Esra Acar, Melanie Irrgang, Dominique Maniry, Frank Hopfgartner

Detecting Violent Content in Hollywood Movies and User-Generated Videos

Open Access via institutional repository of Technische Universität Berlin

Document type

Book chapter | Accepted version

(i. e. final author-created version that incorporates referee comments and is the version accepted for publication; also known as: Author's Accepted Manuscript (AAM), Final Draft, Postprint)

This version is available at

<https://doi.org/10.14279/depositonce-12046>

Citation details

Acar, E., Irrgang, M., Maniry, D., & Hopfgartner, F. (2015). Detecting Violent Content in Hollywood Movies and User-Generated Videos. In: Hopfgartner F. (ed.) Smart Information Systems (pp. 291–314). Advances in Computer Vision and Pattern Recognition. Springer International Publishing.
https://doi.org/10.1007/978-3-319-14178-7_11.

Terms of use

This work is protected by copyright and/or related rights. You are free to use this work in any way permitted by the copyright and related rights legislation that applies to your usage. For other uses, you must obtain permission from the rights-holder(s).

Chapter 11

Detecting Violent Content in Hollywood Movies and User-Generated Videos

Esra Acar, Melanie Irrgang, Dominique Maniry, and Frank Hopfgartner

Abstract Detecting violent scenes in videos is an important content understanding functionality e.g., for providing automated youth protection services. The key issues in designing violence detection algorithms are the choice of discriminative features and learning effective models. We employ low- and mid-level audio-visual features and evaluate their discriminative power within the context of the MediaEval Violent Scenes Detection (VSD) task. The audio-visual cues are fused at the decision level. As audio features, Mel-Frequency Cepstral Coefficients (MFCC), and as visual features dense histogram of oriented gradient (HoG), histogram of oriented optical flow (HoF), Violent Flows (ViF) and affect-related color descriptors are used. We perform feature space partitioning of the violence training samples through k -means clustering and train a different model for each cluster. These models are then used to predict the violence level of videos by employing two-class support vector machines (SVMs). The experimental results in Hollywood movies and short web videos show that mid-level audio features are more discriminative than the visual features, and that the performance is further enhanced by fusing the audio-visual cues at the decision-level.

Esra Acar
Technische Universität Berlin, e-mail: esra.acar@tu-berlin.de

Melanie Irrgang
Technische Universität Berlin, e-mail: melanie.irrgang@campus.tu-berlin.de

Dominique Maniry
Technische Universität Berlin, e-mail: dmaniry@cs.tu-berlin.de

Frank Hopfgartner
Technische Universität Berlin, e-mail: frank.hopfgartner@tu-berlin.de

Babysitting

Nowadays our children are submerged by connected equipment, whether this is at school, at home or even in the car. Notable examples of such equipment include TV, cable or satellite set-top boxes, tablets or the smartphones of the parents, when those let their children play with it.

Think of tablets. Children use it for learning to read or to count; finding where “Wally” is; watching funny videos on streaming websites. And for TVs, Video-On-Demand (VOD) services such as Netflix, Hulu or Amazon Prime Instant, allow them to watch all sorts of video content, including not only “educational” content, but also – and children of course are not aware of it – content which is harmful for them. Parents can always try to track these “unwanted” contents. But sometimes, when they are out of home, they have to rely on babysitters for taking care of their children.

Occasionally, Clara works as a babysitter. A conscientious one! Not the sort of babysitter who would put a child in front of the TV or in front of an iPad and focus on talking for hours on the phone with her friends until the parents finally come back and pay her.

Today, Saturday, Clara is babysitting Ben, the 9-year old child of her neighbors David and Rose. These are celebrating the promotion which Rose just obtained, so they invited many guests in a restaurant, where the parents of Clara are also present.

The dinner is over. “Hey, what would you like to do, Ben?”, asks Clara. Ben is not very talkative. He points to the TV with his finger. “Aha, so you want to watch TV, huh? Okay.” adds Clara. Ben noticed that the famous ’90 movie *Home Alone* which

he heard of recently, is aired on the VOD service. This is a well-known children comedy movie starring child Hollywood star Macaulay Culkin.

Clara then says: “Hey buddy, I know that movie very well, you know! I agree to watch it with you, but only if we skip the bad scenes, the scenes which are not good for a cute child like you!” She remembers that some scenes are very funny but also very violent. Especially the scenes where the burglars first attempt to enter the house. They got badly shot by a toy rifle. Then, one of them gets his scalp burned... Or inside the house, when one of them is hit by a can of paint, or when he walks barefoot on crushed glass and gets hurt. Really funny, but really violent...

Now, with the violent movie content detector plug-in, finding and skipping violent parts has never been made so easy. Clara just has to turn on the violence detection plug-in on the Smart TV and quickly browse the movie before letting Ben watch it. Thanks to the plug-in, she can easily recognize violent parts while watching the movie and jump to the next non-violent scene when necessary. Before a violence detector could offer such possibilities, Smart TV users were provided information mainly based on the Electronic Program guide (EPG).

In their most common forms, EPG information contains a short description of the TV program and a parental guidance (PG) rating for the whole program (referring to the recommended minimal age of the spectator). Unlike an EPG, the plug-in works by analyzing directly the audio and visual content to discover scenes which a spectator would consider as violent. The advantage is that, instead of discarding a whole movie based solely on a PG rate, the plug-in can label the movie scene-by-scene or even shot-by-shot.

At the end of the movie, Clara noticed that the plug-in was really helpful. She thinks, “Wow, almost all violent scenes were found by that app. The kid did not see any of those violent scenes”.

Clara checks what time it is: “Only 7? What else am I gonna do with the kid until the parents are back?” Indeed, Ben is allowed to stay awake until 9 pm on Saturday evenings. He likes to play games on the tablet – crushing objects, throwing birds away and many other games – or watch funny clips on Youtube or Dailymotion – videos sent by his schoolmates via popular social networks. Clara knows that the tablet belongs to the parents and that they do not have any parental control filter installed on it. Therefore, she is not very much pleased with the idea of tablet. However, something suddenly pops in her mind. “Oh yes, that violence detection plug-in on the TV has an equivalent application for the tablet. I remember that it can also be downloaded for the tablet from the Internet”, she remembers.

She installs it, sets violent scene detection on and hands the tablet to Ben. Convinced by the performance of the plug-in on the TV, she is confident that Ben risks nothing when such a feature is turned on. He can play with it until it's time to go to bed...

11.1 Introduction

As the amount of available multimedia content becomes more and more abundant, the use of automatic multimedia analysis solutions in order to find relevant semantic search results or to identify illegal content present on the World Wide Web has reached a critical importance. In addition, the advances in digital media management techniques have facilitated delivering digital videos to consumers. Therefore, accessing online movies through services such as VOD has become extremely easy. As a result, parents are not able to constantly and precisely monitor what their children watch. Children are, consequently, exposed to movies, documentaries, or reality shows which have not necessarily been checked by parents, and which might contain inappropriate content. Violence constitutes one example of such inappropriate content. Psychological studies have shown that violent content in movies has harmful impacts, especially on children [4]. As a consequence, there is a need for automatically detecting violent scenes in videos, where the legal age ratings are not available.

Like any other research challenge, tackling the problem of violence detection begins with establishing a framework, in particular adopting a definition of violence to work with. Since the concept of violence is highly subjective (i.e., person-dependent) – not everybody would indeed evaluate a particular scene of a movie as violent, one of the challenges within the context of multimedia violent content detection is to properly delimit the boundaries of what can be designated as a “violent” scene. In our work, we aim at sticking to the two definitions of violence as described in [8]: the objective and subjective points of view. *Objective violence* is defined as “physical violence or accident resulting in human injury or pain”, whereas *subjective violent* scenes are “those which one would not let an 8 years old child see because they contain physical violence”.

In this context, the MediaEval Violent Scene Detection (VSD) task [9] provides a consistent evaluation framework to the research community and enables different approaches to be evaluated by using the same violence definitions and a standardized annotated dataset. Detailed description of the task, the dataset, the ground truth and evaluation criteria are given in the paper by Demarty et al. [9]. The task ascribes to a use case from the company *Technicolor*¹. The French producer of video content and entertainment technologies adopted the aim of helping users to select movies that are suitable to watch with their children.

In spite of the existence of institutions, the task of which is to assign a recommended age to movies in France, the ratings determined by those institutions are not as strict and differentiated as the ones from the German FSK (*Freiwillige Selbstkontrolle der Filmwirtschaft*). One explanation of the sources of discrepancies is the fact that employees from the film making industry are allowed to participate in the recommendation process in France.

¹ <https://research.technicolor.com/rennes/>

A lot of movies labeled as FSK 16 (i.e., recommended for an audience of age higher than 16 years old) in Germany are released without restrictions in France². There also is no equivalence for the FSK 6 label in France, where movies are recommended only for the age of 0, 12, 16 and 18³. This seems to be the main motivation behind the introduction of the challenge, for which one additional sub-task in 2013 is to detect scenes with subjective violence. Another illustration of differences between countries is the age rating labels used in the USA, where one example of label is “NC-17”. The latter does not mean that audience should be at least 17 years old, but that audience should not be 17 or under 17, while, for instance, FSK 16 means audience should be at least 16. This can also be a source of confusion among consumers.

The degree of violence one is able or willing to bear might vary strongly even within a group of persons of identical age. That is probably why parents should get from *Technicolor* information which is not limited to rating only but also a preview of the most violent scenes, in order to help them decide if the movie is adequate to be watched by their child.

Next to the issue of definition, another important step in the task of movie violent content detection is the representation of movie segments. Many of the existing works (e.g., [5, 14]) proposed for violence detection represent videos using low-level representations, especially for the representation of audio signals. Inferring abstract representations is more suitable than directly using low-level features in order to bridge the semantic gap between the features and high-level human perception of violence. However, high-level semantics are more difficult to detect and state-of-the-art detectors are far from perfect. Therefore, the use of mid-level representations may help modeling video segments one step closer to human perception.

This chapter aims at assessing the discriminative power of mid-level audio-visual features to model violence in Hollywood movies. We also investigate the effects of combining mid-level audio-visual features with low-level audio-visual features for the detection of violent content and show that promising results are obtained by fusing audio-visual cues at the decision level.

The chapter is organized as follows. Section 11.2 explores the recent developments and reviews methods which have been proposed in the literature in order to detect violence in movies. In Section 11.3, we introduce our method and the functioning of its various components. We provide and discuss evaluation results obtained on Hollywood movies in Section 11.4. In Section 11.5, we present our browser-based visualization tool which provides an intuitive way of using our solution. Concluding remarks and future directions to expand our current approach are presented in Section 11.6.

² <http://fsf.de/jugendmedienschutz/international/filmfreigaben/>

³ <http://www.fsk.de/>

11.2 Related Work

Although video content analysis has been studied extensively in the literature, violence analysis of movies or of user-generated videos is restricted to a few studies only. We discuss here some of the most representative ones which use audio and/or visual cues. A difficulty arises regarding the definition of violence. In some of the works presented in this section, the authors do not explicitly state their definition of violence. In addition, nearly all papers in which the concept is defined consider a different definition of violence; therefore, whenever possible, we also specify the definition adopted in each work discussed in this section.

First, we briefly discuss uni-modal (i.e., based exclusively on the audio or visual modality) violence detection methods. Giannakopoulos et al. [13] 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 (ZCR), spectral flux and roll-off are employed. A polynomial SVM is used as the classifier. The main issue with 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 and speech).

In their paper [7], de Souza et al., similarly to other works related to violence detection, adopt their own definition of violence, and designate violent scenes as those containing fights (i.e., aggressive human actions), regardless of the context and the number of people involved. Their approach is based on the use of Bag-of-Words (BoW), where local Spatial-Temporal Interest Point Features (STIP) are used as the feature representation of video shots. They compare the performance of STIP-based BoW with SIFT-based BoW on their own dataset, which contains 400 videos (200 violent and 200 non-violent videos). The STIP-based BoW solution has proven to be superior to the SIFT-based one.

Hassner et al. [18] present a method for real-time detection of breaking violence in crowded scenes. They define violence as sudden changes in motion in a video footage. The method considers statistics of magnitude changes of flow-vectors over time. These statistics, collected for short frame sequences, are represented using the Violent Flows (ViF) descriptor. ViF descriptors are then classified as either violent or non-violent using a linear SVM. The authors also introduce a new dataset of crowded scenes on which their method is evaluated. According to the presented results, the ViF descriptor outperforms the Local Trinary Patterns (LTP) [38], histogram of oriented gradient (HoG) [23], histogram of oriented optical flow (HoF) [23] descriptors as well as the histogram of oriented gradient and optical flow (HNF) descriptor [23]. The ViF descriptor is also evaluated on well-known datasets of videos of non-crowded scene such as the Hockey dataset [28] and the ASLAN dataset [22] in order to assess its performance in action-classification tasks of “non-textured” videos (i.e., non-crowded). With small vocabularies, the ViF descriptor outperforms the LTP and STIP descriptors, while with larger vocabularies,

STIP outperforms ViF. However, this performance gain comes with a higher computational cost.

In [35], Xu et al. propose to use Motion SIFT (MoSIFT) descriptors to extract a low-level representation of a video. Feature selection is applied on the MoSIFT descriptors using kernel density estimation. The selected features are subsequently summarized into a mid-level feature representation based on a BoW model using sparse coding. The method is evaluated on two different types of datasets: crowded scenes [18] and non-crowded scenes [28]. Although Xu et al. do not explicitly define violence, they study fights or sudden changes in motion as violence-related concepts. The results show that the proposed method is promising and outperforms HoG-based and HoF-based BoW representations on both datasets.

Second, we review multi-modal methods, which constitute the most common type of approach used in violent content detection in videos, and which consist in fusing audio and visual cues at either feature or decision level. Aiming at detecting horror, Wang et al. [5] apply Multiple Instance Learning (MIL; MI-SVM [3]) using color, textual and MFCC features. The authors do not explicitly state their definition of horror. Therefore, assessing the performance of their method and identifying the situations on which it properly works is difficult. 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 MFCCs are used for the audio representation. More specifically, mean, variance and first-order differential of each dimension of MFCCs are employed for the audio representation. As observed from their results [5], using color and textural information in addition to MFCC features slightly improves the performance.

Giannakopoulos et al. [14], in an attempt to extend their approach based solely on audio cues [13], 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 and MFCCs 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 [15], a three-stage method is proposed. In the first stage, the authors apply a semi-supervised cross-feature learning algorithm [37] on the extracted audio-visual features such as motion activity, ZCR, MFCCs, pitch and 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. Although not explicitly stated, the authors define violent scenes as those which contain action and violence-related concepts such as gunshots, explosions and screams. 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 and Wang [24] 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, MFCCs, 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 signal 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. Violence-related concepts studied in this work are fights, murders, gunshots and explosions. This method was also evaluated only on action movies. Therefore, the performance of this solution in genres other than action is uncertain.

Chen et al. [6] proposed a two-phase solution. According to their violence definition, a violent scene is a scene that contains action and blood. In the first phase, where average motion, camera motion, and average shot length are used for scene representation and SVM for classification, video scenes are classified into action and non-action. In the second phase, using the “Viola-Jones” face detector, faces are detected in each keyframe of action scenes and the presence of blood pixels near detected human faces is checked using color information. The approach is compared with the method of Lin and Wang [24] because of the similar violence definitions, and is shown to perform better in terms of precision and recall.

Ding et al. [11] observe that most existing methods identify horror scenes only from independent frames, ignoring the context cues among frames in a video scene. In order to consider contextual cues in horror scene recognition, they propose a Multi-view MIL (M^2IL) model based on a joint sparse coding technique which simultaneously takes into account the bag of instances from the independent view and from the contextual view. Their definition of violence is very similar to the definition in [5]. They perform experiments on a horror video dataset collected from the Internet and the results demonstrate that the performance of the proposed method is superior to other existing well-known MIL algorithms.

The works discussed in the following paragraphs employ the same definitions of violence (i.e., *objective* and/or *subjective* violence) adopted in the MediaEval 2013 VSD task. Penet et al. [29] propose to exploit temporal and multi-modal information for objective violence detection at video shot level. In order to model violence, different kinds of Bayesian network structure learning algorithms are investigated. The proposed method is tested on the dataset of the MediaEval 2011 VSD Task. Experiments demonstrate that both multimodality and temporality add valuable information into the system and improve the performance in terms of MediaEval cost

function [9]. The best performing method achieves 50% false alarms and 3% missed detection, ranking among the best submissions to the MediaEval 2011 VSD task.

In [21], Ionescu et al. address the detection of objective violence in Hollywood movies. The method relies on fusing mid-level concept predictions inferred from low-level features. The mid-level concepts used in this work are gory scenes, presence of blood, firearms and cold weapons (for the visual modality); presence of screams and gunshots (for the audio modality); and car chases, presence of explosions, fights and fire (for the audio-visual modalities). The authors employ a bank of multi-layer perceptrons featuring a dropout training scheme in order to construct 10 violence-related concept classifiers. The predictions of these concept classifiers are then merged to construct the final violence classifier. The method is tested on the dataset of the MediaEval 2012 VSD task and ranked first among 34 other submissions, in terms of precision and F-measure.

In [16], Goto and Aoki propose a violence detection method which is based on the combination of visual and audio features extracted at the segment level, using machine learning techniques. Violence detection models are learned via multiple kernel learning. The authors also propose mid-level violence clustering in order to implicitly learn mid-level concepts without using manual annotations. The proposed method is trained and evaluated on the MediaEval 2013 VSD task using the official MediaEval metric Mean Average Precision at 100 (MAP@100). The results show that the method outperforms the approaches which use no external data (e.g., Internet resources) in the MediaEval 2013 VSD task.

Derbas and Quénot [10] explore the joint dependence of audio and visual features for violent scene detection. They first combine the audio and the visual features and then determine statistically joint multi-modal patterns. The proposed method mainly relies on an audio-visual BoW representation. The experiments are performed in the context of the MediaEval 2013 VSD task. The obtained results show the potential of the proposed approach in comparison to methods which use audio and visual features separately, and to other fusion methods such as early and late fusion.

11.3 The Violence Detection Method

In this section, we discuss (1) the representation of video segments, and (2) the learning of a violence model, which are the two main components of our method.

11.3.1 Video Representation

Sound effects and background music in movies are essential for stimulating people's perception [33]. Therefore, the audio signals are important for the representation of videos. Visual content of videos provides complementary information for the detection of violence in videos. We represent the audio content using mid-level represen-

Fig. 11.1: The generation process of SC-based audio and visual representations for video segments. Each video segment is of length 0.6 second. Separate dictionaries are constructed and used for MFCC, HoG and HoF to generate 1024-dimensional representations. Each HoG and HoF descriptor is 144-dimensional.

tations, whereas the visual content is represented at two different levels: low-level and mid-level.

11.3.1.1 Mid-level Audio Representation

Mid-level audio representations are based on MFCC features extracted from the audio signals of video segments of 0.6 second length as illustrated in Figure 11.1. In order to generate the mid-level representations for video segments, we apply an abstraction process which uses an MFCC-based Bag-of-Audio Words (BoAW) approach with sparse coding (SC) as the coding scheme.

The construction of the SC-based audio dictionary is illustrated in Figure 11.2. We employ the dictionary learning technique presented in [26]. The advantage of this technique is its scalability to very large datasets containing millions of training samples which makes the technique well suited for our work. In order to learn the dictionary of size k ($k = 1024$ in this work) for sparse coding, $400 \times k$ MFCC feature vectors are sampled from the training data (experimentally determined figure). In the coding phase, we construct the sparse representations of audio signals by using the LARS algorithm [12]. 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 segments, which is a set of MFCC feature vectors, we apply the *max-pooling* technique.

Fig. 11.2: The generation of audio and visual dictionaries with sparse coding. Each video segment is of length 0.6 second. Low-level features are MFCCs, densely sampled HoG and HoF descriptors.

11.3.1.2 Low-level Visual Representation

Film-makers usually make use of motion in order to elicit some particular perception in the audience [33]. Therefore, we use motion-related descriptors for the visual representation of video segments. One of the motion descriptors is ViF which is an efficient motion descriptor. We computed a ViF descriptor for each video segment to represent statistics of flow-vector magnitude changes over time. For a detailed explanation of the computation of this descriptor, the reader is referred to [18].

In addition to motion information, static content of video frames is also important for evoking some particular perception in the audience [33]. We, therefore, also use static content representations in our work. More specifically, we employ affect-related static visual descriptors. Inspired by the work presented in [25], we compute mean and standard deviation of saturation, brightness and hue in the HSL color space. We also compute the colorfulness of the keyframe of video segments using the method in [17], where the keyframe is deemed to be the frame in the middle of a video segment.

11.3.1.3 Mid-level Visual Representation

Mid-level visual representations are based on HoG and HoF features extracted from the visual content of video segments of 0.6 second length. HoG and HoF descriptors are densely sampled and computed for subvolumes of video segments (HoG descriptors are subsampled every 6 frames and HoF descriptors are subsampled every 2 frames as recommended in [32]). The Horn-Schunck method [20] is applied to compute optical flow vectors which are used for the extraction of HoF descriptors. The resulting HoG and HoF descriptors are subsequently used to generate mid-level HoG and HoF representations separately, which is illustrated in Figure 11.1. The construction of the SC-based HoG and HoF dictionaries is illustrated in Figure 11.2.

11.3.2 Violence Detection Model

“Violence” is a concept, which can be expressed in diverse manners. For instance, both explosions and scream scenes are labeled as violent according to the definition that we adopted. However, these scenes might highly differ from each other in terms of audio-visual appearance depending on their characteristics of violence. Therefore, instead of learning a unique model for violence detection, learning multiple models constitutes a more judicious choice. This justifies that we first perform feature space partitioning by clustering video segments of 0.6 second length in our training dataset and learn a different model for each violence sub-concept (i.e., cluster). We use two-class SVMs in order to learn violence models. An overview of the generation of violence detection models is presented in Figure 11.3.

In the learning step, the main issue is the problem of imbalanced data. This is caused by the fact that, in the training dataset, the number of non-violent video shots is much higher than the number of violent ones. This phenomenon results in the learned boundary being too close to the violent instances. Consequently, the SVM tends 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 [19] for a comprehensive survey), we choose, in the current framework, to perform random undersampling to balance the number of violent and non-violent samples (with a balance ratio of 1:2). This method proposed by Akbani et al. [2] appears to be particularly adapted to the application context of our work. In [2], different under- and oversampling strategies are compared. According to the results, SVM with the 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, training is easily scalable to large datasets similar to the ones used in the context of our work.

In the test phase, the main challenge is to combine the classification results of the violence models. We perform a classifier selection to solve this. More specifically, we first determine the nearest cluster to a video segment of the test set using Eu-

Fig. 11.3: The generation of violence detection models with feature space partitioning.

Fig. 11.4: An overview of the test phase of our classification method. Each video segment is of length 0.6 second.

clidean distance measures. Once the classifier for the video sample is determined, the output of the chosen model is used as the final prediction for that video sample. An overview of the test phase of our method is presented in Figure 11.4.

11.4 Performance Evaluation

The experiments presented in this section aim at comparing the discriminative power of low- and mid-level audio-visual representations and feature space partitioning through clustering. A direct comparison of our results with other works discussed in Section 11.2 is not straightforward due to the differences in the definition of “violence” in published works. However, we compare our method with one of the methods in the MediaEval VSD task of 2013 which also sticks to the same “violence” definition and provides evaluation results at the video segment level.

11.4.1 Dataset and Ground Truth

We used two different types of dataset in our experiments: (1) a set of 24 Hollywood movies which were the movies of the MediaEval 2013 VSD task (the “Hollywood movie dataset”), and (2) a set of 86 short YouTube web videos under Creative Commons licenses which were the short web videos of the MediaEval 2014 VSD task (the “Web video dataset”). The 24 movies of the Hollywood movie dataset are from different genres (ranging from extremely violent movies to movies without violence). Each movie is split in a multitude of video segments, where each video segment is of length 0.6 second. In total, the Hollywood movie dataset consists of 289,699 video segments, where each video segment is labeled as violent or non-violent.

A total of 17 movies from the Hollywood set are dedicated to the training process: *Armageddon*, *Billy Elliot*, *Eragon*, *Harry Potter 5*, *I am Legend*, *Leon*, *Midnight Express*, *Pirates of the Caribbean 1*, *Reservoir Dogs*, *Saving Private Ryan*, *The Sixth Sense*, *The Wicker Man*, *The Bourne Identity*, *The Wizard of Oz*, *Dead Poets Society*, *Fight Club* and *Independence Day*. The remaining 7 movies – *Fantastic Four 1*, *Fargo*, *Forrest Gump*, *Legally Blond*, *Pulp Fiction*, *The God Father 1* and *The Pianist* – serve as the test set for the main task which is to detect violence in Hollywood movies. In terms of number of video segments, the training set (17 movies) consists of 201,216 video segments and the test set (7 movies) consists of 88,483 video segments. Table 11.1 presents the main characteristics of the dataset in more detail. The movies of the training and test sets were selected in such a manner that both training and test data contain movies of variable violence levels (extreme to none). On average, in both datasets, around 12.5% of segments are annotated as violent.

The ground truth of the Hollywood dataset was generated by 9 human assessors, partly by developers, partly by potential users. Violent movie segments are annotated at the frame level. Automatically generated shot boundaries with their corresponding key frames are also provided for each movie. A detailed description of the Hollywood dataset and the ground truth generation are given in [9]. For the generalization task which is to detect violence in short web videos, the ground truth was

Table 11.1: The characteristics of training and test movies of the Hollywood movie dataset (The number of movies and video segments, the number and percentage of violent and non-violent video segments).

Dataset	Movies	Video Segments	Violent	Non-violent
<i>Train</i>	17	201,216	24,517 (12%)	176,699 (88%)
<i>Test</i>	7	88,483	11,594 (13%)	76,889 (87%)
<i>Total</i>	24	289,699	36,111 (12.5%)	253,588 (87.5%)

created by several human assessors⁴ who followed the subjective definition of violence as explained in Section 11.1. A detailed description of the Web video dataset and the ground truth generation are given in [31].

11.4.2 Experimental Setup

We employed the MIR Toolbox v1.4⁵ to extract the MFCC features (13-dimensional). Frame sizes of 40 ms without overlap are used to align with the 25-fps frames. The Matlab toolbox⁶ provided by Uijlings et al. [32] was used to extract dense HoG and HoF features. Features are extracted as explained in Section 11.3.

We employed the SPAMS toolbox⁷ in order to compute sparse codes which are used for the generation of the mid-level audio and visual representations. The VLFeat⁸ open source library is used to perform k -means clustering ($k = 10$ in this work).

We trained the two-class SVMs with an RBF kernel using libsvm⁹ as the SVM implementation. Training was performed using audio and visual features extracted at the video segment level. SVM parameters were optimized by 5-fold cross-validation on the training data.

11.4.3 Evaluation Metrics

We used two different evaluation metrics in our evaluation: (1) *average precision at 100 (AP@100)* which is the official metric used in the MediaEval 2013 VSD task, and (2) *average precision (AP)* which is the official metric used in the MediaEval

⁴ Annotations were made available by *Fudan University*, *Vietnam University of Science*, and *Technicolor*.

⁵ <https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/>

⁶ <http://homepages.inf.ed.ac.uk/juijling/index.php?page=software/>

⁷ <http://spams-devel.gforge.inria.fr/>

⁸ <http://www.vlfeat.org/>

⁹ <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

2014 VSD task. Although the AP@100 metric is no longer the official metric of the MediaEval VSD task, we think that providing a ranked list of violent video shots to the user is still important for our use case. Additionally, including the AP@100 metric allows a comparison with potential other works which would present their results based on AP@100 solely.

11.4.4 Results and Discussion

Table 11.2 reports the mean AP and AP@100 metrics on the Hollywood movie dataset. We observe that the mid-level audio representation based on MFCC and sparse coding provides promising performance in terms of average precision and outperforms all other representations that we use in this work. We also note that the performance is further improved by fusing these mid-level audio cues with low- and mid-level visual cues at the decision level by linear fusion.

Table 11.2: The mean Average Precision (MAP) and MAP@100 of our method with different video representations (i.e., *audio-only* where we use MFCC-based mid-level audio representations, *visual-only* where HoG- and HoF-based mid-level features and ViF descriptors are used, *affect-only* where we use affect-related color features, and *multi-modal* where the previous three representations are linearly fused at the decision level) on the Hollywood movie dataset.

Dataset	MAP	MAP@100
<i>Audio-only</i>	0.363	0.476
<i>Visual-only</i>	0.327	0.439
<i>Affect-only</i>	0.209	0.140
<i>Multi-modal</i>	0.422	0.539

Table 11.3 reports the mean AP and AP@100 metrics on the Web video dataset. We observe results which are similar to the ones obtained on the Hollywood movie dataset (Table 11.2). We used the same violence detection models which were trained using the 17 Hollywood movies, and evaluated these models on short Web videos. The results in terms of AP and AP@100 are still encouraging and even demonstrate superior results compared to the ones obtained on the Hollywood movie dataset. Therefore, we can conclude that our violence detection method generalizes particularly well to other types of video content not used for training the models. Another interesting observation is that affect-related color features seem to provide better results in terms of AP metrics on the Web video dataset in comparison to the Hollywood movie dataset. One final remark is that the linear fusion of audio-visual features does not help improving the performance in the case of short web videos. This might be the consequence of the weights used in the fusion process, as more weight is given to *visual-only* representations. On the contrary, affect-related features perform better in web videos.

Table 11.3: The mean Average Precision (MAP) and MAP@100 of our method with different video representations (i.e., *audio-only* where we use MFCC-based mid-level audio representations, *visual-only* where HoG- and HoF-based mid-level features and ViF descriptors are used, *affect-only* where we use affect-related color features, and *multi-modal* where the previous three representations are linearly fused at the decision level) on the Web video dataset.

Dataset	MAP	MAP@100
<i>Audio-only</i>	0.582	0.582
<i>Visual-only</i>	0.478	0.478
<i>Affect-only</i>	0.495	0.495
<i>Multi-modal</i>	0.567	0.567

In order to allow the interested reader to have a closer look at the evaluation results, we provide movie-specific MAP@100 values of our best performing method which is the one with the multi-modal representation (Table 11.4). One significant point which can be inferred from the overall results is that the average precision variation of the method is high for movies of varying violence levels.

Table 11.4: The movie-specific MAP@100 values of our best performing method (i.e., the one with a *multi-modal* representation) on the Hollywood movie dataset. Each video segment is of length 0.6 second.

Movie Name	MAP@100	# of Violent Video Segments
<i>Fantastic Four 1</i>	0.615	2102
<i>Fargo</i>	0.416	1426
<i>Forrest Gump</i>	0.548	1139
<i>Legally Blond</i>	0.000	0
<i>Pulp Fiction</i>	0.992	3720
<i>The God Father 1</i>	0.290	987
<i>The Pianist</i>	0.910	2220

In Figure 11.5, the precision-recall (PR) curves of our method with different video representations are provided. As seen from the resulting PR curves, our method performs better on short web videos (Figure 11.5b).

Table 11.5 provides a comparison of our best performing method (i.e., the one with a *multi-modal* representation) in terms of MAP and MAP@100 metrics with the method introduced in [30] and an SVM-based unique violence detection model (i.e., a model where no feature space partitioning is performed). We can conclude that our method provides promising results and more importantly, outperforms the SVM-based detection method where the feature space is not partitioned and all violent and non-violent samples are used to build a unique model.

Finally, we can observe from the overall results provided in this section that our method performs better when violent content is better expressed in terms of audio features (a typical example would be a gun shot scene). This is an indication that

Fig. 11.5: Precision-Recall curves (a) on the Hollywood movie dataset (Area Under Curve (AUC) for Main1: 0.2985, Main2: 0.2757, Main3: 0.2066, and Main4: 0.3255), and (b) on the Web video dataset (AUC for Gen1: 0.5606, Gen2: 0.4447, Gen3: 0.4500, and Gen4: 0.5571) using different representations (Main/Gen1: *audio-only*, Main/Gen2: *visual-only*, Main/Gen3: *affect-only* and Main/Gen4: *multi-modal*).

Table 11.5: The mean Average Precision (MAP) and MAP@100 of our best performing method (i.e., the one with a *multi-modal* representation), the work of Penet et al. [30] and an SVM-based unique violence detection model (i.e., no feature space partitioning) on the Hollywood movie dataset.

Method	MAP	MAP@100
<i>Our method (multi-modal)</i>	0.422	0.539
<i>Penet et al. [30]</i>	0.353	0.448
<i>SVM-based unique violence detection model</i>	0.257	0.356

we need more discriminative visual representations for detecting violent content in movies and short web videos to further improve the performance of our method.

11.5 Application

We present in this section a browser-based visualization tool that allows users to explore movies and online videos based on the detected violence levels. In this tool, currently, only the objective violence definition of the MediaEval VSD task is used to model violence. The system offers the visualization of annotations and results of the MediaEval 2012 VSD task [9] and can interactively download and analyze content from video hosting sites such as YouTube.

The development and evaluation of VSD creates the need for a detailed visualization to assess the strengths and weaknesses of algorithms. Our visualization tool

[27] consists of three parts: the *Ranked List* view shows the results on the test set of the MediaEval 2012 VSD task, the *Annotations* view shows the annotations of the MediaEval 2012 VSD training set and the *Online Analysis* carries out our analysis pipeline [1] to arbitrary online videos.

11.5.1 The Method

Among the plurality of audio features, MFCCs are shown to be indicators of the excitement level of video segments [36]. Therefore, we employ them as low-level audio features. For the representation of video segments, we use mid-level audio features based on MFCCs in a BoAW scheme. We apply the BoAW approach with two different coding schemes; as an alternative to sparse coding (introduced in Section 11.3), we also carried out vector quantization. We train a pair of two-class SVMs in order to learn violence models using both mid-level feature representations. Normally, in a basic SVM, only class labels or scores are output. The class label results from thresholding the score, which is not necessarily a probability measure. The scores output by the SVM are converted into probability estimates using the method explained in [34].

11.5.2 Ranked List

The user first selects the run (algorithm and parameters), of which the results will be visualized. The user can also select a specific test movie or the whole test set. The *Ranked List* view (Figure 11.6) then shows the thumbnails of all segments with an overlay of the violence score (i.e., the probability of violence), time information and a notice whether the classification matches the ground truth. If a segment is classified as violent, the thumbnail is highlighted with an orange frame around it. This enables the user to interpret the results easily and quickly. A click on the thumbnail plays the given segment without leaving the *Ranked List* view. The user can sort the list by the violence scores returned by the algorithm, or can sort it by time to see the classification results chronologically from the beginning to the end of the movie. We also added a button which, when pressed, jumps to a random part of the list to enable a more dynamic exploration experience.

11.5.3 Annotations

The training set of the MediaEval 2012 VSD task provides annotations for 15 Hollywood movies. The annotations mark the presence of audio, visual and audio-visual concepts such as explosions, gunshots, screams, blood, fights, car chases,

Violent Scene Detection in Hollywood Movies
VISUALIZATION TOOL

Ranked List	Annotations
<p>RUN:</p> <ul style="list-style-type: none"> ◆ MFCC-based Bag-of-Audio-Words (with 100k samples) ○ MFCC-based Bag-of-Audio-Words (with 200k samples) ○ MFCC-based Bag-of-Audio-Words (with 300k samples) ○ MFCC-based Bag-of-Audio-Words (with 400k samples) ○ MFCC-based Low-Level Representation ○ Bag-of-Audio-Words based on Sparse Codes of MFCC (with 100k samples) ○ Bag-of-Audio-Words based on Sparse Codes of MFCC (with 200k samples) ○ Bag-of-Audio-Words based on Sparse Codes of MFCC (with 300k samples) ○ Bag-of-Audio-Words based on Sparse Codes of MFCC (with 400k samples) ○ Ground Truth 	
<p>MOVIE:</p> <ul style="list-style-type: none"> ◆ All Test Movies ○ Dead Poets Society (1989) ○ Fight Club (1999) ○ Independence Day (1996) 	

SUBMIT

Fig. 11.6: *Ranked List* view of our visualization tool.

fire, firearms, cold weapons and gore. The user can query any or all movies for any of these concepts (e.g., show all segments with fire in *Saving Private Ryan*). The annotations are then displayed in a view (Figure 11.7) similar to the one of the *Ranked List*.

11.5.4 Online Analysis

The *Online Analysis* (Figure 11.8) executes our VSD pipeline to any video hosted by YouTube (or any other site supported by the youtube-dl script). After the user entered the URL, the video is downloaded, transcoded and split into segments. The MFCC feature vectors of the audio of each segment are subsequently computed and used to build mid-level features with sparse coding and vector quantization as explained in [1]. Both mid-level feature representations are used to classify the segment and produce two violence scores. Even though our methods only use audio features, the *Online Analysis* pipeline can be applied to any method using audio, visual or audio-visual features. In addition to the *Ranked List* view, the *Online Analysis* produces a summary box with the maximum violence score and the number of violent segments. The results are cached so that a query of a previously seen video can return the results immediately without downloading or classifying again.

Fig. 11.7: *Annotations* view of our visualization tool.

11.5.5 Discussion

On one hand, our audio-based method is able to suitably detect violent content such as fights and disasters with explosions. Video segments which contain no excitement (e.g., containing a man giving a speech or featuring strong music in the background) are also easily classified as non-violent. On the other hand, the method wrongly classifies a video segment as violent when the segment contains very strong sounds or exciting moments such as a plane taking off or a bell ringing loudly. The most challenging violent segments to detect are the ones which are “violent” according to the objective definition of violence given in the MediaEval VSD task, but which actually contain only actions such as self-injuries, or other moderate actions such as an actor pushing or hitting slightly another actor. Our method is also unable to detect violent video segments which are “violent” according to the objective definition of violence, but which contain no audio cues exploitable for the identification of violence (e.g., a man bleeding). More detailed discussion on the performance of our method is given in [1].

Fig. 11.8: Screenshot of *Online Analysis* view of our visualization tool (showing classification results).

11.6 Conclusions and Future Work

In this chapter, we presented an approach for the detection of violent content in movies and short web videos at the video segment level. We employed low- and mid-level audio-visual features to represent videos. The mid-level audio and visual representations are based on BoW where we first extract audio features (MFCC) and visual features (dense HoG and HoF), and subsequently apply sparse coding on each feature descriptor separately. We used ViF and affect-related color features as low-level visual representation of videos.

Since “violence” is a very diverse concept, we first performed feature space partitioning through clustering video segments instead of learning a unique violence detection model. We then learned a different model for each violence sub-concept. In order to combine the classification results of the violence models in the test phase, we performed a classifier selection. More specifically, we labeled a video segment with the output of the classifier whose cluster center is closest to the video segment in terms of Euclidean distance.

To demonstrate the wide applicability of our solution, we evaluated our method on two different datasets: one dataset of Hollywood movies and one dataset of user-generated videos.

We showed that the mid-level audio representation based on MFCC and sparse coding provides very promising performance in terms of AP and AP@100 metrics

and also outperforms visual representations that we used in this work. We also fused these mid-level audio cues with low- and mid-level visual cues at the decision level using linear fusion for further improvement and achieved better results than unimodal video representations in terms of the AP metrics.

Different from Hollywood movies, user-generated videos are more challenging, since they are not professionally edited, e.g., in order to enhance dramatic scenes. We also demonstrated the performance of our system on the challenging web video dataset which contains short web videos from YouTube. The evaluation results on the short web videos were similar to the ones on the Hollywood movie dataset and hence, showed that our violence detection method generalizes well to different types of video content.

We observed from the overall evaluation results that our method performs better when violent content is better expressed in terms of audio features. Hence, as a future work, we need to extend/improve our visual representation set with more discriminative representations. Another possibility for future work is to further investigate the feature space partitioning concept and optimize the distribution or number of sub-concepts in order to enhance the classification performance of our method.

Acknowledgements The research leading to these results has received funding from the European Community FP7 under grant agreement number 261743 (NoE VideoSense). 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. Our thanks also go to *Fudan University* and *Vietnam University of Science* for providing the ground truth of the Web video dataset.

References

1. Esra Acar, Frank Hopfgartner, and Sahin Albayrak. Detecting violent content in hollywood movies by mid-level audio representations. In *CBMI 2013*. IEEE, 2013.
2. R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. *Machine Learning: ECML 2004*, pages 39–50, 2004.
3. S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15:561–568, 2002.
4. B.J. Bushman and L.R. Huesmann. Short-term and long-term effects of violent media on aggression in children and adults. *Archives of Pediatrics & Adolescent Medicine*, 160(4):348, 2006.
5. Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su. Horror video scene recognition via multiple-instance learning. In *ICASSP*, 2011.
6. Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su. Violence detection in movies. In *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*, pages 119–124. IEEE, 2011.
7. Fillipe Dias Moreira de Souza, Guillermo Camara Chávez, EA do Valle, and A de A Araujo. Violence detection in video using spatio-temporal features. In *Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on*, pages 224–230. IEEE, 2010.
8. Claire-Hélène Demarty, Bogdan Ionescu, Yu-Gang Jiang, Vu Lam Quang, Markus Schedl, and Cédric Penet. Benchmarking violent scenes detection in movies. In *Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on*, pages 1–6. IEEE, 2014.

9. Claire-Hélène Demarty, Cédric Penet, Guillaume Gravier, and Mohammad Soleymani. The MediaEval 2012 Affect Task: Violent Scenes Detection. In *Working Notes Proceedings of the MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.
10. Nadia Derbas and Georges Quénot. Joint audio-visual words for violent scenes detection in movies. In *Proceedings of International Conference on Multimedia Retrieval*, page 483. ACM, 2014.
11. Xinmiao Ding, Bing Li, Weiming Hu, Weihua Xiong, and Zhenchong Wang. Horror video scene recognition based on multi-view multi-instance learning. In *Computer Vision-ACCV 2012*, pages 599–610. Springer, 2013.
12. Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
13. T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis. Violence content classification using audio features. *Advances in Artificial Intelligence*, pages 502–507, 2006.
14. 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*, pages 91–100, 2010.
15. 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*, pages 317–326, 2008.
16. Shinichi Goto and Terumasa Aoki. Violent scenes detection using mid-level violence clustering. *Computer Science*, 2014.
17. David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Electronic Imaging 2003*, pages 87–95. International Society for Optics and Photonics, 2003.
18. Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 1–6. IEEE, 2012.
19. H. He and E.A. Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.
20. Berthold K Horn and Brian G Schunck. Determining optical flow. In *1981 Technical Symposium East*, pages 319–331. International Society for Optics and Photonics, 1981.
21. Bogdan Ionescu, Jan Schlüter, Ionut Mironica, and Markus Schedl. A naive mid-level concept-based fusion approach to violence detection in hollywood movies. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 215–222. ACM, 2013.
22. Orit Kliper-Gross, Tal Hassner, and Lior Wolf. The action similarity labeling challenge. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):615–621, 2012.
23. Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
24. 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*, pages 930–935, 2009.
25. Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia*, pages 83–92. ACM, 2010.
26. Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.
27. Dominique Maniry, Esra Acar, Frank Hopfgartner, and Sahin Albayrak. A visualization tool for violent scenes detection. In *ICMR'14: Proceedings of ACM Conference on Multimedia Retrieval*, pages 522–523. ACM, 04 2014.
28. Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns*, pages 332–339. Springer, 2011.

29. Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, and Patrick Gros. Multimodal information fusion and temporal integration for violence detection in movies. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 2393–2396. IEEE, 2012.
30. Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, Patrick Gros, et al. Technicolor/inria team at the mediaeval 2013 violent scenes detection task. *MediaEval 2013 Working Notes*, 2013.
31. Mats Sjöberg, Bogdan Ionescu, Yu-Gang Jiang, Vu Lam Quang, Markus Schedl, and Claire-Hélène Demarty. The MediaEval 2014 Affect Task: Violent Scenes Detection. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, Barcelona, Spain, October 16-17 2014.
32. Jasper RR Uijlings, IC Duta, Negar Rostamzadeh, and Nicu Sebe. Realtime video classification using dense hof/hog. In *Proceedings of International Conference on Multimedia Retrieval*, page 145. ACM, 2014.
33. Hee Lin Wang and Loong-Fah Cheong. Affective understanding in film. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(6):689–704, 2006.
34. Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, 2004.
35. Long Xu, Chen Gong, Jie Yang, Qiang Wu, and Lixiu Yao. Violent video detection based on mosift feature and sparse coding. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3538–3542. IEEE, 2014.
36. M. Xu, N.C. Maddage, C. Xu, M. Kankanhalli, and Q. Tian. Creating audio keywords for event detection in soccer video. In *ICME'03*. IEEE, 2003.
37. 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*, volume 1, pages 657–663. IEEE, 2005.
38. Lahav Yeffet and Lior Wolf. Local trinary patterns for human action recognition. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 492–497. IEEE, 2009.