A Knowledge Engineering Approach for Complex Violence Identification in Movies

Thanassis Perperis[1] and Sofia Tsekeridou[2]

[1] University of Athens, Greece a.perperis@di.uoa.gr
[2] Athens Information Technology, Greece sots@ait.edu.gr

Abstract. Along with the rapid increase of available multimedia data, comes the proliferation of objectionable content such as violence and pornography. We need efficient tools for automatically identifying, classifying and filtering out harmful or undesirable video content for the protection of sensitive user groups (e.g. children). In this paper we present a multimodal approach towards the identification and semantic analysis of violent content in video data. We propose a layered architecture and focus on ontological and knowledge engineering aspects of video analysis. We demonstrate the development of two ontologies defining violent hints hierarchy that low level analysis, in visual and audio modality, respectively should identify. Violence domain ontology, as a reality representation, defines higher-level semantics. Taking under consideration extracted violent hints, spatio-temporal relations and behavior patterns higher-level semantics automatic inference is possible.

1 Introduction

Psychological researches on media violence prove its negative effects on behavior, attitude and emotional state of vulnerable user groups (especially children). In the age of internet technologies and digital television, dissemination of dangerous content seems uncontrollable. Common users and industry demand intelligent, human-like methods, to automatically detect, annotate and filter out any violence hidden in video data, thus enabling high level parental control. The main obstacle towards this direction is the inability of machines to grasp high level semantic concepts from multimedia data (multimedia semantic gap). In order to tackle this problem we need efficient audio and visual medium level concept/event detectors, higher level domain knowledge represented in a formal way and tools to optimally handle their interoperation.

Previous research towards bridging multimedia semantic gap follow either a *unimodal* or a *multimodal* approach. The former case consists in classifying low level feature vectors, extracted from a *single* modality, in a set of predefined classes and the later in *fusing* each modalitys' low level analysis results. Multimodal fusion schemes either combine single modality features in a multimodal representation and feed those to machine learning algorithms to extract combined semantics (early fusion) or couple each modalitys' medium level semantics (extracted using single modality analysis techniques) to achieve higher level of abstraction and improve semantics extraction accuracy (late fusion).

Ontologies and MPEG-7 tackle the problem of knowledge, content and semantics representation and annotation, which arise in either multimedia semantics analysis techniques. MPEG-7 [2] defines metadata descriptors for structural and low level aspects of multimedia documents, as well as, high level description schemes (Multimedia Description Schemes) encapsulating multimedia content semantics. Deploying Semantic Web trends and ontologies on multimedia data, which are complex in nature, multi-modal, of significant size, requiring extensive and efficient analysis to reduce the data space and extract representative features and descriptions, is a very challenging task. On the on hand multimedia ontologies, tackling the data characteristics, implement definition models for low- to medium-level descriptions, as the ones dictated by the Audio and Visual Parts of MPEG-7. On the other hand domain ontologies represent domain knowledge in the form of high-level concepts, hierarchies and relations. The interoperation of Multimedia and Domain Ontologies, along with the optimal definition of the latter, to support automated semantic annotation of multimedia data is a major research focus [5] in various application domains.

In this paper we propose a late fusion scheme for automatic identification and annotation of complex violent scenes in video data. We overlook medium level semantic extraction techniques (we use them as black boxes) and focus our attention on representing and inferring single and cross modality semantics using an ontological and knowledge engineering approach. Previous, mainly low level analysis based 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), in order to ease the solution. Our approach by focusing on tackling an extensive range of complex violent acts in video data, based on violence domain knowledge representation, using ontologies and reasoning on or inferring from results obtained by multimodal analysis of both visual [10] and audio [8] modalities, attempts to proceed further.

The utmost goal is to automatically detect complex multimodal semantics around violence hidden in video data, annotate them accordingly and enable filtering of content for parental control. A crucial step in the overall methodology is the best possible definition of the underlying ontologies. To optimally combine multimedia descriptions with the violence domain ontology, the knowledge representation process has involved the definition of modality violence ontologies (audio, visual) that essentially map low-level analysis results to simple violence events and objects (medium-level semantics), as well as a violence domain ontology that defines the hierarchy of violence concepts and inherent relationships, irrespective of data, starting from abstract and complex ones to more simple and concrete ones (shared by all ontologies). The latter is used as input to the inference engine that undertakes the fusion of results from medium level semantics to lead to higher level ones and to infer knowledge about existing violence.

The paper is organized as follows. Section 2 briefly describes the overall architecture of the proposed solution and analyzes developed ontologies in detail. Section 3 discusses the creation of a violent movies corpus and the related ground truth. Finally, discussion for further work and conclusions are drawn in Section 4.

2 Knowledge Engineering Methodology for Violence Identification in Movies

Violence is a very abstract concept describing actions and situations, which may cause physical or mental harm to one or more persons, injury to animals, or damage to non-living objects. Violent content in video data refer to scenes that include such actions. Previous research towards violence identification in video data is limited and in most cases examines only low level features to extract simple semantics. In [7] the design of a simple feature space, for scene categorization based on the degree of action (i.e. degree of violence), is presented. In [3] detection of person on person violence, such as fist fighting, kicking, hitting with objects, in video data captured using a stationary camera, is accomplished using motion trajectory and orientation information of a person and its limbs. Audio data for violence detection is used as an additional feature in TV drama and movies in [6], where abrupt changes (i.e. explosions) in energy level of the audio signal are detected using the energy entropy criterion.

Combining violent objects, actions and events extracted from visual and audio modality respectively, towards composing and identifying higher level violent behaviors, seems a very promising approach for the problem at hand. A crucial step for this approach is to represent the violence domain knowledge as effectively as possible, in all its complexity, abstractness and hierarchy depth. A formal representation of the violence domain, to drive violent acts detection, has never been attempted before. We make a step forward towards this direction. Our overall goal is to devise a multimodal analysis, fusion and inferencing methodology towards automatic semantic violence extraction and annotation of video scenes, aiming further at content filtering and enabling parental control. The conceptual architecture of our overall methodology is shown in Fig. 1.

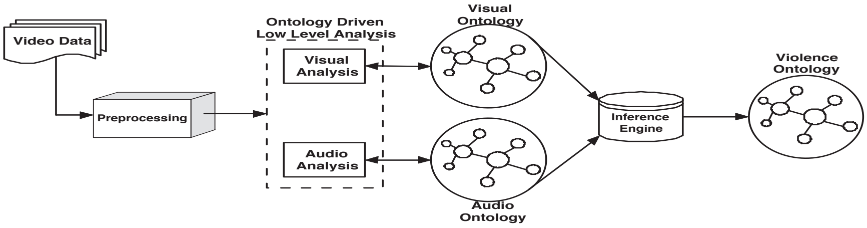

Fig. 1. Conceptual Architecture

A preprocessing step tackles the task of temporal video segmentation and feeds the low level audio and visual analysis algorithms with video shots and audio segments respectively. A visual and an audio ontology defining the hierarchy and relations of violent objects and primitive actions/events drive the corresponding segment based low level analysis procedures. Having recognized some audio and visual violent hints and instantiate the corresponding MPEG-7 descriptors, the inference engine using probabilistic reasoning, spatiotemporal relations and behavioral patterns maps sets and sequences of violent hints to higher level violent concepts represented in the domain ontology. In the following paragraphs we neglect the preprocessing and low level analysis steps (considered as black boxes) and we focus on the higher level of analysis, by presenting the corresponding ontologies and sketching the inference procedure.

2.1 Violence Domain Knowledge

As we pinpointed previously, effective violence domain representation, in all its complexity, abstractness and hierarchy depth is crucial towards tackling the problem at hand. Combining psychologists' view of violence and violent acts and extended investigation through observation of video data, we make the first attempt to conceptualize violence in an organized way. We define complex semantics of extensive violent acts, also found in movie data, along with cross-modal relations of medium level semantics, deploying Semantic Web languages, in violence domain ontology. The violence domain ontology as a knowledge representation can be further exploited by researchers and organizations investigating violence (i.e. psychologists, pedagogists, police).

Although our ontology comprises a generic representation of violence we will focus our analysis in the movie violence domain. In a movie scene containing violence (e.g. torture, fight, war) a spectator can quickly grasp the form of violence (e.g. fighting with weapons), recognize a sequence of violent (e.g. shooting, stabbing), of generic (e.g. running, walking) and of consequence (e.g. falling, crawling, scream) actions. The direct application of this process demonstrates how the hierarchy (taxonomy) of violent actions, along with their inter-relations (e.g. a stabbing is followed by a scream), is constructed, formulating the violence domain ontology. The presented movie violence ontology is implemented in OWL-DL [4] using Protégé. In Fig. 2 we present the higher level concepts of the violence ontology (left part) and an instantiation of the ontology (right part) for the violent action "stab" and the violent action "punch", demonstrating hierarchical and temporal concept's relations. The medium level classes (actions) of the ontology are strongly related with the inference engine, since they represent multimodal actions inferred using reasoning by relating to the visual and audio ontology and the single modality analysis and classification results. Additionally the simplest, more concrete, concepts (e.g. weapon, scream) are further represented, along with their low level feature descriptions, in the visual and audio ontologies, thus defining the association mechanisms of the three ontologies.

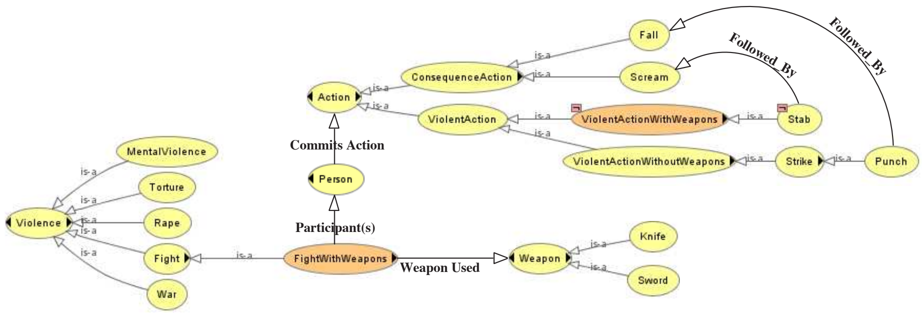

Fig. 2. Violence Ontology

2.2 Visual Semantics for Violence

Every violent behavior described in the domain ontology is usually composed of some primitive actions (performed from a person), or events (either visual or auditory) and include a set of objects (e.g. knifes, swords). The visual ontology (Fig. 3) for violence defines a taxonomy of moving objects (e.g. people, weapons, body parts, military vehicles), stationary (contextual) objects (e.g. walls, fences, furniture), abstract objects (e.g. explosions, fire, injuries) and primitive actions (e.g. crawling, running, falling). We note that the detection of some of the aforementioned concepts does not directly imply violence (e.g. bottle), but in the context of violence (e.g. hit on the head with bottle) its identification might be very important. The visual ontology further includes the MPEG-7 visual descriptors and MPEG-7 MDS (Multimedia Description Schemes), which describe visual features such as color, texture, shape and motion and semantic information of video respectively, associated with the above mentioned taxonomy entries. Furthermore it drives low level analysis algorithms towards extracting the specified objects and actions from the video data, along with their low level features. Thus the identified concepts and the corresponding features are instantiated based on the ontology. Following the previously reported examples of "punch" and "stab" recognition, in the marked area of Fig. 3 we demonstrate the description instantiation of the moving body part "arm", the visual object "knife" and the injury related concept "wound" (as a consequence of stab). This example further demonstrates the linking between the violence and the visual ontology through common terms, from a different viewpoint.

2.3 Audio Semantics for Violence

In violent scenes one can recognize a set of audio events indicative of violence like gunshots, screams, explosions, hit sounds. Moreover indication of violence can be drawn from the background music (e.g. action scenes with multiple persons fighting are accompanied with intense music) or the emotional speech of an actor

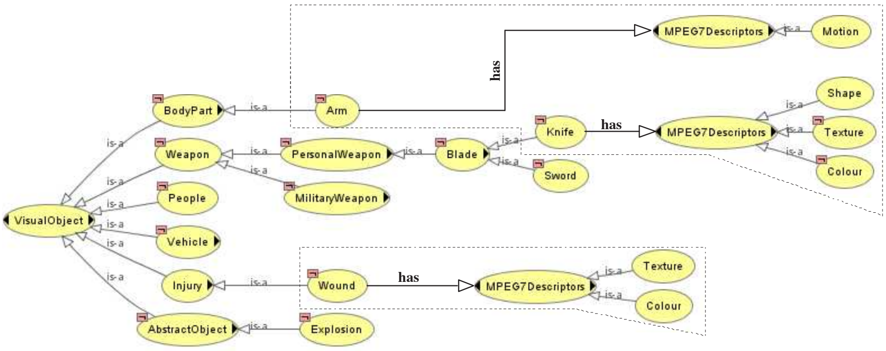

Fig. 3. Visual Ontology

(e.g. angry speech might be followed by some sort of fight). Thus, we have further implemented the aforementioned audio events (Fig. 4) in a taxonomic way, defining the audio ontology for violence. The ontology is extended with MPEG-7 audio descriptors and MPEG-7 MDS (Multimedia Description Schemes), to specify low level features and semantically describe the audio data respectively. As in the case of the visual ontology a set of classification algorithms [8] (e.g. Bayesian networks, SVMs) is responsible for instantiating audio descriptions in compliance with the ontology including the corresponding sound segments, their categorization and the values of their low level features (e.g. spectrum, timbre, energy, volume). In the marked area of Fig. 4 we demonstrate the instantiation potentials of ”punch” and ”scream” following the aforementioned example. We note that the high level actions (punch, scream) are also represented both in the violence domain and the audio ontology. Thus, as in the case of the visual ontology, the association between the two ontologies is evident.

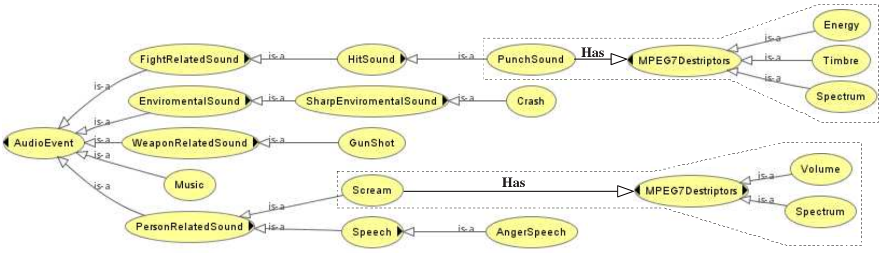

Fig. 4. Audio Ontology

2.4 Inference Engine Design

Having the adequate ontological descriptions for single modality medium level semantics and the corresponding low level extraction algorithms, we need efficient methods that taking under consideration spatio-temporal relations behavior patterns and uncertainty of extraction fuse and interchange medium level semantics from different modalities (audio and visual) to infer/reason about higher level violence behaviors (e.g. more complex, abstract and extensive violence cases). For example, the "punch" concept can be automatically extracted based on the initial analysis results and on the sequence or synchronicity of audio or visual detected events such as two person in visual data, the one moving towards the other, a hand is moving fast towards a face, while a punch sound and scream of pain is detected in the audio data.

To support reasoning mechanisms, it is required that apart from the ontological descriptions for each modality, there is a need for a cross-modality ontological description which interconnects all possible relations from each modality and constructs rules that are cross-modality specific and must tackle the following issues:

- Account for intra- and cross-modality spatial, temporal or spatio-temporal relationships.
- Represent the priorities/ordering among modalities for any multimodal concept.
- Take under consideration cross-modality synchronicity relationship (simultaneous semantic instances in different modalities).
- Handle the issue of importance of each modality for identifying a concept or semantic event.
- Capture uncertainty of extracted medium level semantics and support reasoning with partial, imprecise information.

Rule construction, either in some logic form (FOL, F-logic) or in the form of sequential if-then rules, seems an ideal solution for all sort of relationships' representation (spatio-temporal, ordering, synchronicity). Significance weights can represent the importance of each modality for identifying a concept or semantic event. Bayesian networks or fuzzy logic could handle the uncertainty imported in medium level semantics from low level analysis algorithms.

3 Experimental data setup

We have collected a corpus of 10 movies in MPEG-4 format, containing a variety of violent scenes, composed of both auditory and visual clues. We are in the process of producing manual annotations to form the essential ground truth, as MPEG-7 description instances, based on the violence terms and concepts existent in all defined ontologies. This ground truth data will be used by all processes involved, semantic audio analysis and violent events identification, semantic visual analysis and violent events identification, as well as late fusion methodology and inferencing for complex violent events identification, in order to assess their performance and identification accuracy.

4 Conclusions and Future Work

We have proposed an ontological and knowledge representation approach to define the underlying semantics for violence characterization in video data. This is the first step before providing the inferencing mechanisms in order to automatically identify violent scenes and their context in video data. Thus, this work has to further tackle the question: *how to fuse and interchange semantics from different modalities?*. We are in the process of exploring the usage of basic probabilistic inference methods (Bayesian/belief networks, HMMs), probabilistic reasoning (probabilistic logic, PR-OWL) and rule construction. Furthermore we intend to subsequently apply a similar approach by defining the corresponding ontologies to identify and filter out pornographic content.

References

1. Chandrasekaran B, Josephson J R, Benjamins R V: What Are Ontologies, and Why Do We Need Them? IEEE Intelligent Systems, 14, 1 (1999), 20-26.
2. Manjunath B S, Salembier P, Sikora T: Introduction to MPEG-7: Multimedia Content Description Interface. John Wiley and Sons / England (2002).
3. Datta A, Mubarak S, Lobo N: Person-on-Person Violence Detection in Video Data. Proc. of ICPR2002, Quebec City, Canada, Aug. (2002), 433-438.
4. Smith M K, Welty C, McGuinness D L: OWL Web Ontology Language Guide. W3C Recommendation 10 February 2004, www.w3.org/TR/owl-guide/.
5. Hunter J: Enhancing the Semantic Interoperability of Multimedia through a Core Ontology. IEEE Transactions on Circuits and Systems for Video Technology. Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Description, 13, 1 (2003), 49-58.
6. Nam J, Tewfik A H: Event-driven video abstraction and visualisation. Multimedia Tools and Applications, 16(1-2), 55-77, 2002.
7. Vasconcelos N, Lippman A: Towards semantically meaningful feature spaces for the characterization of video content. Proc. of ICIP1997, Washington, DC, USA, Oct 1997, vol.1, 25-28.
8. Giannakopoulos T, Kosmopoulos D, Aristidou A, Theodoridis S: Violence Content Classification Using Audio Features. Proc. of 4th Hellenic Conference on Artificial Intelligence (SETN'06), Heraklion, Crete, Greece, May 18-20, 2006.
9. Pratikakis I, Tsekeridou S: Use Case : Semantic Media Analysis for Intelligent Retrieval. W3C Multimedia Semantics Incubator Group, www.w3.org/2005/Incubator/mmsem/wiki/Semantic_Media_Retrieval_Use_case.
10. Makris A, Kosmopoulos D, Perantonis S, Theodoridis S: Hierarchical feature fusion for visual tracking. Accepted to be published in Proceedings of IEEE International Conference on Image Processing 2007 (ICIP2007).