors [19] and modeled via Gaussian mixture models (GMMs) [20].

Most frequently used classifiers include support vector machines (SVM) [21] and deep neural networks (DNN) [22]; one team used a multilayer perceptron (MLP) [23]. The recent popularity of DNN was reflected by several runs using them either directly for classification or fusion, or to generate visual features which were then used with a traditional SVM classifier.

C. Evaluation and Results

In the following, we present the results of the 2014 edition of the *Violent Scenes Detection* task in *MediaEval*, which may serve as a baseline for prospective users of the *VSD2014* dataset. As evaluation metric, we used an adapted version of mean average precision (MAP), which we call MAP2014. It avoids the problem that participating teams could artificially increase their algorithms' MAP scores by predicting more, but shorter segments within the boundaries of a violent scene. Hence, MAP inherently penalizes algorithms that identify larger segments, while such algorithms should rather be rewarded, because larger segments more likely correspond to a whole scene and ease interpretation. Concretely, we define the MAP2014 measure as follows: All segments marked as violent by the algorithm under consideration are sorted in descending order, according to the given confidence scores. An algorithm's prediction of violence is considered a hit if the predicted segment overlaps with the corresponding ground truth segment by more than 50% (or the other way round). To counteract the above mentioned problem when predicting many short segments, several hits on the same ground truth segment only count as one true positive. The others are ignored, thus not counted as false positives either.

The results of the eight participating teams are summarized in Tables III and IV for the main and the generalization task[14], respectively. In addition to MAP2014 figures, we report precision, recall, and MAP@100. Algorithms are sorted with respect to MAP2014. The best result in the main task was achieved by using a novel variant of deep neural networks for both classification and fusion [32]. The other aspect that stood out in the main task was the power of auditory features. In fact, some teams achieved their best result by using only audio. Comparing the numbers between Tables III and IV it would appear that algorithms performed better in the generalization task. However, when interpreting these results, it is important to consider the much higher percentage of violence in the generalization task subset. Likewise, the performance of a random run, which is 29.4% MAP2014 on the *YouTube* dataset, but only 5.5% on the Hollywood dataset, has to be taken into account. The random runs were generated by alternating non-violent and violent scenes with normally distributed lengths. The distribution parameters for non-violent and violent segments were estimated separately from the training set, and scaled down by the change in average clip length for the *YouTube* videos. Nevertheless, running the generalization task can be considered a success, since most systems generalized well from the rather different training set. The popularity of user-generated videos and easier distribution of the videos, compared to Hollywood movies, also makes this an attractive task.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a dataset, *VSD2014*, intended for benchmarking violence detection in Hollywood movies and *YouTube* videos. The dataset includes manual annotations of violent and non-violent segments as well as of violence-related concepts. It was validated during the 2014 edition of the *Violent Scenes Detection* task as part of the *MediaEval* campaign. The high number of submitted runs and the variety of approaches make *VSD2014* a substantial validation dataset for violence detection.

Future extensions of this dataset will mainly focus on: (i) addressing fine-grained violence-related concepts instead of binary violence/non-violent annotations, e.g., pushing in a soccer game vs. brutal torturing; (ii) extending the definition of violence to other scenarios, e.g., sexual violence or mental violence; (iii) investigating different use case scenarios that may address different age classes; and (iv) considering more freely distributable footage.

[14]Due to the ambiguity of evaluating completely non-violent video sequences (26 in gen. task), these were excluded from the results reported here.

ACKNOWLEDGMENTS

This work is supported by the following projects: Austrian Science Fund (FWF) P25655, UEFISCDI SCOUTER grant 28DPST/30-08-2013, ESF POSDRU/159/1.5/S/132395 InnoRESEARCH program, National Natural Science Foundation of China grants 61201387 and 61228205, Vietnam National University Ho Chi Minh City grant B2013-26-01, and EU FP7-ICT-2011-9 project 601166 ("PHENICX").

We would further like to thank all annotators and participants of the *MediaEval Violent Scenes Detection* tasks for their vital contributions to creating and using the dataset.

REFERENCES

[1] M. Sjöberg, B. Ionescu, Y.-G. Jiang, V. L. Quang, M. Schedl, and C.-H. Demarty, "The MediaEval 2014 Affect Task: Violent Scenes Detection," in *Working Notes Proc. of the MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, October 2014.

[2] E. Acar, F. Hopfgartner, and S. Albayrak, "Violence Detection in Hollywood Movies by the Fusion of Visual and Mid-level Audio Cues," in *Proc. ACM Multimedia*, Barcelona, Spain, October 2013, pp. 717–720.

[3] B. Ionescu, J. Schlüter, I. Mironica, and M. Schedl, "A Naive Mid-level Concept-based Fusion Approach to Violence Detection in Hollywood Movies," in *Proc. ICMR*, Dallas, TX, USA, April 2013, pp. 215–222.

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

[5] 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, May 2014.

[6] 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, 2014, pp. 185-208.

[7] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su, "Violence Detection in Movies," in *Proc. CGIV*, Singapore, August 2011, pp. 119–124.

[8] F. De Souza, G. Chávez, E. do Valle, and A. De A Araujo, "Violence Detection in Video Using Spatio-Temporal Features," in *Proc. SIBGRAPI*, Gramado, Rio Grande do Sul, Brazil, August–September 2010, pp. 224–230.

[9] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence Detection in Video Using Computer Vision Techniques," in *Computer Analysis of Images and Patterns*. Springer, 2011, pp. 332–339.

[10] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-Visual Fusion for Detecting Violent Scenes in Videos," in *Artificial Intelligence: Theories, Models and Applications*, ser. Lecture Notes in Computer Science, S. Konstantopoulos et al., Ed. Springer, 2010, vol. 6040, pp. 91–100.

[11] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Detecting violent scenes in movies by auditory and visual cues," in *Advances in Multimedia Information Processing - PCM 2008*, ser. Lecture Notes in Computer Science, Y.-M. Huang et al., Ed. Springer Berlin / Heidelberg, 2008, vol. 5353, pp. 317–326.

[12] A. Datta, M. Shah, and N. Da Vitoria Lobo, "Person-on-person Violence Detection in Video Data," in *Proc. IEEE Pattern Recognition*, Quebec City, Canada, August 2002, pp. 433–438.

[13] W. Zajdel, J. D. Krijnders, T. Andringa, and D. M. Gavrila, "CASSANDRA: Audio-video Sensor Fusion for Aggression Detection," in *Proc. IEEE AVSS*, London, UK, September 2007, pp. 200–205.

[14] M. Stricker and M. Orengo, "Similarity of Color Images," *SPIE Storage and Retrieval for Image and Video Databases III*, vol. 2420, no. 2, pp. 381–392, 1995.

[15] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP Human Detector with Partial Occlusion Handling," *Proc. IEEE ICCV*, September 2009.

[16] C.-H. Demarty, B. Ionescu, Y.-G. Jiang, V. Quang, M. Schedl, and C. Penet, "Benchmarking Violent Scenes Detection in Movies," in *Proc. CBMI*, Klagenfurt, Austria, June 2014.

[17] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *Proc. IEEE ICCV*, Sydney, Australia, December 2013.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[19] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," in *Proc. ECCV*, Crete, Greece, September 2010, pp. 143–156.

[20] K. K. Yiu, M. wai Mak, and C. kwong Li, "Gaussian Mixture Models and Probabilistic Decision-Based Neural Networks for Pattern Classification: A Comparative Study," *Neural Computing and Applications*, vol. 8, pp. 235–245, February 1999.

[21] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[23] D. Ruck, S. Rogers, M. Kabrisky, M. Oxley, and B. Suter, "The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function," *IEEE Transactions on Neural Networks*, vol. 1, no. 4, pp. 296–298, December 1990.

[24] Q. Dai, Z. Wu, Y.-G. Jiang, X. Xue, and J. Tang, "Fudan-NJUST at MediaEval 2014: Violent Scenes Detection Using Deep Neural Networks," in *Working Notes Proc. MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, October 2014.

[25] V. Lam, D.-D. Le, S. Phan, S. Satoh, and D. A. Duong, "NII-UIT at MediaEval 2014 Violent Scenes Detection Affect Task," in *Working Notes Proc. MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, October 2014.

[26] M. Sjöberg, I. Mironică, M. Schedl, and B. Ionescu, "FAR at MediaEval 2014 Violent Scenes Detection: A Concept-based Fusion Approach," in *Working Notes Proc. MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, October 2014.

[27] B. Zhang, Y. Yi, H. Wang, and J. Yu, "MIC-TJU at MediaEval Violent Scenes Detection (VSD) 2014," in *Working Notes Proc. MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, October 2014.

[28] S. Avila, D. Moreira, M. Perez, D. Moraes, I. Cota, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "RECOD at MediaEval 2014: Violent Scenes Detection Task," in *Working Notes Proc. MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, October 2014.

[29] D. Castán, M. Rodríguez, A. Ortega, C. Orrite, and E. Lleida, "ViVoLab and CVLab - MediaEval 2014: Violent Scenes Detection Affect Task," in *Working Notes Proc. MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, October 2014.

[30] E. Acar and S. Albayrak, "TUB-IRML at MediaEval 2014 Violent Scenes Detection Task: Violence Modeling through Feature Space Partitioning," in *Working Notes Proc. MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, October 2014.

[31] B. do Nascimento Teixeira, "MTM at MediaEval 2014 Violence Detection," in *Working Notes Proc. MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, October 2014.

[32] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring Inter-feature and Inter-class Relationships with Deep Neural Networks for Video Classification," in *Proc. ACM Multimedia*, Orlando, FL, USA, November 2014.