Expert Systems with Applications 38 (2011) 14102–14116

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Multimodal and ontology-based fusion approaches of audio and visual processing for violence detection in movies

Thanassis Perperis [a,*], Theodoros Giannakopoulos [a], Alexandros Makris [b], Dimitrios I. Kosmopoulos [b], Sofia Tsekeridou [c], Stavros J. Perantonis [b], Sergios Theodoridis [a]

[a] Dept. of Informatics and Telecommunications, University of Athens, GR 15784, Greece
[b] NCSR Demokritos, Inst. of Informatics and Telecommunications, GR 15310, Greece
[c] Athens Information Technology (AIT), 0.8 km Markopoulou Ave., GR 19002 Peania, Athens, Greece

ARTICLE INFO

Keywords:
Violence
Movie
Multimodal fusion
Learning
Ontology
Knowledge representation
Reasoning

ABSTRACT

In this paper we present our research results towards the detection of violent scenes in movies, employing advanced fusion methodologies, based on learning, knowledge representation and reasoning. Towards this goal, a multi-step approach is followed: initially, automated audio and visual analysis is performed to extract audio and visual cues. Then, two different fusion approaches are deployed: (i) a multimodal one that provides binary decisions on the existence of violence or not, employing machine learning techniques, (ii) an ontological and reasoning one, that combines the audio-visual cues with violence and multimedia ontologies. The latter reasons out not only the existence of violence or not in a video scene, but also the type of violence (fight, screams, gunshots). Both approaches are experimentally tested, validated and compared for the binary decision problem of violence detection. Finally, results for the violence type identification are presented for the ontological fusion approach. For evaluation purposes, a large dataset of real movie data has been populated.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

In the age of broadband and next generation Media Internet, digital television (IPTV, Video on Demand) and streaming video services (e.g. YouTube.com), the dissemination of objectionable audio-visual content, such as violence and pornography, become uncontrollable. Although psychological research on media violence (Ángel Vidal, Clemente, & Espinosa, 2003; Kevin D Browne, 2005, 2002), has proven its negative effects on behavior, attitude and emotional state of children, highlighting thus the need for automatic tools, to filter out this content, research towards that direction is limited. Some previous approaches tackle the problem at hand by detecting a limited and simple set of violence actions and semantics (i.e. kicking, fist fighting, explosions, gunshots) mostly in controlled environments (surveillance applications, stationary cameras, absence of editing effects etc.).

In this work, we exploit medium level semantics, towards devising a late fusion scheme based on a kNN Binary classifier and further investigate the usage of an ontological and knowledge engineering approach towards representing and inferring complex violence semantics. Our overall goal is to compare and explore the potentials of single and multimodal analysis approaches towards automatic semantic violence identification and annotation of video scenes, aiming further at content filtering and enabling parental control.

Content filtering seems the most important application of our method. However direct application of the proposed methodology either on movies or in other domains (i.e. sports, news) could enhance overall viewers experience either in the sense of personalization services or by automated description generation. Automatic annotation of existing unstructured multimedia databases, for semantic search and retrieval services, is another potential application.

The main innovative features of the proposed system, which will be further described in the next sections, are:

- Extraction of some discriminative audio and visual features, specific for violence detection.
- Fusion of the audio and visual modalities, using a meta-classification scheme.
- Cross modality ontological/inferencing framework for violence identification.
- First attempt for complete ontological definition of the movie violence domain.

In the next section we survey the recent developments in the field. In Section 3 we shortly describe the proposed methodology

* Corresponding author. Tel.: +307275320.
E-mail addresses: a.perperis@di.uoa.gr (T. Perperis), tyiannak@di.uoa.gr (T. Giannakopoulos), amakris@iit.demokritos.gr (A. Makris), dkosmo@iit.demokritos.gr (D.I. Kosmopoulos), sots@ait.edu.gr (S. Tsekeridou), sper@iit.demokritos.gr (S.J. Perantonis), stheodor@di.uoa.gr (S. Theodoridis).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.219

towards violence detection. In Sections 4 and 5 we describe the audio and video analysis procedures. In Section 6, an audiovisual fusion technique is presented for the task of detecting violence in movies. In Section 7, the ontological analysis for the same problem is described. Finally, in Section 8 the experimental results are provided and we conclude this work with Section 9.

2. Related work

2.1. Audio analysis driven violence detection

In movies, most violent scenes are characterized by specific audio events (e.g. explosions, gunshots). The literature related to the detection of violent content is limited. In Nam and Tewfik (2002) the energy entropy is used as additional information to visual data. Energy entropy can be used in order to detect abrupt changes in the audio signal, which, in general, may characterize violence content. Though, the usage of this feature for violent detection can only be used in combination with other cues, since it only detects abrupt changes and it could therefore lead to the classification of a non violent impulsive noise (e.g. a thunder) as violent. In Rasheed and Shah (2002), a film classification method is proposed that is mainly based in visual cues. The only audio feature adopted in this paper is the signal's energy.

A more detailed examination of the audio features for discriminating between violent and non-violent sounds was presented in our initial approach in Giannakopoulos, Kosmopoulos, Aristidou, and Theodoridis (2006). In particular, seven audio features, both from the time and frequency domain, have been used, while the binary classification task (violent vs non violent) was accomplished via the usage of Support Vector Machines. In our later work in Giannakopoulos, Pikrakis, and Theodoridis (2007), a multi-class classification algorithm for audio segments from movies has been proposed. Bayesian networks along with the one vs all architecture has been used, while the definition of the classes has been violence-oriented (three violent classes have been adopted).

In this work we have used a variant of the classifier proposed in Giannakopoulos et al. (2007), on a segment basis, in order to generate a sequence of class labels. We extend this work, by detecting violent content using a meta-classification process, instead of simply generating a multi-class classification decision. Further to that, in this meta-classification process the fusion of visual-based features is also introduced, as will be presented in Section 6.

2.2. Video analysis driven violence detection

The reported research that uses visual features for violence detection is limited as well. Most of these works concern surveillance cameras and use background subtraction techniques to detect the people in the scene (Vasconcelos & Lippman, 1997; Zajdel, Krijnders, Andringa, & Gavrila, 2007). These approaches however are not suitable for movies where the camera moves abruptly and there are many shot changes. In Vasconcelos and Lippman (1997), a generic approach to determine the presence of violence is presented. Two features are used, which measure the average activity and the average shot-length. Experiments with movie trailers show that the features are able to discriminate violent from non-violent movies. However, no attempt to characterize the specific segments of the movie which contain the violence is carried out. In Nam, Alghoniemy, and Tewfik (1998), three visual features are used, indicating shot changes, the presence of gunfires/explosions and the presence of blood.

We extend the previous works, by defining novel features that represent the level of activity and visually observable gunshots. We also exploit the human presence through a person detector.

2.3. Multimodal analysis driven violence detection

Although multimodal analysis is extensively examined in constrained domains like sports or news (i.e. Babaguchi, Kawai, & Kitahashi, 2002; Cheng & Hsu, 2006; Hsu et al., 2004; Lin Huang, Chia Shih, & Yuan Chao, 2006; Leonardi, Migliorati, & Prandini, 2004) little progress is achieved on complex movie domains. In Iyengar, Nock, Neti, and Franz (2002), a multimodal late fusion approach is presented. A set of scores representing the existence of every concept (fire, sky, outdoors, face, etc.) in each video shot is produced from Audio, Visual and Text models. These scores are combined to form a simple vector that is classified using a Support Vector Machine approach. Similarly a late fusion approach is presented in Li and Tan (2005) for event detection. Features from the visual and auditory modalities are initially extracted and analyzed to generate mid level concepts (i.e. faces, screams, speech). The logistic regression and Bayesian belief network are then employed to fuse the information and detect the news and sports events of interest. In Lehane, O'Connor, and Murphy (2004), high level film-making knowledge is combined with low and mid level digital video analysis into a state machine in a late fusion scheme detecting action sequences in movies. In this paper, we extend the previous work, by proposing a late fusion approach, which combines the probabilistic outputs of the single-modality classifiers.

2.4. Knowledge-based semantics extraction for violence detection

The knowledge-based approaches, employing multimedia and domain ontologies for multimodal video analysis, presented promising results in the past in the medical, sports and surveillance domains. The only ontological approach towards movie violence identification, was our preliminary work presented in Perperis, Tsekeridou, and Theodoridis (2007). Bao, Cao, Tavanapong, Honavar, and Honavar (2004) Fan, Luo, Gao, and Jain (2007) defined domain ontologies for medical applications: in the former case to boost hierarchical video classifier training through exploiting the strong correlations between the video concepts and multitask learning and in the latter case to integrate with independent (i.e. multimedia) ontologies towards colonoscopy video annotations. Bai, Lao, Jones, and Smeaton (2007) proposed a video semantic content analysis framework based on ontologies as well. A Soccer Domain ontology was used to define high level semantic concepts and their relations. Low-level features, video content analysis algorithms and Description Logic rules were integrated into the ontology to enrich video semantic analysis results. Bertini, Del Bimbo, and Torniai (2005) developed a pictorially enriched ontology and Reidsma, Kuper, Declerck, Saggion, and Cunningham (2003) introduced an ontology, created from newspapers descriptions and video speech transcriptions, for soccer video annotations. Few attempts appear in the literature for semantic video analysis in more abstract domains. In Snidaro, Belluz, and Foresti (2007) domain ontologies and SWRL rules were used towards reasoning about entities and their interactions in a surveillance application. In Neumann and Möller (2008) high-level scene interpretation was attempted, using Description Logics as a knowledge representation and reasoning system. An ontological *ad hoc* approach for video event detection applied mostly in surveillance, physical security and meeting understanding applications emerged as a result of the ARDA event taxonomy challenge project (Bolles & Nevatia, 2004a, 2004b). Towards creating this taxonomy a novel language for video event representation was proposed (Francois, Nevatia, Hobbs, & Bolles, 2005).

In this work we make a step forward towards employing ontologies and rules in such an abstract and complex domain like movie violence identification. A complete ontological framework is proposed combining domain and modality specific ontologies with

Download English Version:

https://daneshyari.com/en/article/385415

Download Persian Version:

https://daneshyari.com/article/385415

Daneshyari.com