Neurocomputing 208 (2016) 225–237

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Breaking down violence detection: Combining divide-et-impera and coarse-to-fine strategies

Esra Acar [a,*], Frank Hopfgartner [b], Sahin Albayrak [a]

[a] Technische Universität Berlin, Distributed Artificial Intelligence Laboratory, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
[b] University of Glasgow, Humanities Advanced Technology and Information Institute, University Gardens, Glasgow, UK

ARTICLE INFO

Article history:
Received 26 September 2015
Received in revised form
20 April 2016
Accepted 1 May 2016
Available online 3 June 2016

Keywords:
Event detection
Violence concept
Ensemble learning
Feature space partitioning
Coarse-to-fine violence analysis
Support vector machine

ABSTRACT

In today's society where audio–visual content is ubiquitous, violence detection in movies and Web videos has become a decisive functionality, e.g., for providing automated youth protection services. In this paper, we concentrate on two important aspects of video content analysis: Time efficiency and modeling of concepts (in this case, violence modeling). Traditional approaches to violent scene detection build on audio or visual features to model violence as a single concept in the feature space. Such modeling does not always provide a faithful representation of violence in terms of audio–visual features, as violence is not necessarily located compactly in the feature space. Consequently, in this paper, we target to close this gap. To this end, we present a solution which uses audio–visual features (MFCC-based audio and advanced motion features) and propose to model violence by means of multiple (sub)concepts. To cope with the heavy computations induced by the use of motion features, we perform a coarse-to-fine analysis, starting with a coarse-level analysis with time efficient audio features and pursuing with a fine-level analysis with advanced features when necessary. The results demonstrate the potential of the proposed approach on the standardized datasets of the latest editions of the *MediaEval Affect in Multimedia: Violent Scenes Detection* (*VSD*) *task* of 2014 and 2015.

1. Introduction

The amount of multimedia content accessible to consumers becomes more and more abundant. This creates a need for automatic multimedia analysis solutions which can be used to find relevant semantic search results or to identify illegal content present on the Internet. In parallel, the developments in digital media management techniques have simplified delivering digital videos to consumers. As a consequence, gaining access to online film productions offered on platforms such as Video-On-Demand (VOD) services has literally become a *child's play*, with the risk that children be exposed to movies or reality shows which have not been checked by parents. Thus, these might contain inappropriate content, as one cannot expect that parents constantly and precisely monitor what their children are viewing. Violence constitutes one typical example of such inappropriate content, whose negative effects have been evidenced [1]. Consequently, a need for automatically detecting violent scenes in videos (e.g., movies, Web videos) has appeared.

Nowadays, movies receive different ratings in different countries (e.g., age of 0, 12, etc.). Even if there is an agreement between different national rating institutes, the perception of violence can still differ from person to person, even within a group of persons of identical age. Due to the subjective nature of the "violence" concept, one of the challenges is to adequately delimit the boundaries of what can be designated as a "violent" scene. Therefore, one preliminary step is the adoption of a definition of violence to work with. We adhere to the definition of violence as described in [2]: *subjective violence*. According to [2], *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 Affect in Multimedia: Violent Scenes Detection* (VSD) task [3], held yearly since 2011, has provided a consistent evaluation framework to the research community and enabled various approaches to be evaluated and compared by using the same violence definition and a standardized dataset. Interested readers will find a comprehensive description of the task, dataset, ground truth and evaluation criteria in [3].

* Corresponding author.
E-mail address: esra.acar@tu-berlin.de (E. Acar).

http://dx.doi.org/10.1016/j.neucom.2016.05.050
0925-2312/

The task stems from a use case attributed to 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. This helps them decide if, according to their own criteria, the movie is adequate to be watched by their child.

For the reasons we stated above, an effective violence detection solution, which is designed to automatically detect violent scenes in movies (or in videos in general), is highly desirable. Such an automated solution requires working with a proper representation of data which is an essential processing step. Recently, solutions using mid-level feature representations have gained popularity. These solutions shifted away not only from the traditional approaches which represented videos using low-level features (e.g., [4,5]) but also from the use of state-of-the-art detectors designed to identify high-level semantic concepts (e.g., "a killing spree"). The earlier solutions could not carry enough semantic information, and the latter ones have not reached a sufficient level of maturity. Hinted by these recent developments, we adopt here mid-level audio and motion representations as they may help modeling video segments one step closer to human perception. As a basis for the mid-level audio and motion representations, we employ MFCC and dense trajectory features, respectively. Using simultaneously audio and visual information is computationally expensive. We approach this issue by exploiting audio and visual information in a coarse-to-fine setup to reduce computations and boost the velocity of violence detection. In addition, this can be used for designing scalable solutions, i.e., adjustable depending on the processing power or accuracy requirements.

In parallel to the progress in feature representation, machine learning techniques are constantly improved in order to effectively use features. A development in this direction is feature space partitioning [6]. A classifier is usually trained on a given dataset to detect a unique class (e.g., the concept of violence). However, such a class might not be expressed in a "compact" manner in the feature space. Partitioning the feature space to build multiple models that correspond to the same concept might help in properly recognizing a given concept. Therefore, instead of building a unique model to detect violence, we use feature space partitioning. This presents several advantages. It enables a faithful modeling of "violence". It also constitutes a data-driven operation, as it does not require defining manually several "violence" concepts (e.g., there is no need to have a separate concept for "explosion", "fire" or other similar concepts), as it directly builds on the data. Finally, this aspect is not hardwired to "violence" only, but can be extended to other concepts.

The paper is organized as follows. Section 2 explores the recent developments by reviewing video violent content detection methods which have been proposed in the literature, and presents the contributions of the paper. In Section 3, we introduce our method and the functioning of its various components. We provide and discuss evaluation results obtained on the latest MediaEval datasets of 2014 and 2015 in Section 4. Concluding remarks and future directions to expand our current approach are presented in Section 5.

2. Related work and contributions

2.1. Related work

Although video content analysis has been extensively covered in the literature, violence analysis of movies or of user-generated videos does not enjoy a comparable coverage and is restricted to a few studies. We present here a selection of the most representative ones, from a machine learning and classification perspective. As a preliminary remark, we would like to emphasize that, with respect to prior art studies, the definition of violence poses a difficulty. 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.

One popular type of approach adopted in the literature is classification based on SVM models. An illustration to SVM-based solutions is the work by Giannakopoulos et al. [7], where violent scenes are defined as those containing shots, explosions, fights and screams, while non-violent content corresponds to audio segments containing music and speech. Frame-level audio features both from the time and the frequency domain are employed and a polynomial SVM is used as the classifier. In [8], de Souza et al. 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 SVM approach is based on the use of Bag-of-Words (BoW), where local Spatial-Temporal Interest Point Features (STIP) are used as feature representations. 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). Hassner et al. [9] 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 using the Violent Flows (ViF) descriptor. ViF descriptors are then classified as either violent or non-violent using a linear SVM. In [10], Gong et al. propose a three-stage method. In the first stage, they apply a semi-supervised cross-feature learning algorithm [11] on the extracted audio–visual 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. Chen et al. [12] 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, faces are detected in each keyframe of action scenes and the presence of blood pixels near detected faces is checked using color information. Aiming at improving SVM-based classification, Wang et al. [4] apply Multiple Instance Learning (MIL; MI-SVM [13]) using audio–visual features in order to detect horror. 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. In [14], 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 multiple kernel learning.

Next to SVM-based solutions, approaches which make use of other types of learning-based classifiers exist. Yan et al. [15] adopt a Multi-task Dictionary Learning approach to complex event detection in videos. Based on the observation that complex events

[1] https://research.technicolor.com/rennes/