Acar, E., Hopfgartner, F., & Albayrak, S.

Detecting violent content in Hollywood movies by mid-level audio representations

Acar, E., Hopfgartner, F., & Albayrak, S. (2013). Detecting violent content in Hollywood movies by mid-level audio representations. In 2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI). IEEE. https://doi.org/10.1109/cbmi.2013.6576556

Detecting Violent Content in Hollywood Movies by Mid-level Audio Representations

Esra Acar [1], Frank Hopfgartner [2], Sahin Albayrak [3]

DAI Laboratory, Technische Universität Berlin
Ernst-Reuter-Platz 7, TEL 14, 10587 Berlin, Germany
[1] esra.acar@dai-labor.de [2] frank.hopfgartner@dai-labor.de[3] sahin.albayrak@dai-labor.de

***Abstract*—Movie violent content detection e.g., for providing automated youth protection services is a valuable video content analysis functionality. Choosing discriminative features for the representation of video segments is a key issue in designing violence detection algorithms. In this paper, we employ mid-level audio features which are based on a Bag-of-Audio Words (BoAW) method using Mel-Frequency Cepstral Coefficients (MFCC). BoAW representations are constructed with two different methods, namely the vector quantization-based (VQ-based) method and the sparse coding-based (SC-based) method. We choose two-class support vector machines (SVMs) for classifying video shots as (non-)violent. Our experimental results on detecting violent video shots in Hollywood movies show that the mid-level audio features provide promising results. Additionally, we establish that the SC-based method outperforms the VQ-based one. More importantly, the SC-based method outperforms the unimodal submissions in the MediaEval Violent Scenes Detection (VSD) task except one visual-based method in terms of average precision.**

I. Introduction

Equipments including DVB set top boxes (terrestrial, cable or satellite), Tablet PCs, high-speed Internet access or digital media-streaming devices are now part of the facilities usually found in the home of many families. Therefore, accessing online movies through services such as Video-On-Demand has become extremely easy. Children are, consequently, exposed to movies, documentaries, or reality shows which have not necessarily been checked by parents, and which potentially comprise inappropriate content.

One of these inappropriate contents is violence whose harmful effect, especially on children has been shown by psychological studies (e.g., [1]). Therefore, there is a need for automatically detecting violent scenes in videos, where the legal age ratings are not available. Defining the term "violence", when applied to characterize movies, is a hard task and subjective (i.e., person-dependent). In our work, we aim at sticking to the common definition of violence: "physical violence or accident resulting in human injury or pain" which is the definition of "violence" in the MediaEval VSD task [2].

Representing movie segments is an important step in the task of movie violent content detection as in any pattern recognition task. Existing works for video content understanding construct higher level representations from the low-level ones in order to model the relationship between low-level features and high-level human perception of videos [3]. However, high-level semantics are difficult to derive and state-of-the-art detectors are far from perfect. Therefore, using mid-level representations may help modeling video segments one step closer to human perception. Among the plurality of mid-level representations, bags of features, spatial pyramids, and the upper units of convolutional networks or deep belief networks are the popular examples of mid-level representations [4].

The aim of this paper is to investigate the discriminative power of mid-level audio features which are BoAW representations based on MFCC and constructed with two different methods (i.e., VQ-based and SC-based) for modeling violence in Hollywood movies. We show that promising results are obtained by both methods, while the SC-based method performs slightly better than the VQ-based one.

The paper is organized as follows. Section II explores the recent developments and reviews methods which have been proposed to detect violence in movies. In Section III, we introduce our method for violent content detection. We provide and discuss evaluation results on Hollywood movies in Section IV. Finally, we present concluding remarks and future directions to expand our current approach in Section V.

II. Related Work

In this section, we briefly discuss methods which are audio based or audio-visual based for violent content detection in videos with an emphasis on the audio analysis part of the discussed methods.

In [5], Giannakopoulos et al. define violent scenes as those containing shots, explosions, fights and screams, whereas non-violent content corresponds to audio segments containing music and speech. Frame-level audio features both from the time and the frequency domain such as energy entropy, short time energy, zero crossing rate, spectral flux and roll-off are employed. Polynomial SVM is used as the classifier. The main issue of this work is that audio signals are assumed to have already been segmented into semantically meaningful non-overlapping pieces (i.e., shots, explosions, fights, screams, music, speech). In order to overcome this audio segmentation issue, we segment audio signal using the visual shots of movies in our work.

The most common type of approach used violent content detection in videos is fusing audio and visual cues at either feature- or decision-level. Wang et al. [6] apply Multiple Instance Learning (MIL) (MI-SVM [7]) using color, textual and MFCC features. Video scenes are divided into video shots, where each scene is formulated as a bag and each shot as an instance inside the bag for MIL. Color and texture features

are used for the visual representation of video shots, while MFCC is used for the audio representation. More specifically, mean, variance and first-order differential of each dimension of MFCC are employed for the audio representation. As observed from their results [6], using color and textural information in addition to MFCC slightly improves the performance.

Giannakopoulos et al. [8], in an attempt to extend their approach based solely on audio cues [5], propose to use a multi-modal two-stage approach. In the first step, they perform audio and visual analysis of segments of one second duration. In the audio analysis part, audio features such as energy entropy, ZCR, MFCC are extracted and the mean and standard deviation of these features are used to classify scenes into one of seven classes (violent ones including shots, fights and screams). In the visual analysis part, average motion, motion variance and average motion of individuals appearing in a scene are used to classify segments as having either high or low activity. The classifications obtained in this first step are then used to train a k-NN classifier.

In [9], a three-stage method is proposed. In the first stage, the authors apply a semi-supervised cross-feature learning algorithm [10] on the extracted audio-visual features such as motion activity, ZCR, MFCC, pitch, rhythm features for the selection of candidate violent video shots. In the second stage, high-level audio events (e.g., screaming, gun shots, explosions) are detected via SVM training for each audio event. In the third stage, the outputs of the classifiers generated in the previous two stages are linearly weighted for final decision. The method was only evaluated on action movies. However, violent content can be present in movies of all genres (e.g., drama). The performance of this method in genres other than action is, therefore, unclear.

Lin et al. [11] train separate classifiers for audio and visual analysis and combine these classifiers by co-training. Probabilistic latent semantic analysis is applied in the audio classification part. Spectrum power, brightness, bandwidth, pitch, MFCC, spectrum flux, ZCR and harmonicity prominence features are extracted. An audio vocabulary is subsequently constructed by k-means clustering. Audio clips of one second length are represented by the audio vocabulary. This method also constructs mid-level audio representations with a technique derived from text analysis. However, this approach presents the drawback of only constructing a dictionary of twenty audio words, which prevents having a precise representation of the audio signals of video shots. In the visual classification part, the degree of violence of a video shot is determined by using motion intensity, the (non-)existence of flame, explosion and blood appearing in the video shot. This method was also evaluated only on action movies. Therefore, the performance of this solution in genres other than action is uncertain.

To summarize, MFCC is proven to be effective in video content analysis. The following two points define possibilities for improvement. First, the video violent content analysis methods that employ MFCC to represent video segments mostly use low-level features such as mean and standard deviation based on MFCC. Second, some problems exist concerning the construction of mid-level audio representations as in [11]. In our current framework, we only exploit the audio modality of videos to detect violent segments, since sound effects are essential elements which film-makers make use of in order to stimulate people's perception. Our approach differs from the aforementioned works in the following aspects: (1) we stick to a broad definition of "violence" [12], (2) we evaluate our approach on a diverse benchmarking dataset [2] (i.e., not a restricted dataset which contains only action movies), (3) we construct mid-level audio representations by a BoAW approach with two different coding schemes (i.e., vector quantization and sparse coding), and (4) we show that these mid-level audio representations provide promising results and the sparse coding-based BoAW method outperforms the unimodal submissions in the MediaEval VSD task except one visual-based method in terms of average precision.

III. The Violence Detection Method

The representation of video shots and the learning of a violence model are the two main components of the method which we discuss in detail in the following two subsections.

A. Representation of Videos

Among the plurality of audio features, MFCC features are shown to be indicators of the excitement level of video segments [13]. Therefore, we employ these features as low-level audio features. For the representation of video shots, we use mid-level audio features based on MFCC (i.e., BoAW approach). We apply the BoAW approach with two different coding schemes, namely vector quantization-based (VQ-based) and sparse coding-based (SC-based). The feature extraction process is illustrated in Figure 1(a).

1) Audio Representation by Vector Quantization: The construction of the VQ-based audio dictionary is illustrated in Figure 2. We follow an unsupervised way of constructing the audio dictionary. First, we cluster MFCC feature vectors extracted from video shots with a k-means clustering, in which the centroid of each of the k clusters is treated as an audio word. For the dictionary construction, we sampled $400 \times k$ MFCC feature vectors from the training data (this figure has experimentally given satisfactory results).

Once an audio vocabulary of size k (k = 1000 in this work) is built, each MFCC feature is assigned to the closest audio word in terms of Euclidean distance. Subsequently, a histogram is computed for each video shot extracted from movies in the training dataset and the related video shot is represented by a BoAW histogram representing the audio word occurrences.

2) Audio Representation by Sparse Coding: The construction of the SC-based audio dictionary is also illustrated in Figure 2. We employ the dictionary learning technique presented in [14]. The advantage of this technique is its scalability to very large datasets with millions of training samples which makes the technique well suited for our work. In order to learn the dictionary of size k (k = 1000 in this work) for sparse coding, $200 \times k$ MFCC feature vectors are sampled from the training data (this figure has experimentally given satisfactory results). In the coding phase, we construct the sparse representations of audio signals by using the LARS algorithm [15]. Given an audio signal and a dictionary, the LARS algorithm returns sparse representations for MFCC feature vectors. In order to generate the final sparse representation of video shots which are a set of MFCC feature vectors, we apply the *max-pooling* technique.

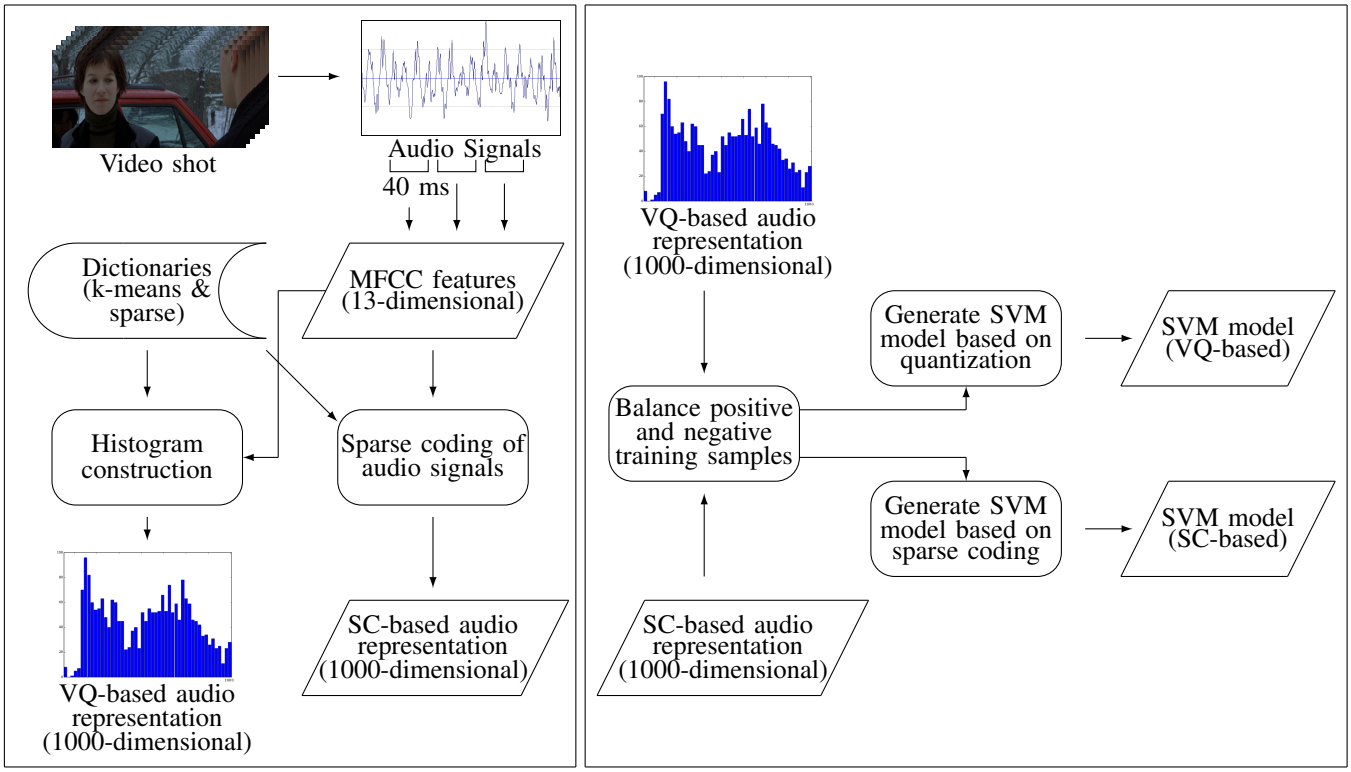

Fig. 1. (a) The generation process of vector quantization-based (VQ-based) and sparse coding-based (SC-based) audio representations for video shots of movies. (b) The learning phase of the method.

B. Violence Detection Model

We train a pair of two-class SVMs in order to learn violence models. One SVM model is constructed using VQ-based audio features, and the other model is constructed using SC-based audio features. In the learning step, the main issue to deal with is the problem of imbalanced data. In the training dataset, the number of non-violent video shots is much higher than the number of violent ones. This results in the learned boundary being too proximate to the violent instances. Consequently, the SVM shows the tendency to classify every sample as non-violent. Different strategies to "push" this decision boundary towards the non-violent samples exist. Although more sophisticated methods dealing with the imbalanced data issue have been proposed in the literature (see [16] for a comprehensive survey), we choose to perform random undersampling to balance the number of violent and non-violent samples in the current framework. The undersampling method which was proposed by Akbani et al. [17] appears to be particularly adapted to the application context of our work. In [17], different under- and oversampling strategies are compared on 10 different UCI datasets[1]. According to the results, SVM with undersampling strategy provides the most significant performance gain over standard two-class SVMs. In addition, the efficiency of the training process is improved as a result of the reduced training data and, hence, is scalable to large datasets comparable to the ones used in the context of our work.

[1] http://archive.ics.uci.edu/ml/

IV. Performance Evaluation

The experiments presented in this section aim at comparing the discriminative power of VQ- and SC-based mid-level audio representations for the detection of violent content in movies. We also compare our method with a baseline method provided by the MediaEval VSD task organizers [2] and the methods in the MediaEval VSD task which also stick to the same violence definition. Approaches discussed in Section II follow a different definition of "violence" such that a direct comparison could be misleading.

A. Dataset and Ground Truth

The dataset consists of 32.708 video shots from 18 Hollywood movies of different genres (ranging from extremely violent movies to movies without violence), where each video shot is labeled as violent or non-violent. The dataset is divided into a training set consisting of 26.138 video shots from 15 movies and a test set consisting of 6.570 video shots from the remaining 3 movies. Table I summarizes the main characteristics of the dataset in more detail. The movies of the training and test set were selected in such a manner that both training and test data contain movies of variable violence levels (extreme to none).

The ground truth for the dataset was generated by 7 human assessors. Violent movie segments are annotated at the frame level (i.e., violent segments are defined by their starting and ending frame numbers). In the dataset, automatically generated shot boundaries with their corresponding key frames are also

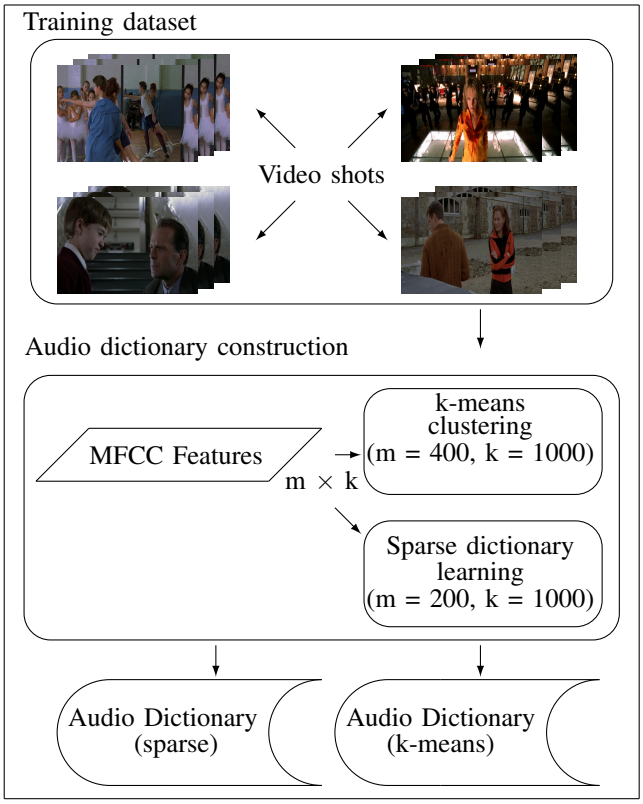

Fig. 2. The generation of two different audio dictionaries: one by using vector quantization and another dictionary for sparse coding.

TABLE I. THE CHARACTERISTICS OF TRAINING AND TEST DATASETS (THE NUMBER OF MOVIES AND VIDEO SHOTS, THE NUMBER AND PERCENTAGE OF VIOLENT AND NON-VIOLENT VIDEO SHOTS)

Dataset	Movies	Shots	Violent	Non-violent
Train	15	26.138	3.201 (12.25%)	22.937 (87.75%)
Test	3	6.570	715 (10.88%)	5.855 (89.12%)
Whole	18	32.708	3.916 (11.97%)	28.792 (88.03%)

provided for each movie. A detailed description of the dataset and the ground truth are given in [12].

B. Experimental Setup

We employed the MIR Toolbox v1.4[2] to extract the 13-dimensional MFCC features. Frame sizes of 40 ms without overlap are used to align with the 25-fps video frames. The features are extracted as explained in Section III.

We employed the SPAMS toolbox[3] in order to compute sparse codes which are used for the generation of SC-based mid-level audio representations.

We trained the two-class SVMs with a Radial Basis Function (RBF) kernel using libsvm[4] as the SVM implementation. Training was performed using audio features extracted at the video shot level. More specifically, we trained one SVM using VQ-based mid-level audio features and another SVM using SC-based mid-level audio features as input. SVM parameters were optimized by 5-fold cross-validation on the training data. Our approach was evaluated using a training-test split. In order to account for the problem of imbalanced training data, we performed undersampling by choosing random non-violent samples.

[2]https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

[3]http://spams-devel.gforge.inria.fr/

[4]http://www.csie.ntu.edu.tw/~cjlin/libsvm/

C. Evaluation

Precision and recall are metrics based on the results obtained for the whole list of video shots of the movies. Metrics other than precision and recall are, however, required to compare the performance of the VQ- and SC-based mid-level representations, since the ranking of violent shots is more important for our use case (i.e., providing a ranked list of violent video shots to the user). As evaluation metrics, therefore, we used *average precision at 20* and *100* which are also official metrics used in the MediaEval VSD task and *R-precision* which can be seen as an alternative to the *precision at k* metric in information retrieval. The values 20 and 100 for the computation of *average precision at k* are reasonable, since a user will only have a look at the video shots that are presented in the first few pages of the returned list.

D. Results and Discussions

Table II reports the *average precision at 100* values for the baseline method (i.e., random classification) and for our approach based on VQ- and SC-based mid-level audio representations. In Table II, *Dead Poets Society* is a movie in the test dataset having the lowest number of violent video shots (approximately 2% of all video shots within the movie). *Fight Club* is a movie in the test dataset having more violent video shots compared to *Dead Poets Society* (around 13% of all video shots). *Independence Day* is a movie having the most violent video shots in the test dataset (around 14.5% of all video shots). The results show that significant improvement is achieved with our approach compared to the baseline method in terms of *average precision at 100*.

TABLE II. AVERAGE PRECISION AT 100 FOR THE BASELINE AND OUR MID-LEVEL AUDIO METHODS (VQ: VECTOR QUANTIZATION, SC: SPARSE CODING)

Movie	Baseline	VQ-based Audio	SC-based Audio
Dead Poets Society	2.17%	15.6%	13.1%
Fight Club	13.27%	29.2%	41.03%
Independence Day	13.98%	72.2%	79.1%

Although currently we only exploit the audio and disregard the visual modality of videos to detect violent segments, the method where we represent video shots with VQ-based mid-level audio features (*VQ-based Audio* in Table II), manages to be in the top 35% of the methods in the MediaEval VSD task. The other method where we represent video shots with SC-based mid-level audio features (*SC-based Audio* in Table II), manages to be in the top 30% of the methods in the MediaEval VSD task. In addition, the *SC-based Audio* method (*SC-based Audio* in Table II), among approaches making use of only one modality (unimodal, i.e., either audio or visual taken alone), ranks second among 16 other unimodal submissions in the MediaEval VSD task in terms of average precision. Table III provides a comparison of our approach with the best run of participating teams in the MediaEval VSD task.

TABLE III. AVERAGE PRECISION (AP) AT 100 FOR THE BEST RUN OF TEAMS IN THE MEDIAEVAL VSD TASK AND OUR METHODS (VQ: VECTOR QUANTIZATION, SC: SPARSE CODING, SIFT: SCALE INVARIANT FEATURES TRANSFORM, STIP: SPATIAL-TEMPORAL INTEREST POINTS) [18]

Team	Features	Modality	Method	APat100
ARF	Color, texture, audio and concepts	audio-visual	Multi-layer perceptron	0.651
Shanghai-Hongkong	trajectory-based features, SIFT, STIP, MFCC	audio-visual	SVM with chi-squared kernel + temporal smoothing	0.624
TEC	color, motion, acoustic	audio-visual	Bayesian network with temporal integration post-processing	0.618
TUM	Acoustic energy and spectral, color, texture, optical flow	audio-visual	SVM with linear kernel	0.484
SC-based Audio (ours)	BoAW with sparse coding	audio	SVM with RBF kernel	**0.444**
VQ-based Audio (ours)	BoAW with vector quantization	audio	SVM with RBF kernel	**0.387**
LIG-MRIM	color, texture, bag of SIFT and MFCC	audio-visual	Fusion of SVMs and k-NNs with conceptual feedback	0.314
NII	Visual concepts learned from color and texture	visual	SVM with RBF kernel (with chi-square distance)	0.308
DYNI-LSIS	Multi-scale local binary pattern	visual	SVM with linear kernel	0.125

Table IV shows *average precision* (at 20 and 100) as well as *R-precision* for both SVMs with VQ- and SC-based mid-level audio features. From the results in Table IV, we observe that the SC-based mid-level representation provides more precise detections.

TABLE IV. AVERAGE PRECISION (AP) AT K (K = 20 AND 100) AND R-PRECISION (RP) ON THE TEST DATASET

Method	APat20	RPat20	APat100	RPat100
VQ-based Audio	0.489	0.445	0.387	0.355
SC-based Audio	0.537	0.483	0.444	0.366

In order to assess the performance of our VQ- and SC-based method in more detail, we also investigated the *average precision at k* (k = 20 and 100) and *R-precision* values for each movie in the test dataset. In Table V, these values are presented for *Independence Day* in the test set. We observe that both the SC-based method and the VQ-based one performs well, which demonstrates the potential of the mid-level audio representations.

TABLE V. AVERAGE PRECISION (AP) AT K (K = 20 AND 100) AND R-PRECISION (RP) ON *Independence Day*

Method	APat20	RPat20	APat100	RPat100
VQ-based Audio	1	0.907	0.722	0.616
SC-based Audio	0.938	0.925	0.791	0.712

Table VI presents the *average precision at 20* and *at 100* and the corresponding *R-precision* values for *Dead Poets Society*. The VQ-based method leads to slightly better results in terms of precision on this movie.

TABLE VI. AVERAGE PRECISION (AP) AT K (K = 20 AND 100) AND R-PRECISION (RP) ON *Dead Poets Society*

Method	APat20	RPat20	APat100	RPat100
VQ-based Audio	0.230	0.156	0.156	0.156
SC-based Audio	0.148	0.1	0.13	0.1

In Table VII, we provide the *average precision at 20* and *at 100* and the related *R-precision* values for *Fight Club*. Contrary to the results in Table VI, the VQ-based method is outperformed by the SC-based one.

TABLE VII. AVERAGE PRECISION (AP) AT K (K = 20 AND 100) AND R-PRECISION (RP) ON *Fight Club*

Method	APat20	RPat20	APat100	RPat100
VQ-based Audio	0.237	0.273	0.282	0.292
SC-based Audio	0.523	0.427	0.41	0.288

An illustration of the situations where our method performs well and the situations where it fails is provided in Figure 3 and Figure 4, respectively. These samples and their corresponding results demonstrate that our method is able to suitably detect violent content such as fights (e.g., Figure 3(a) showing the keyframe of a man being dragged on the floor by another man and yelling loudly) and disasters with explosions (e.g., Figure 3(b) showing the keyframe of an explosion inside a building). Video shots which contain no excitement, e.g., containing a man giving a speech (Figure 3(c)) or strong music in the background (Figure 3(d)) are also easily classified as non-violent.

On the other hand, the method wrongly classifies a video shot as violent when the video shot contains very strong sounds or exciting moments such as a plane taking off (Figure 4(a)) or loud ringing bells (Figure 4(b)). The most challenging violent video shots to be detected are the ones which are "violent" according to the definition of violence within the MediaEval VSD task, but actually only contain actions such as self-injuries, or other moderate actions such as an actor pushing or hitting slightly another actor (e.g., Figure 4(d)). Our method is also unable to detect violent video shots which are "violent" according to the definition of violence, but which contain no audio cues exploitable for the identification of violence (e.g., Figure 4(c) showing the keyframe of a man bleeding).

(a) (b) (c) (d)

Fig. 3. The keyframes of sample video shots from the test dataset which are *correctly* classified. Frame (a) represents a man being dragged and yelling (true positive), (b) an explosion inside a building (true positive), (c) a man giving a speech (true negative) and (d) strong background music (true negative)

One significant point which can be inferred from the overall results is that the average precision variation of the proposed method is high for movies of varying violence levels. Additionally, the method performs better when the violence level of a movie is higher. The difference between the results obtained from *Fight Club* and *Independence Day* in the test set (Tables V and VII) is most probably due to the nature

Fig. 4. The keyframes of sample video shots from the test dataset which are *wrongly* classified. Frame (a) represents a plane taking off (false positive) (b) loud ringing bells (false positive) (c) a man bleeding (false negative) and (d) a man pushing slightly another man (false negative)

of the violent content present in these movies. The violent actions present in *Fight Club* are under-represented in the training dataset and, consequently, no related audio word(s) could be extracted for these actions. In other words, violent content in *Fight Club* has no proper representation in terms of audio words.

To summarize, experiments presented in this section show that, on the one hand, the VQ- and SC-based mid-level audio representations provide promising results in terms of average precision, and, on the other hand, that the SC-based one outperforms both the VQ-based and the unimodal submissions in the MediaEval VSD task except one visual-based method in terms of average precision.

V. Conclusions and Future Work

In this paper, we presented an approach for movie violent content detection at video shot level. We employed mid-level audio features based on BoAW, where we first extract MFCC features and subsequently apply two coding schemes (i.e., vector quantization and sparse coding). We have shown that mid-level audio features provide promising results and that the sparse coding-based BoAW outperforms the unimodal submissions in the MediaEval VSD task except one visual-based method in terms of average precision. Incited by the promising results obtained for this work, we currently investigate the construction of more sophisticated mid-level representations for video content analysis. The current method only exploits the audio modality for content representation. An interesting research question is whether augmenting the feature set by including visual features (both low-level and mid-level ones) helps further improving classification. Hence, in future work, we plan to study the representation of videos with multi-modal features. In addition, we aim to extend our approach to user-generated videos. Different from Hollywood movies, these videos are not professionally edited, e.g., in order to enhance dramatic scenes. Thus, focusing on such content will shed light on the significance of actual sounds that are produced in real violent scenes such as street fights or explosions.

Acknowledgment

We would like to thank *Technicolor* (http://www.technicolor.com/) for providing the ground truth, video shot boundaries and the corresponding keyframes which have been used in this work.

References

[1] B. Bushman and L. Huesmann, "Short-term and long-term effects of violent media on aggression in children and adults," *Archives of Pediatrics & Adolescent Medicine*, vol. 160, no. 4, p. 348, 2006.

[2] (2012) Violence detection task. [Online]. Available: http://www.multimediaeval.org/mediaeval2012/violence2012/

[3] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 41, no. 6, pp. 797–819, 2011.

[4] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2559–2566.

[5] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features," *Advances in Artificial Intelligence*, pp. 502–507, 2006.

[6] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su, "Horror video scene recognition via multiple-instance learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011.

[7] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," *Advances in neural information processing systems*, vol. 15, pp. 561–568, 2002.

[8] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-visual fusion for detecting violent scenes in videos," *Artificial Intelligence: Theories, Models and Applications*, pp. 91–100, 2010.

[9] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Detecting violent scenes in movies by auditory and visual cues," *Advances in Multimedia Information Processing-PCM 2008*, pp. 317–326, 2008.

[10] R. Yan and M. Naphade, "Semi-supervised cross feature learning for semantic concept detection in videos," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 657–663.

[11] J. Lin and W. Wang, "Weakly-supervised violence detection in movies with audio and video based co-training," *Advances in Multimedia Information Processing-PCM 2009*, pp. 930–935, 2009.

[12] C. Demarty, C. Penet, G. Gravier, and M. Soleymani, "The MediaEval 2012 Affect Task: Violent Scenes Detection in Hollywood Movies," in *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.

[13] M. Xu, N. Maddage, C. Xu, M. Kankanhalli, and Q. Tian, "Creating audio keywords for event detection in soccer video," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 2. IEEE, 2003, pp. II–281.

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[15] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[16] H. He and E. Garcia, "Learning from imbalanced data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, no. 9, pp. 1263–1284, 2009.

[17] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," *Machine Learning: ECML 2004*, pp. 39–50, 2004.

[18] (2013) Mediaeval multimedia benchmark workshop, violent scenes detection task, 2012. [Online]. Available: http://ceur-ws.org/Vol-927/