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.

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

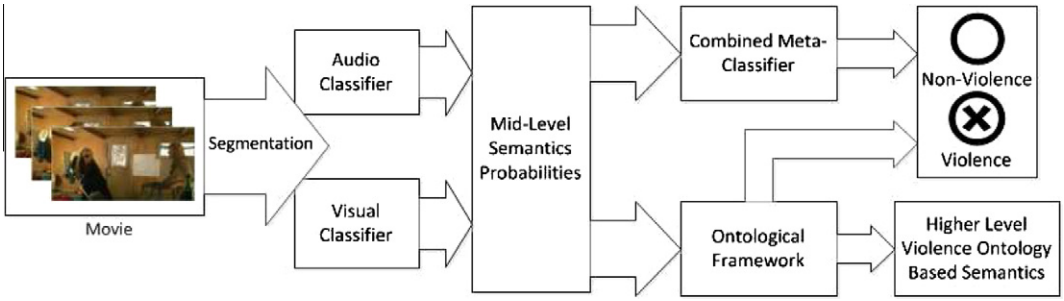

Fig. 1. Proposed methodology for violence detection.

cross modality ontological representation in the form of SWRL rules. Towards tackling the knowledge generation lifecycle problem a 5-Step procedure interchanging ontological and rule reasoning is employed. Moreover, the first violence domain ontological implementation, attempting to capture violence's full depth and abstractness, is presented.

3. General description of the proposed methodology

Our research in violence detection aims towards exploiting mid-level semantics extracted from the audio and the visual modality either in a meta-classifier or in a knowledge representation and reasoning framework (see Fig. 1). A preprocessing step tackles the task of temporal video segmentation and feeds the low level audio and visual analysis algorithms with fixed length audiovisual segments of 1 second, bypassing thus the problem of common timeline representation. The audio classification process exploits a 12-D feature vector in a "One-vs-All" (OVA) classification scheme towards extracting three violence and four non-violence audio classes. The visual classification process involves the definition of motion, person and gunshot detection features and two independent weighted kNN classifiers for activity and gunshot detection. The single modality procedures lead to a 11-D vector, which feeds a kNN meta-classifier and instantiate the corresponding ontologies of the presented ontological framework. The multimodal fusion approach tackles the binary (violence, non-violence) detection problem while the ontological approach further extracts higher level of semantics. In the following sections we further elaborate on each module of the proposed methodology.

4. Audio classification for violence hints

4.1. Audio class definition

In order to create an audio-based characterization scheme, we have defined seven audio classes, three from which are associated to violence and four non-violence. The class definitions have been motivated by the nature of the audio signals met in most movies. In addition, the audio classes are defined, in order to cover all possible audio events met in a film.

The classes that indicate lack of violence are: *Music*, *Speech*, *Others* 1, and *Others* 2. Classes Others 1 and Others 2 are environmental sounds met in movies. Others 1 contains environmental sounds of low energy and almost stable signal level (e.g. silence, background noise, etc.). Others 2 is populated by environmental sounds with abrupt signal changes, e.g. a door closing, thunders, etc. The *violent*-related classes are: *Shots*, *Fights* (beatings) and *Screams*. A detailed description of all audio classes, extracted from the audio features, is given in Table 1.

4.2. Audio feature extraction

For audio-based classification, 12 audio features are extracted for each segment on a short-term basis. In particular, each segment is broken into a sequence of non-overlapping short-term windows (frames). For each frame 12 feature values are calculated. This process leads to 12 feature sequences, for the whole segment. In the sequel, a *statistic* (e.g. standard deviation, or average value) is calculated for each sequence, leading to a 12-D feature vector for each audio segment. The features, the statistics and the window lengths adopted are presented in Table 2. For more detailed descriptions of those features, the reader is referred to Giannakopoulos et al. (2007).

The selection of the particular audio features, and of the respective statistics, was a result of extensive experimentation. However, most of the adopted features have an obvious physical meaning for the task of classifying an audio sample in the particular seven

Table 1
Audio classes definitions.

	Class name	Class description
1	Music	Music from film soundtrack
2	Speech	Speech segments from various speakers, languages and emotional states. Speech is usually mixed with other types of audio classes.
3	Others 1	Environmental sounds of low energy and almost stable signal level (e.g. silence, background noise, wind, rain, etc.)
4	Others 2	Environmental sounds with abrupt changes in signal energy (e.g. a door closing, a sound of a thunder, an object breaking, etc.).
5	Gunshots	Sounds from several types of guns. Contains both short abrupt and continuous gunshots.
6	Fights	Sounds from human fights – beatings.
7	Screams	Sounds of human screams.

Table 2
Window sizes and statistics for each of the adopted features.

	Frame feature	Sequence statistic	Window (msecs)
1	Spectrogram	σ^2	20
2	Chroma 1	μ	100
3	Chroma 2	*median*	20 (mid term:200)
4	Energy entropy	*max*	20
5	MFCC 2	σ^2	20
6	MFCC 1	*max*	20
7	ZCR	μ	20
8	Sp. RollOff	*median*	20
9	Zero pitch ratio	–	20
10	MFCC 1	max/μ	20
11	Spectrogram	*max*	20
12	MFCC 3	*median*	20

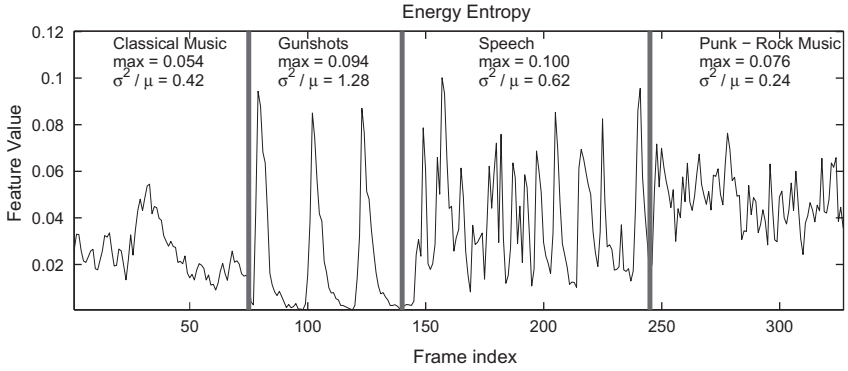

Fig. 2. Example of energy entropy sequence for an audio signal that contains four successive homogenous segments: classical music, gunshots, speech and punk-rock music.

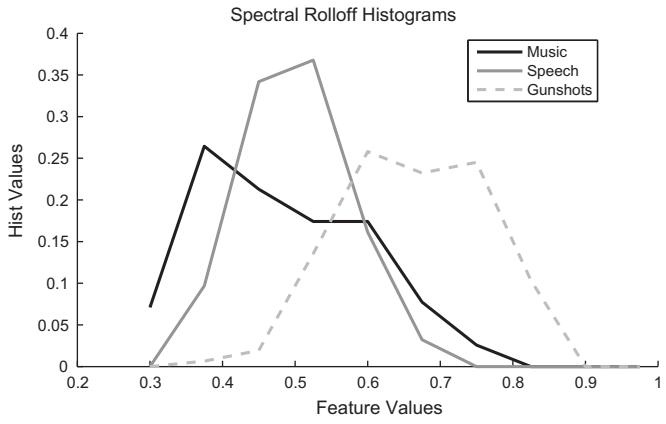

Fig. 3. Median values of the spectral rolloff sequences for three classes of audio segments: music, speech and gunshots.

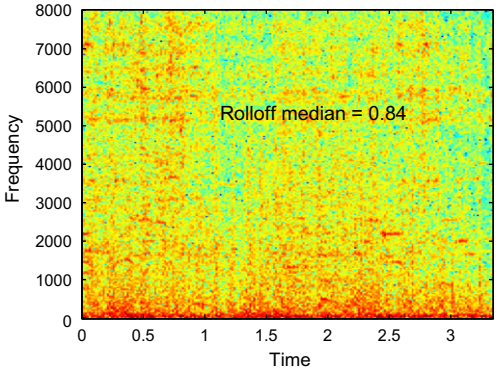

Fig. 4. Gunshots spectrogram.

classes. For example, in Fig. 2 an example of an Energy Entropy sequence is presented for an audio stream that contains: classical music, gunshots, speech and punk-rock music. Also, the maximum value and the $\frac{\sigma^2}{\mu}$ ratio statistics are presented. It can be seen that the maximum value of the energy entropy sequence is higher for gunshots and speech. This is expected, since the energy entropy feature (see Giannakopoulos et al., 2007) has higher values for audio signals with abrupt energy changes (such as gunshots).

Another example of differentiation of the feature values for several audio classes is presented in the sequel. In particular, we present how the *median* values of the spectral rolloff sequences (feature No 8) change for three audio classes. Note that the spectral rolloff is the frequency below which certain percentage of the magnitude distribution of the spectrum is concentrated (Giannakopoulos et al., 2007; Theodoridis & Koutroumbas, 2008). In Fig. 3, the histograms of this feature for music, speech and gunshots audio segments are presented. One important observation is that for the "gunshots" segments the adopted statistic is significantly higher. This is expected, since a gunshot is a sound that is characterized by a widely distributed spectrogram (see spectrograms of a gunshot and a music segment in (Figs. 4 and 5)). Finally, experiments have shown that in 96% of the gunshot segments the median value of the spectral rolloff sequence was higher than 0.5, while the same percentage was 40% for the music and 48% for the speech segments.

4.3. Class probability estimation

In order to achieve multi-class classification, the "One-vs-All" (OVA) classification scheme has been adopted. This method is based on decomposing the K-class classification problem into K binary sub-problems (Rifkin & Klautau, 2004). In particular, K binary classifiers are used, each one trained to distinguish the samples of a single class from the samples in the *remaining* classes. In the current work, we have chosen to use Bayesian Networks (BNs) for building these binary classifiers.

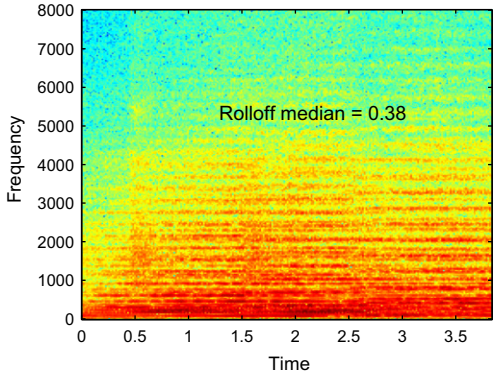

Fig. 5. Music spectrogram.

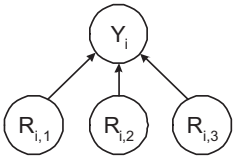

Fig. 6. BNC architecture.

At a first step, the 12 feature values described in Section 4.2, are grouped into three 4-D separate feature vectors (feature subspaces). In the sequel, for each one of the 7 binary sub-problems, three k-Nearest Neighbor classifiers are trained on the respective feature sub-space. This process leads to three binary decisions for each binary classification problem. Thus, a 7×3 kNN decision matrix R is computed. $R_{i,j}$ is 1 if the input sample is classified in class i, given the j^{th} feature sub-vector, and it is equal to 0 if the sample is classified in class "not i".

In order to decide to which class the input sample belongs, according to R, BNs have been adopted: each binary subproblem has been modeled via a BN which combines the individual kNN decisions to produce the final decision. In order to classify the input sample to a specific class, the kNN binary decisions of each subproblem (i.e. the rows of matrix R) are fed as input to a separate BN, which produces a probabilistic measure for each class. In this work, the BN architecture shown in Fig. 6, has been used as a scheme for combining the decisions of the kNN individual classifiers. Discrete nodes $R_{i,1}$, $R_{i,2}$ and $R_{i,3}$ correspond to the binary decisions of the kNN individual classifiers for the i-th binary subproblem and are called hypotheses of the BN, while node Y_i is the *output* node and corresponds to the true binary label. Y_i, like the elements of R, is 1 if the input sample really belongs to class i, and it is *0*, otherwise.

For each sample k, each BN i, makes the final decision for the i-th binary subproblem, based on the conditional probability:

$$P_i(k) = P\left(Y_i(k) = 1 | R_{i,1}^{(k)}, R_{i,2}^{(k)}, R_{i,3}^{(k)}\right) \quad (1)$$

This is the probability that the input sample's true class label is i, given the results of the individual kNN classifiers. After the probabilities $P_i(k)$, $i = 1,\ldots,7$ are calculated for all binary subproblems, the input sample k is classified to the class with the largest probability, i.e.:

$$WinnerClass(k) = \arg\max_i P_i(k) \quad (2)$$

This combination scheme can be used as a classifier, though, in this work we use it as a probability estimator for each one of the seven classes.

Table 3
Visual features description.

Feature	Description
AM	Average overall motion calculated using motion vectors.
MOV	Variance of the motion vectors orientations.
OTD	Average degree of overlap of the detected people.
MLD	Maximum luminance difference.
MLI	Maximum luminance interval.

5. Visual classification for violence hints

The problem of violence detection in videos is challenging because of the big variability of the violent scenes, and the unconstrained camera motion. It is impossible to accurately model the scene and the objects within. Instead, we define classes that represent the amount of human activity in the scene and use features that can discriminate the video segments between these classes. The amount of activity is strongly correlated with the existence of violence as can be seen from Fig. 7. The histograms in the Figure originate from a randomly selected dataset comprised of 25 movie segments each of them with duration of about 1 min. The dataset was manually labeled with respect to activity (high, normal, low) and with respect to violence (violence – non violence) to estimate the amount of correlation between them. We observed that most of the segments that contain high activity contain violence and vice versa.

5.1. Visual class definition

For the visual based characterization we define a label with respect to the amount of *activity* (low, normal, high). The "activity" classes are defined by the amount of human activity in the scene as no-normal-high activity. The 'low' label contains scenes that do not show humans or that show humans that are completely inactive. The 'normal' label indicates scenes in which one or more persons perform an activity that cannot be characterized as violent (e.g. walking, jogging, chatting). The 'high' label, which is strongly correlated with violence, contains scenes with people with abrupt

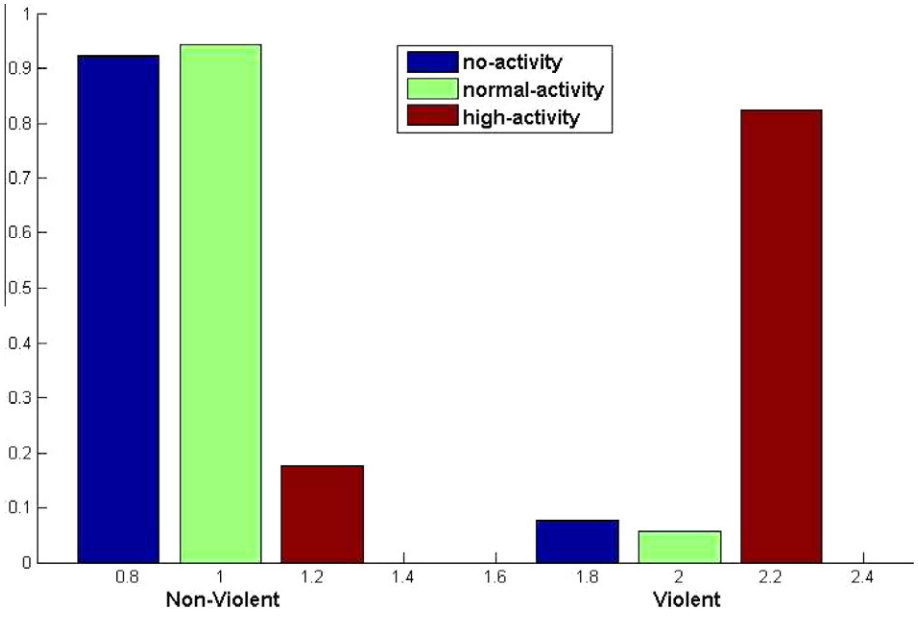

Fig. 7. Correlation between violence and activity classes: the first set of bars represents the percentage of segments with no, normal, and high activity which are labeled as non-violent whereas the second set contains the violent. As can be seen most of the segments that contain high activity are violent and vice versa. The plots are derived from a randomly selected dataset of hand-labeled movie segments.

motion (e.g. fighting, falling). Scenes with high activity tend to be violent as we validated by the experiment shown in Fig. 7. Furthermore, we defined one label indicating the presence of *gunshots* or not. The scenes containing gunshots are dominantly violent. In the following we describe the features that were used to extract the activity and gunshots-related labels.

5.2. Visual features

The used features can be split in three categories: the first is used to represent motion (AM, MOV). The motivation for the motion features is that most scenes with high activity contain abrupt motion. The second category indicates the presence of people and their trajectory in the scene (OTD). The third category is used to detect gunshots (MLD, MLI).

The visual signal is split into segments. Each segment comprises several video frames, and the visual features are calculated on every frame. We average over the values of each frame to derive the value of the feature which characterizes the whole segment. The visual features are summarized in Table 3 and are further analyzed in the following subsections.

5.2.1. Motion features

- *Average Motion (AM)*: This is the average motion of the frame. The frame is split in blocks of which we calculate the motion vectors using the previous frame as reference. The feature is derived by averaging the motion vector's magnitude. The motion vectors magnitude and the block size are determined as fractions of the frame size so that the feature will be invariant to frame scale changes. The feature is defined by:

$$AM = \frac{1}{N_b}\sum_{i=1}^{N_b} v_i \tag{3}$$

where N_b is the number of blocks and v_i is the length of the motion vector of the i^{th} block.

- *Motion Orientation Variance (MOV)*: The feature measures the variance of the motion vectors orientations. The mean orientation is derived by calculating the mean motion vector. Then the variance is calculated by the following:

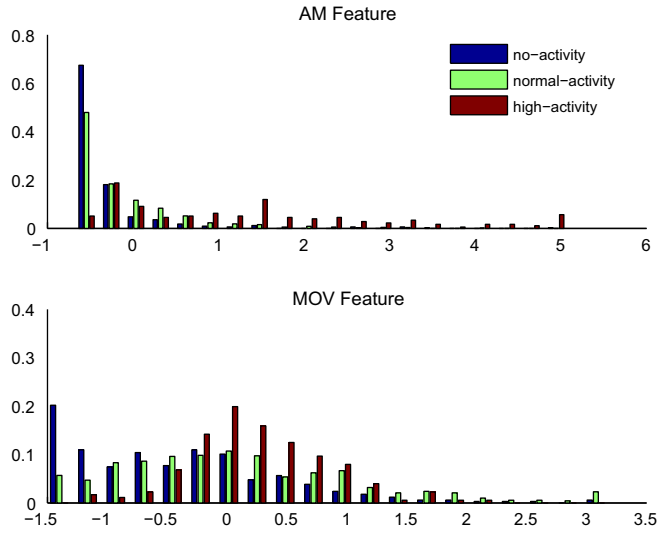

Fig. 8. *AM* and *MOV* feature histograms. The histograms show that the features contain information about the activity classes. The segments belonging to the high-activity class tend to have higher values for both features. The plots are derived from a randomly selected dataset of hand-labeled movie segments.

$$MOV = \frac{1}{N_b}\sum_{i=1}^{N_b} d_a^2(a_i, a_\mu) \tag{4}$$

where a_i is the orientation of the i-th motion vector, a_μ is the orientation of the mean motion vector and $d_a()$ denotes the difference between the two orientations in $(-\pi, \pi)$.

An example of AM and MOV feature histograms is given in Fig. 8.

5.2.2. Person detection features

Using an object recognition method we detect the presence of people or faces in the scene. To increase the robustness of the approach a color based skin detector is used to filter the previous results. The detected visual objects are tracked to establish the correspondences between them in consecutive frames. The tracked objects are used to derive a feature that measures their motion in the scene.

We used the object detection algorithm of Lienhart and Maydt (2002) which exploits shape information to detect people in the scene. The detector uses a set of haar-like features and a boosted classifier. For the training phase of the classifier the features are extracted from several positive and negative images. Then for a given image, features are extracted from several windows within the image at several scales and are passed to the classifier. The classifier's output for every window is '1' if the object is present or '0' otherwise. We trained four different classifiers to recognize frontal faces, profile faces, full body and, upper body and we run the detectors independently on each frame. The algorithm results in bounding boxes containing the detected objects in each frame.

To detect the skin colored regions we use the histogram based method described in Jones and Rehg (2002). Using an image dataset skin and non-skin pixels are manually annotated and the respective histograms in the RGB space are constructed. With these skin and non-skin histograms we can compute the probability that a given color value belongs to the skin and non-skin classes:

$$P(rgb|skin) = \frac{s[rgb]}{T_s}, \quad P(rgb|\neg skin) = \frac{n[rgb]}{T_n} \tag{5}$$

where $s[rgb]$ is the pixel count contained in bin rgb of the skin histogram, $n[rgb]$ is the equivalent count from the non-skin histogram, and T_s and T_n are the total counts contained in the skin and non-skin histograms, respectively. The next step is the computation of $P(skin|rgb)$, which is given by Bayes rule:

$$P(skin|rgb) = \frac{P(rgb|skin)P(skin)}{P(rgb|skin)P(skin) + P(rgb|\neg skin)P(\neg skin)} \tag{6}$$

A particular RGB value is labeled as skin if $P(skin|rgb) > \Theta$ where Θ is a threshold. $P(skin)$ and $P(\neg skin)$ are the prior probabilities for any color value being skin or non-skin, respectively. We only need to estimate one of these priors since they sum up to one. One choice for the prior probability of skin is the ratio of the total skin pixels in the histogram to the total of all the pixels:

$$P(skin) = \frac{T_s}{T_s + T_n} \tag{7}$$

For a given frame the output of the skin detector is a binary bitmap with the dimensions of the frame and with values one for the pixels that contain skin and zero for the rest.

To increase the robustness of our approach we filter the results of the face detection using the output of the skin detection algorithm. To this end, for each bounding box b_f containing a face we compute the ratio:

$$R = \frac{I_s(b_f)}{I_s(b_f) + I_n(b_f)} \tag{8}$$

where $I_s(b_f)$, $I_n(b_f)$ is the number of skin and non-skin pixels respectively contained in the b_f bounding box. If this is below a threshold T_f which is determined empirically the detected face is discarded.

The detected objects are tracked to establish the correspondences between detections in different frames. The detector runs in standard intervals and each time the produced bounding boxes are used to initialize the trackers. When the next interval starts we measure the degree of overlap between a tracked object and each detected object of the same type (e.g. face). If there is a significant overlap then the detected object is linked to the tracked object. This way the final output of the system is a set of objects, which belong to one of the four detected object types. Each object has a trajectory from the frame where first detected up to the frame where it has been lost by the tracker or the position of the tracker does not match with any new detection. The tracking algorithm used is the one described in Makris, Kosmopoulos, Perantonis, and Theodoridis (2007).

The tracked objects are used to derive a metric for their motion. The metric is calculated as the average degree of overlap over all the tracked objects between two consecutive frames. The degree of overlap for a tracked object is defined as:

$$OTD = \frac{1}{2}\left[\frac{I^t(b) - I^t_{int}(b)}{I^t(b)} + \frac{I^{t-1}(b) - I^t_{int}(b)}{I^{t-1}(b)}\right] \quad (9)$$

where: $I^t(b)$ is the number of pixels of the b^{th} tracked object's bounding box at time t, and $I^t_{int}(b)$ is the number of pixels of the intersection between the bounding boxes of the b^{th} tracked object at times t and $t-1$.

5.2.3. Gunshot detection features

A set of features to detect gunshots in the scene have been developed. The features detect the abrupt change in the illumination intensity caused by the fire that comes through the barrel during a gunshot. Therefore, to detect a gunshot with these features the gun should be within the field of view of the camera when it fires. The features are detailed in the following and an example is given in Fig. 9:

- *Maximum Luminance Difference (MLD):* This feature measures the amount of the short-term increase in luminance in an image region. To calculate the feature the image is split into blocks. The feature is defined as:

$$MLD = \max_b[MLD(b)] \quad (10)$$

where $MLD(b)$ denotes the value of the feature for block b defined as:

$$MLD(b) = \max_{i\in h(t)}\left[L^t(b) - L^i(n(b))\right] \quad (11)$$

where $h(t)$ denotes the set of previous (history) frames of t, $n(b)$ denotes the neighboring blocks to b, and $L^t(b)$ denotes the mean luminance of block b at frame t. This formula gives higher values for image blocks that have a large increase in their luminance compared to the luminance in their neighborhood on the previous frames.

- *Maximum Luminance Interval (MLI):* This feature measures the amount of short-term luminance fluctuations in an image region. It is defined by:

$$MLI(b) = \max_b[MLI(b)] \quad (12)$$

where $MLI(b)$ denotes the value of the feature for block b defined as:

$$MLI(b) = \min\left[\max_{i\in h(t)}\left(L^t(b) - L^i(n(b))\right), \max_{i\in f(t)}\left(L^i(n(b)) - L^t(b)\right)\right] \quad (13)$$

where $f(t)$ denotes the set of feature frames of t

5.3. Video class probability estimation

Two independent classifiers are used to classify each segment in one of the three activity and two gunshot classes. The classification of the segments in the activity classes is performed using a weighted kNN (k-Nearest Neighbor) classifier using the two motion features (*AM*, *MOV*) and the people overlap feature (*OTD*) described in Section 5.2. For the classification in the gunshot classes we use another weighted kNN classifier using the two gunshot related features (*MLD*, *MLI*). The activity classifier is trained using a dataset of hand-labeled scenes containing no, normal, or high human activity. Similarly the gunshot classifier is trained using a dataset containing scenes with and without gunshots. The output of the activity and gunshot classifiers is considered as an approximation to the probability of a given sample belonging to an activity or a gunshot class respectively.

6. Machine learning – based fusion

6.1. Fused feature vector

The seven audio class probabilities described in Section 4, along with the two visual-based class probabilities, described in Section 5, are combined in order to extract a final binary decision. This process is executed on a mid-term basis, i.e. in a sequence of successive segments from the original stream. In particular, *for each mid-term window of* 1 second *length, a 11-D feature vector is created* with the following elements:

- Elements 1–7: the *seven audio probabilities* of Eq. 1.
- Element 8: the label of the *winner audio* class
- Elements 9–11: the visual cues described in Section 5. In particular, they consist of the two soft classification decisions related to the *human activity* (i.e., the probability for normal and high activity), along with the *gunshots probability*.

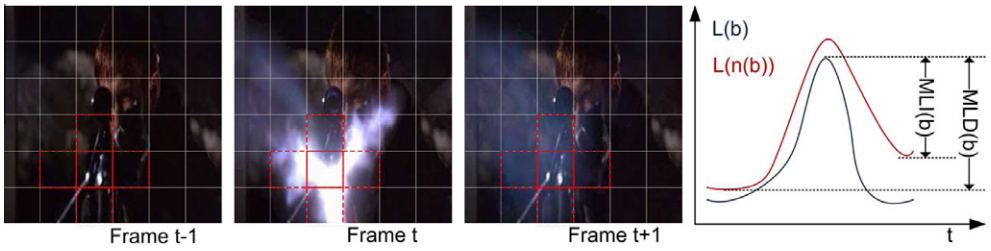

Fig. 9. Gunshot detection features. Three consecutive frames of a gunshot are shown. The frames are split in blocks. The diagram shows how the luminance on the highlighted block ($L(b)$) and its neighborhood ($L(n(b))$) change over time along with its respective $MLD(b)$ and $MLI(b)$ values for time t.

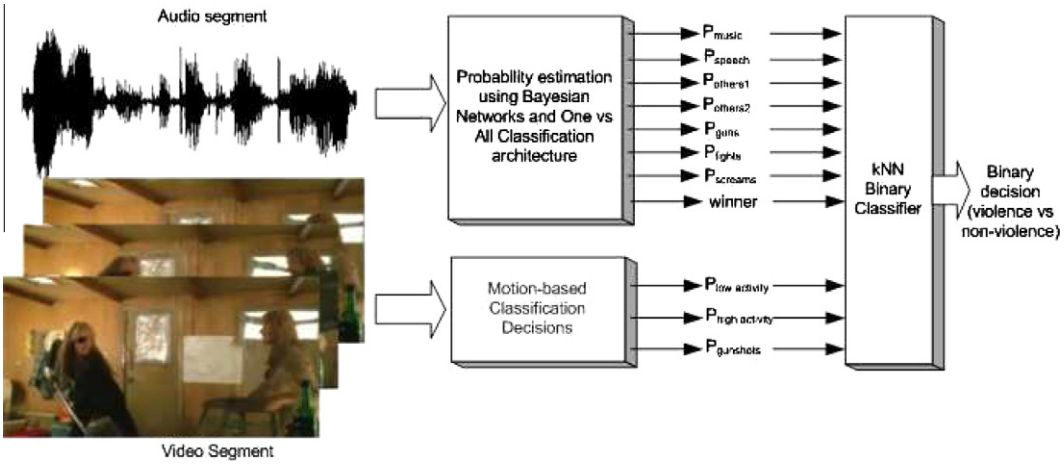

Fig. 10. Multi-modal fusion process: the probabilities of the seven audio classes, the winner audio class, and the probabilities of the three visual classes are used as a feature vector in the binary (violence vs non violence) meta-classification process.

6.2. Meta-classifier

The combined 11-D feature vector is used by a k-Nearest Neighbor classifier, which extracts the final binary (violence vs non-violence) decision, for the respective mid-term segment. The same process is repeated for all 1 second segments of the whole video stream. In Fig. 10 a scheme of this process is presented.

For comparison purposes, apart from the fused classifier, two individual kNN classifiers, an audio-based and a visual-based, have been trained, in order to distinguish between violence and non-violence, on a mid-term basis. In other words, these two individual classifiers have been trained on the *8*D feature sub-space (audio-related) and on the 3D feature sub-space (visual-related) respectively. In Fig. 11, an example of the violence detection algorithm is presented, when using (a) only audio features (b) only visual features and (c) the fused feature vector.

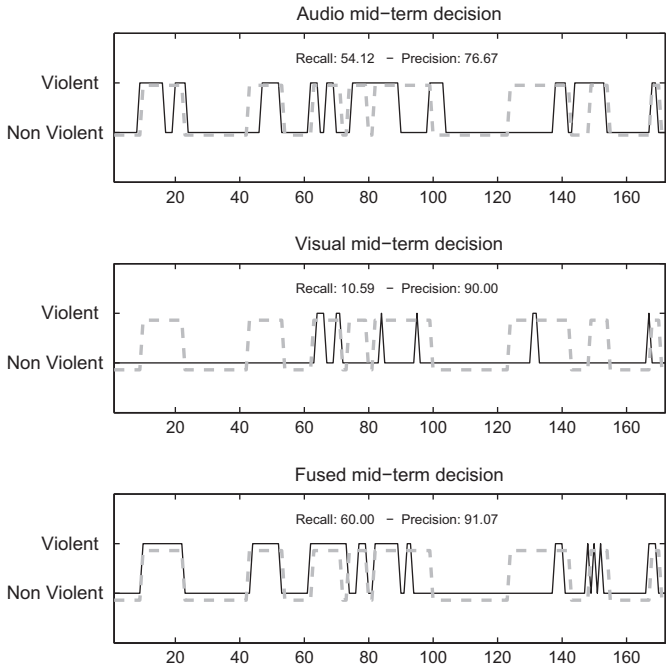

Fig. 11. Example of all three violence detection algorithms (audio, visual and fused) for a movie audio stream. The gray line corresponds to the true class labels over time. Also, for each case, the precision and recall rates are presented (computed on a mid-term basis: for example, as explained in the next Paragraph, mid-term based precision is the percentage of mid-term segments that have been classified as violence, and are indeed violence).

7. Ontological fusion

The ontological approach examined in this paper is a direct application of the ontological framework presented in Perperis et al. (2007) elaborating further the corresponding ontologies and the inferencing procedure. Deploying such an approach in a violence identification application does not aim towards boosting the overall detection accuracy, but towards extracting higher level of semantics and detailed descriptions of the movie at hand. According to the ontological methodology (Fig. 12), visual (Section 5) and audio analysis steps (Section 4) apply on one second audiovisual segments extracting thus violent clues, objects and primitive events/actions. The produced results populate segment based individuals to the corresponding ontologies. Finally the inference process exports semantics at the desired level.

7.1. Ontological framework

The ontological framework (Fig. 12) involves the definition of four different ontologies, interconnected using common terms, Web Ontology Language (OWL) property axioms, OWL restrictions and SWRL rules. 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). Ontological and rule reasoning combined in the inference engine map sets and sequences of violent hints to higher level violent events and concepts represented in the Violence Domain Ontology, thus instantiating automated violent-specific annotations for the movie in question. Although the ontology-based fusion approach focuses on tackling an *extensive range of complex violent actions* in video data, in this paper our primary target remains the binary violence detection problem.

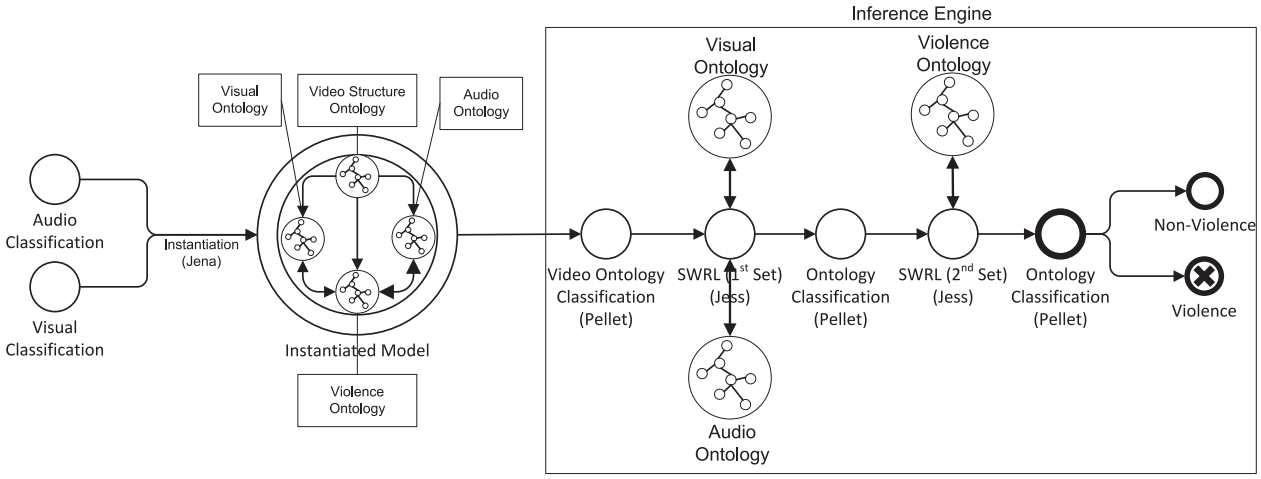

Fig. 12. Ontological framework: low level audio and visual algorithms produce mid-level semantics for every 1 s segments generating thus proper ontology individuals. Ontological and SWRL reasoning combined in the 5-Step inference engine exploit the instantiated model towards classifying segments either as violent or as non-violent.

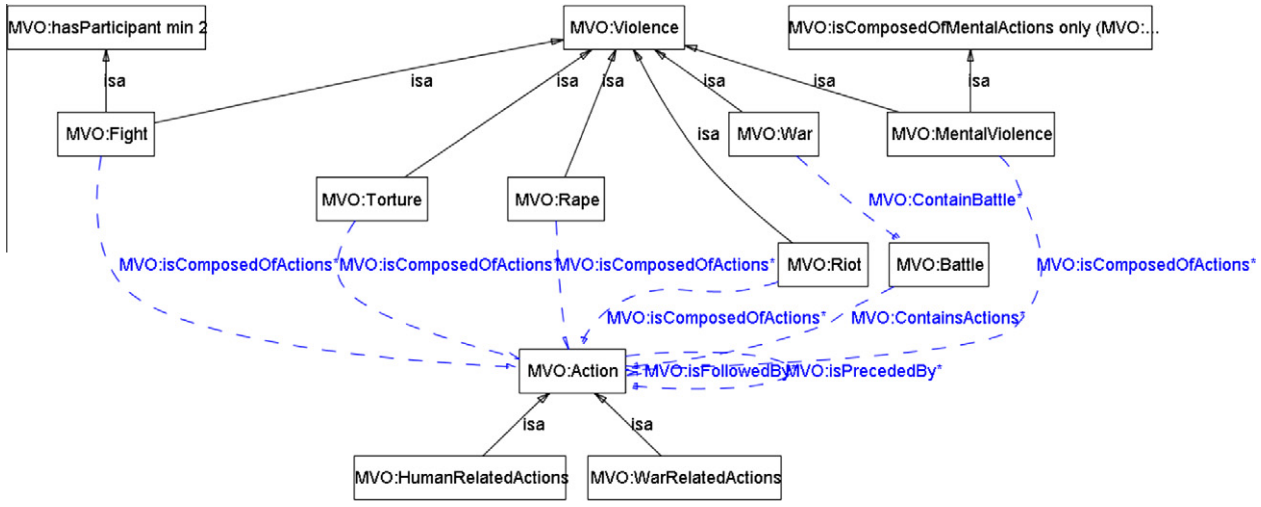

Fig. 13. Violence ontology: represent violence domain as an hierarchy of concepts composed from spatiotemporal related objects, events and actions. In this figure we demonstrate the upper level concepts of the Violence ontology.

7.2. Violence ontology definition

Defining the abstract concept of violence as an hierarchy of high level concepts is not an easy task. We had to take a lot of decisions, starting from the elementary question *What is violence?* We may define as violence any action or situation, that may cause pain, suffering, physical or mental harm to one or more persons, injury to animals, fire, explosion or destruction of non-living objects. Violent content in video data refers to scenes that include such actions. An effective formal representation of the violence domain knowledge in all its complexity, abstractness and hierarchy depth, to drive violent acts identification, has never been attempted before. We make a step forward towards this direction. The ontology definition arises as the result of extensive investigation through observation of violent acts in video data collected from movies, streaming videos, news and security camera captures.

Violence Domain Ontology[1] involves the definition (Fig. 13) of a broad class of concepts as a set of related spatiotemporal entities (i.e. actions, events, objects) irrespective of low level analysis algorithms. However low level analysis algorithms are not yet capable of extracting the broad set of modality specific semantics (i.e. guns, swords, vehicles, contextual objects, body parts, emotional state, simple actions, events, etc.) the ontological definition prescribes. Thus we are forced to constrain the set of Violence Domain Ontology concepts exploited based on the aforementioned (Section 4, 5) mid-level semantics. Further taking under consideration open world reasoning in OWL (i.e. everything is true unless there is a definition that proves is false), we cannot identify non-violence directly using simple negation as failure reasoning. Thus in violence domain ontology we have also incorporated the definition of non violent classes. Summarizing the violent classes of interest are *MVO:Fighting*, *MVO:ShotsExplosions*, *MVO:Screaming* and the non-violent *MVO:Dialogue*, *MVO:ActionScenes* (i.e. people running, car chases etc.), *MVO:Scenery* (audiovisual segments not displaying people and with no action).

7.3. Visual semantics for violence

The Visual Ontology[2] for violence prescribes, in a hierarchical way, a set of objects/events – possibly related with violence – that low level visual analysis algorithms may automatically extract.

[1] MVO: Stands for *movie violence ontology* and used as prefix for every element of the violence domain ontology.

[2] MVisO: Stands for *Movie Visual Ontology* and used as prefix for every element of the Visual ontology.

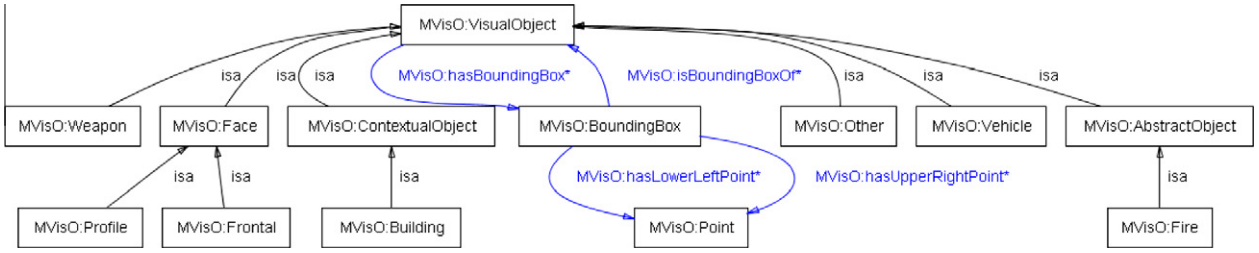

Fig. 14. Visual objects: part of the visual ontology representing the hierarchy of objects extracted from visual modality.

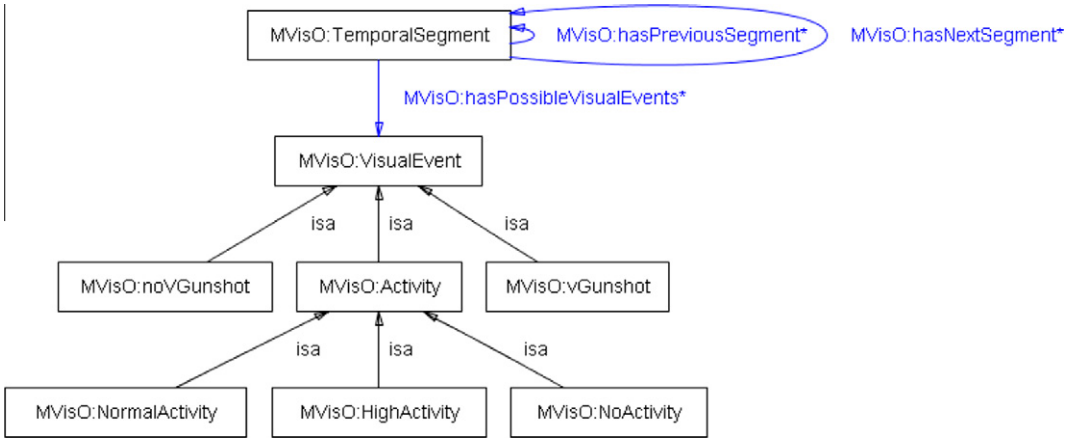

Fig. 15. Visual Events: part of the visual ontology representing the hierarchy of mid-level events extracted from visual modality.

Although a much broader set of visual objects and events is already defined, for interoperability reasons with the low level visual analysis step reasons we focus our attention on the extracted visual objects/events. Thus the simple human visual events (Fig. 15) (See Fig. 16) and visual objects classes (Fig. 14) of interest (as defined in Section 5.1) are *MVisO:HighActivity*, *MVisO:NormalActivity*, *MVisO:NoActivity*, *MVisO:Gunshot*, *MVisO:noGunshot* and *MVisO:Face*.

Visual object detection and tracking from the low level visual analysis perspective is a complex procedure producing uncertain results. Although such an approach is informal (OWL does not yet support uncertainty since the development of PR-OWL (Costa et al., 2008), is still in progress) we have defined the *MVisO:hasProbability* datatype (float) property to capture and exploit this uncertainty in our inferencing procedure.

7.4. Audio semantics for violence

Additional clues, relative to violence, increasing the accuracy of violence detection exist in the auditory modality. Contrary to visual semantics in audio semantics every class that is defined in the ontology is also extracted from the audio classification algorithms. Thus (as defined in Section 4.1) the audio classes of interest are *MSO:Screams*,[3] *MSO:Speech*, *MSO:Gunshot*, *MSO:SharpEnviromentalSound*, *MSO:SmoothEnviromentalSound* and *MSO:Fights*.

Low level, fixed segment, classification algorithms (Section 4) extract the aforementioned semantics and instantiate the corresponding audio semantic classes. The estimated probability of existence is stored using the *MSO:hasProbability* datatype (float) axiom.

7.5. Video structure ontology

For interoperability among the domain and modality specific ontologies reasons the overall ontological methodology demands for a Video Structure Ontology[4] capturing video temporal and structural semantics along with authoring information. In Fig. 17 we demonstrate the core elements of VSO.

VSO:MultimediaDocument: The class of documents containing combinations of audio, visual and textual information. Currently the only implemented subclass is the one representing movies. However this class offers an attachment point for a broad set of interesting subclasses like streaming videos, web pages etc.

VSO:TemporalSegment: Every movie segment exploiting either the auditory or visual modality or both to convey meaning and having some temporal duration. In our case every movie is decomposed in a set of consecutive audiovisual segments of one second. The subclass *VSO:AudioVisualSegment* serves as the main interconnection point with the low level analysis algorithms. This class and its properties are the first to be instantiated in order to initiate the inferencing procedure.

[3] MSO: Stands for *Movie Sound Ontology* and used as prefix for every element of the Sound Ontology.

[4] VSO: Stands for *Video Structure Ontology* and used as prefix for every element of the Video Structure ontology.

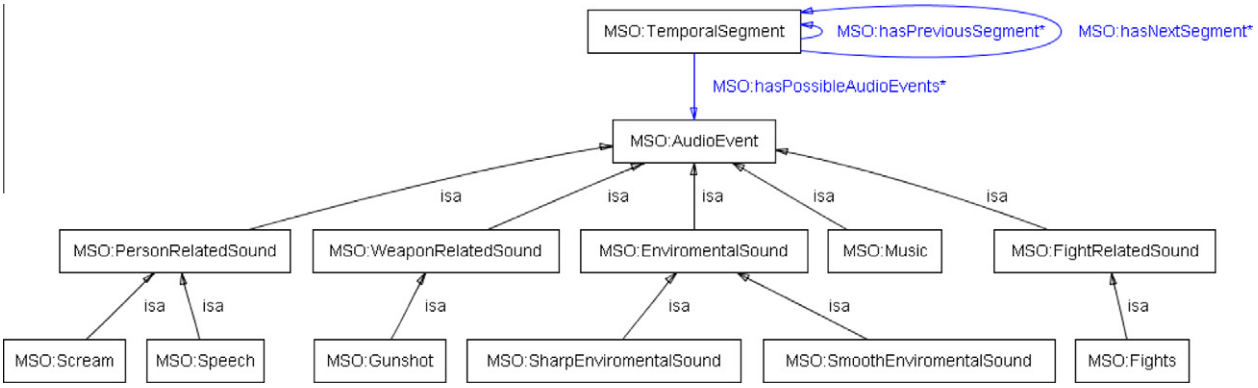

Fig. 16. Audio events: part of the visual ontology representing the hierarchy of mid-level events extracted from audio modality.

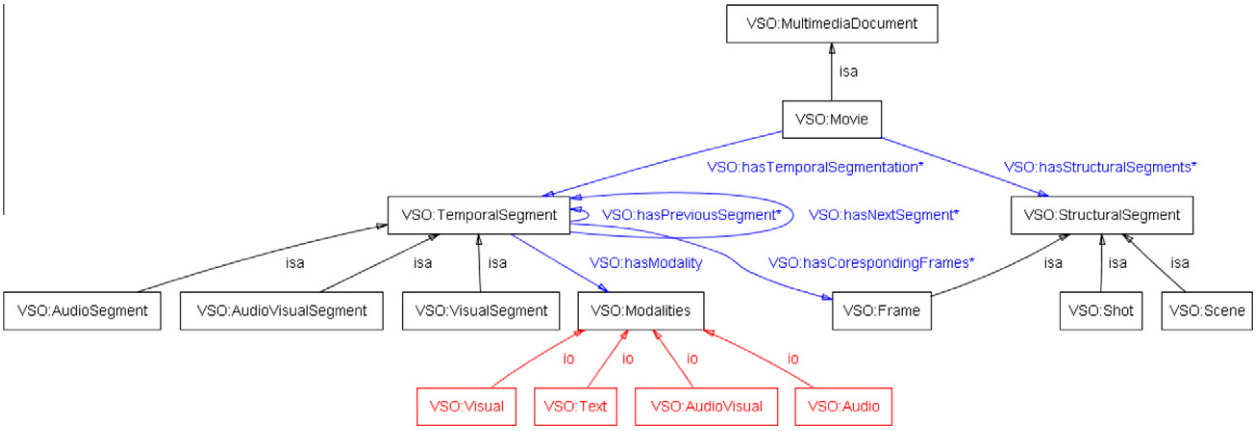

Fig. 17. Video structure ontology: represents authoring information and structural decomposition of the movie (e.g. frames, segments, shots) in question.

VSO:StructuralSegment: The class of elementary structural and logical segments of a movie. Every movie is a sequence of shots (single camera captures) and every shot is a sequence of images (frames). Logically every movie is a sequence of scenes each of one is composed from consecutive shots. Similarly with *VSO:AudioVisualSegment*, *VSO:Frame* and its property axioms serve as an interconnection point with the low level analysis algorithms. This class and it's properties instantiation aim towards initiating the inferencing procedure.

VSO:Movie: Multimedia documents containing at least one temporal segment and at least one Structural segment. Datatype properties like VSO:hasTitle, VSO:hasTotalFrameNumber, VSO:hasFrameRate define the authoring information of the movie.

7.6. Inference engine design

Having the adequate low to medium level semantics extracted from single modality analysis algorithms, and the corresponding loosely coupled, using common terms, equivalent classes and object property axioms ontological definitions, we have to tackle the issue of *fusing and interchanging semantics from different modalities* towards inferring more complex, abstract and extensive violent cases represented in the domain ontology. Thus it is required that apart from each modality's ontological description, there is a need for a cross-modality ontological description increasing the semantic extraction potentials of our application. We investigate the usage of SWRL rules (Horrocks et al., 2004) combined in a 5-step inferencing procedure (Fig. 12) with ontological reasoning.

Ontological reasoning is usually performed using one of the existing reasoners (in our case Pellet Sirin, Parsia, Cuenca Grau, Kalyanpur, & Katz, 2007) that implement consistency checking, classification and instance checking services. SWRL (the standard rule language of the semantic web) written in a form of Horn like rules, can reason about OWL instances (individuals) in terms of OWL classes and properties. Although manual SWRL rule construction is a tough procedure, especially for such a complicated field like violence detection, we use them for cross-modality representation of violent spatio-temporal relations and behavioral patterns. Since SWRL does not support reasoning with uncertainty we model such information by thresholding the corresponding datatype relations with SWRL builtin axioms.

7.7. Inference engine implementation

In this subsection we will further elaborate on the implementation details of the proposed inferencing procedure (Fig. 12). Speaking in terms of Description Logic (DL) modality specific, video structure and violence domain ontologies define the Terminological Box[5] (TBox) of our knowledge based system. Towards

[5] Terminological Box is the conceptual representation of the domain in question.

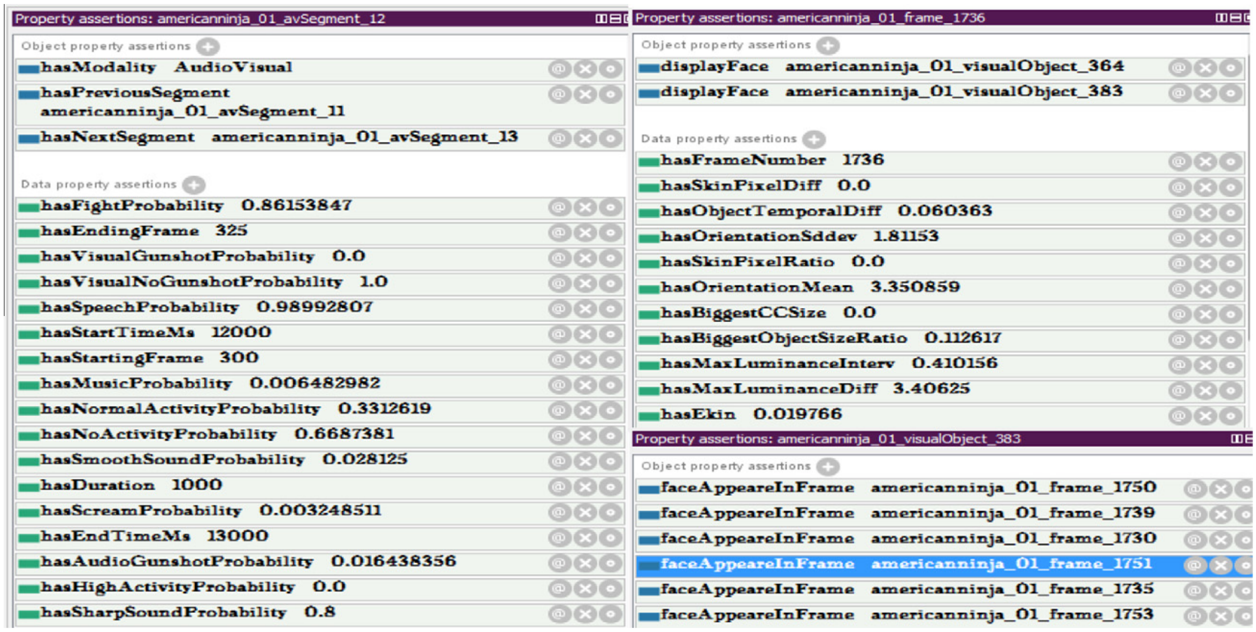

Fig. 18. Instatiation example: object and datatype property instantiations of a video segment, a frame and a visual object as extracted from low level analysis.

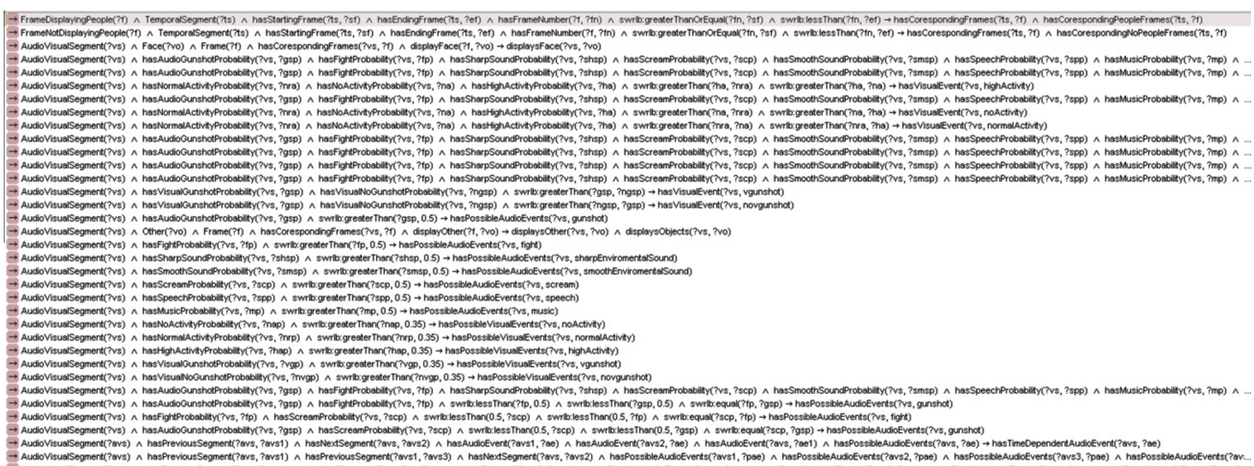

Fig. 19. First SWRL rule set.

applying higher level reasoning services the definition of the Assertional Box[6] is mandatory, taking in consideration that inferencing procedure attempts to reclassify individuals on the violence domain ontology. In the examined field of application the ABox derives directly from the modality specific low level analysis algorithms. Having the exported audiovisual segments, frames, events and objects along with the corresponding low level numerical values and probability results the modality specific and the video structure ontologies are instantiated. In Fig. 18 we demonstrate sample instantiations of a visual segment, a randomly chosen frame and an object appearing in a set of frames.

The instantiated model, essentially captures existing and extracted knowledge for the movie in question. Based on this knowledge a 5-step inferencing procedure combining ontological reasoning and SWRL rules is applied. Each reasoning step on ontologies produces new knowledge that is possible to trigger the execution of a new rule set. Similarly each rule execution produces new knowledge exploited by subsequent ontology reasoning steps.

Step-1: This is the first reasoning service applied on the instantiated model, and in coordination with Step-2 serves as the initialization step. During ontology instantiation every individual is subclassed under owl:Thing, thus the main purpose of this step is to compute the initial class of every individual based on defined TBox conditions, and check the consistency of the instantiated model.

Step-2: The first SWRL rule set (Fig. 19), composed of 35 rules, is applied. In this step arithmetic probabilities of mid level semantics existence are properly assigned with classes of the modality specific ontologies. Additionally instantiation of axioms relating temporal with structural segments is performed (i.e. audiovisual segments with the corresponding frames).

[6] Assertional Box is the set of individual definitions in terms of concepts, roles and axioms defined in the TBox.

Table 4
Identifying a non violent (Action) and a violent (Fighting) segment.

Description	SWRL rule	Description	SWRL rule
If		**If**	
?avs is an Audio Visual Segment	VSO:AudioVisualSegment (?avs) ∧	?avs is an Audio Visual Segment	VSO:AudioVisualSegment (?avs) ∧
?ms is an individual of audio class Music	MSO:Music (?ms) ∧	?f is an individual of audio class Fights	MSO:Fights (?g) ∧
?ha is an individual of visual class HighActivity	MVisO:HighActivity (?ha) ∧	?f is an individual of visual class HighActivity	MVisO:HighActivity (?ha) ∧
In ?avs music is detected	VSO:hasAudioEvent (?avs,?ms) ∧	In ?avs fights are detected	VSO:hasAudioEvent (?avs,?f) ∧
In ?avs high activity is detected	VSO:hasVisualEvent (?avs,?ha) ∧	In ?avs high activity is detected	VSO:hasVisualEvent (?avs,?ha) ∧
Then		**Then**	
?avs is an action segment	→ MVO:ActionScene (?avs)	?avs is a fights segment	→ MVO:Fighting (?avs)

Table 5
Identifying a violent PersonOnPersonFighting and MultiplePersonFighting.

MVO:Fighting Subclasses Definition	Necessary and Sufficient Conditions
MVO:PersonOnPersonFighting	MVO:displaysObjects some MVisO:Face ∧ MVO:displaysObjects exactly 2
MVO:MultiplePersonFighting	MVO:displaysObjects some MVisO:Face ∧ MVO:displaysObjects min 3

Step-3: Having new knowledge (i.e. reclassified audiovisual segments) generated from Step-2, consistency checking and classification services of the new model are applied.

Step-4: The second SWRL rule set (composed of 93 rules) is applied. In Table 4 we demonstrate two sample cases of implemented rules. This is the core step of our inferencing procedure. Mid level semantics, as inferred in previous steps, are combined towards classifying every extracted video segment in one of the violence domain ontology classes. Although in this paper we aim towards tackling the binary classification problem (i.e. violence–non violence) the second SWRL rule set is producing higher level of semantics by classifying segments in three non violent (*MVO:Action, MVO:Dialogue, MVO:Scenery*) and 3 violent (*MVO:Fighting, MVO:Screaming, MVO:ShotsExplosions*) classes. Towards creating this rule set we have exploited common sense logic for cross modal reasoning, temporal relations of extracted mid level semantics and simplified conclusions drawn from the audio classification confusion matrix.

Step-5: In this final step consistency of the ontology is checked once again and classification is performed towards computing violent and non violent segments (instances reclassify from children to parents) from the one hand and extended semantics (instances reclassify from parents to children) from the other hand. Since the first classification case is straightforward we will further describe the second case using a simple example. In Table 5 we demonstrate the definition of *MVO:PersonOnPersonFighting* and *MVO:MultiplePersonFighting* using necessary and sufficient conditions. During classification every fighting segment displaying exactly two persons (identified by their faces) is finally classified as instance of *MVO:PersonOnPersonFighting* and every fighting segment displaying more than three persons is finally classified as instance of *MVO:MultiplePersonFighting* producing thus higher level of semantics. In addition following the path from an instance to the root (owl:Thing) of the ontology a complete description of the corresponding segment is produced.

8. Experimental evaluation

8.1. Implementation issues

The audio classification feature extraction, the multi-class probability estimation processes and the machine learning-based fusion method have been implemented in Matlab. The visual-related features have been extracted using the OpenCV library (http://sourceforge.net/projects/opencvlibrary/).

Regarding the ontological framework, each ontology was defined in OWL using Protégé (http://protege.stanford.edu) and every SWRL rule using the corresponding SWRLtab (http://protege.cim3.net/cgi-bin/wiki.pl?SWRLTab). Pellet (Sirin et al., 2007) is used for ontology reasoning services application and (Jess, 2008) for SWRL rules execution. Jena semantic web framework (Jena, 2002) is used for ontology instantiation and synchronization of the aforementioned reasoning steps.

8.2. Scenario and setup

For training and evaluation purposes, 50 videos have been ripped from 10 different films. The overall duration of the data is 2.5 hours i.e., almost 3 min average duration per video. The video streams have been manually annotated by three humans. In particular, the humans annotated the parts of the videos that contained violent content and the results of this annotation have been used as ground truth for training and evaluating the proposed methods. According to the manual annotations 19.4% of the data (on a mid-term basis) was of violent content. For estimating the performance of the classification procedure, the fused feature values (described in Section 6), along with the respective true class labels (for each 1-second segment) have been used. Therefore, almost 9000 mid-term segments were, in total, available for training and testing. However, the evaluation (as described in the following Section) was carried out on a video stream basis.

8.3. Classification and detection results

In this Section the results of the proposed binary classification method are presented, along with the ontology-based decision method. The results of the fused classification and the ontology-based performance are compared to the individual performances, if only the audio and the visual features were used. In all three methods, the "Leave One Out" evaluation method has been used, on a video file basis, i.e., in each cross-validation loop, the mid-term segments of a single video file have been used for evaluation, while the rest of the data has been used for training purposes.

The following types of performance measures have been computed:

(1) Classification Precision (P): This is the proportion of mid-term segments that have been classified as violence and were indeed violence.

(2) Classification Recall (R): This is the proportion of mid-term violence segments that were finally classified as violent.
(3) Classification F_1 measure: This measure is computed according to the equation: $F_1 = \frac{2 \cdot P \cdot R}{P+R}$.
(4) Detection Precision (Pd): This is the number of detected violence segments, that were indeed violence, divided by the total number of detected violence segments.
(5) Detection Recall (Rd): This is the number of correctly detected violence segments divided by the total number of *true* violence segments.
(6) Detection Fd_1 measure: This measure is computed according to the equation: $Fd_1 = \frac{2 \cdot Pd \cdot Rd}{Pd+Rd}$.

Performance measures P, R and F_1 are associated to the *classification* performance of the algorithm on a mid-term (1 second) basis, while the measures Pd, Rd and Fd_1 are related to the *detection* performance of the algorithm. Note that a violent segment is correctly detected if it overlaps with a true violent segment. In addition, for comparison purposes, we have computed the same performance measures for the random mid-term classifier.

The following observations are directly extracted from the performance results shown in Table 6:

(1) Audio classification and detection is respectively almost 1.5% and 6% better than the visual-based method.
(2) Combined classification achieves a *boosting* at the overall performance by 0.5% for the classification and by almost 5.5% for the detection mode, compared to the best individual method (i.e., the audio-based method).
(3) The overall *detection* method achieves a 84.28% recall rate and a 43.43% precision rate. This means that only 25% of the violent events are not detected, while almost 1 out of 2 detected segments are indeed violent ones.
(4) The ontology-based approach performs better that the visual-based decision, though, it only achieves a small boosting (in terms of mean accuracy) in the case of event detection. However, the advantage of using this approach is that we can extract higher level of semantics using an unsupervised procedure and common sense reasoning. In Table 7 we demonstrate recall and precision results for the ontology based fusion regarding the Fights, Screams and Shots-Explosions classes. These results are drawn only from correctly classified as violence segments. We notice that for the Fight class we achieve the best results and for the Screams class the worst. This happens because for the identification of the Fight class audio and visual analysis algorithms produce the most accurate hints. Contrarily Scream and Shots-Explosions identification are not actually aided from the visual modality results.

Table 6
Classification and detection performance measures.

	Recall (%)	Precision (%)	F_1 (%)	Mean (%)
	Classification performance measures			
Audio-based classification	63.2	45.2	52.7	54.2
Visual-based classification	65.1	40.7	50.1	52.9
Random classification	19	50	28	34.5
Fused classification	60.1	47	53.2	53.5
Ontology-based	71.5	35.8	47.7	53.7
	Detection performance measures			
Audio-based detection	82.9	38.9	53	61
Visual-based detection	75.6	34	46.9	54.8
Fused detection	83	45.2	58.5	64.1
Ontology-based	91.2	34.2	50	62.7

Table 7
Violent event based classification performance measures using the ontology-based approach.

	Recall (%)	Precision (%)	F_1 (%)	Mean (%)
	Classification performance measures			
Fights classification	61.6	68.2	64.8	64.9
Screams classification	41.4	33.5	37.1	37.4
Shots-explosions classification	63.3	38.2	47.6	50.7

9. Conclusions and future work

In this work we presented our research towards the detection of violent content in movies. We have presented a set of violence-related features both from the visual and audio channel. For fusion purposes, we have adopted a simple meta-classification technique and an ontology-based method. A large dataset has been populated for training and testing the respective methods. The experimental procedure indicated that the combined classification achieves a boosting at the overall performance by almost 5.5% for the detection mode, compared to the best individual method (i.e., the audio-based method). Furthermore, the fact that the overall *detection* recall rate was almost 75%, while almost 1 out of 2 detected segments were indeed violent, indicates that the proposed method can successfully be used for detecting violence in real movies. On the other hand, the ontology-based fusion achieves a very small improvement, only as far as the mean accuracy of the detection mode is concerned. However, the advantage of using the ontology approach is that a higher level of semantics can be extracted without any further training (i.e., unsupervised procedure).

In the future, we are planning to implement more sophisticated combination schemes for the meta-classification stage. Furthermore, new visual and/or audio features may be computed, always focused on the detection of violent content. The visual module can be extended to provide accurate detections and trajectories of various visual objects that can be used to infer the existence of violence such as people or weapons. Furthermore visual features that are capable of modeling several specific types of violence that are not detected with the more general features used here should be considered. As far as the ontology-based approach is concerned, taking under consideration that visual analysis does not yet produce semantics at the desired level (i.e. object detection, body parts movement, complex and simple event detection), the ontological fusion approach result seem really promising for the violence identification problem. Future work towards improving ontological fusion results involves from the low level analysis aspect implementation of properly designed object – event/action detectors and from the ontological aspect formal exploitation of uncertainty and automatic SWRL rule learning and generation. Towards increasing interoperability with other applications we examine the potentials of anchoring the Violence Domain Ontology in one of the existing core ontologies and the modality specific ontologies in one of the existing Mpeg-7 ontologies. Further exploiting Mpeg-7 ontologies we aim towards automatically producing complete and standardized movie annotations.

Acknowledgment

This work has been supported by the Greek Secreteriat for Research and Technology, in the framework of the PENED program, Grant No. TP698.

References

Ángel Vidal, M., Clemente, M., & Espinosa, P. (2003). Types of media violence and degree of acceptance in under-18s. *Aggressive Behavior, 29*(5), 381–392.

Babaguchi, N., Kawai, Y., & Kitahashi, T. (2002). Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia, 4*(1), 68–75.

Bai, L., Lao, S., Jones, G. J. F., & Smeaton, A. F. (2007). Video semantic content analysis based on ontology. In *IMVIP '07: Proceedings of the international machine vision and image processing conference* (pp. 117–124). National University of Ireland, Maynooth (NUIM): IEEE Computer Society.

Bao, J., Cao, Y., Tavanapong, W., Honavar, V., & Honavar, V. (2004). Integration of domain-specific and domain-independent ontologies for colonoscopy video database annotation. In *International conference on information and knowledge engineering (IKE 04)* (pp. 82–88). CSREA Press.

Bertini, M., Del Bimbo, A., & Torniai, C. (2005). Automatic video annotation using ontologies extended with visual information. In *Proceedings of the 13th annual ACM international conference on multimedia, MULTIMEDIA'05* (pp. 395–398). Hilton, Singapore: ACM.

Bolles, B., & Nevatia, R. (2004a). *ARDA event taxonomy challenge project final report*.

Bolles, B., & Nevatia, R. (2004b). *A hierarchical video event ontology in OWL, ARDA challenge workshop report*.

Cheng, C.-C., & Hsu, C.-T. (2006). Fusion of audio and motion information on hmm-based highlight extraction for baseball games. *IEEE Transactions on Multimedia, 8*(3), 585–599.

Costa, P.C., Laskey, K.B., & Laskey, K.J. (2008). Pr-owl: A bayesian ontology language for the semantic web. In *Uncertainty reasoning for the semantic web I: ISWC international workshops, URSW 2005–2007, revised selected and invited papers* (pp. 88–107).

Ernest Friedman-Hill – Sandia National Laboratories. (2008). Jess – the rule engine for the java platform. Available: <http://www.jessrules.com/>.

Fan, J., Luo, H., Gao, Y., & Jain, R. (2007). Incorporating concept ontology for hierarchical video classification, annotation, and visualization. *IEEE Transactions on Multimedia, 9*(5), 939–957.

Francois, A. R. J., Nevatia, R., Hobbs, J., & Bolles, R. C. (2005). VERL: An ontology framework for representing and annotating video events. *IEEE MultiMedia, 12*(4), 76–86.

Giannakopoulos, T., Kosmopoulos, D. I., Aristidou, A., & Theodoridis, S. (2006). Violence content classification using audio features. In *Proceedings of the 4th Helenic conference on AI, advances in artificial intelligence, SETN 2006. Lecture notes in computer science* (Vol. 3955, pp. 502–507). Heraklion, Crete, Greece: Springer.

Giannakopoulos, T., Pikrakis, A., & Theodoridis, S. (2007). A multi-class audio classification method with respect to violent content in movies, using bayesian networks. In *IEEE international workshop on multimedia signal processing* (pp. 90–93). Crete.

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). *Swrl: A semantic web rule language combining owl and ruleml*. URL: <http://www.w3.org/Submission/SWRL/>.

Hsu, W., Kennedy, L., Huang, C. W., Chang, S. F., Lin, C. Y., & Iyengar, G. (2004). News video story segmentation using fusion of multi-level multi-modal features in trecvid 2003. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP'04)* (Vol. 3, pp. 645–648).

Iyengar, G., Nock, H., Neti, C., & Franz, M. (2002). Semantic indexing of multimedia using audio, text and visual cues. In *Proceedings of the 2002 IEEE international conference on multimedia and expo (ICME)* (pp. 369–372).

HP Labs. (2002). Jena – A semantic web framework for java. Available: <http://jena.sourceforge.net/index.html>.

Jones, M. J., & Rehg, J. M. (2002). Statistical color models with application to skin detection. *International Journal of Computer Vision, 46*(1), 81–96.

Kevin D Browne, C. H.-G. (2002). American film violence, an analytic portrait. *Journal of Interpesonal Violence, 17*(4), 351–370.

Kevin D Browne, C. H.-G. (2005). The influence of violent media on children and adolescents: A public-health approach. *Lancet, 365*, 702–710.

Lehane, B., O'Connor, N. E., & Murphy, N. (2004). Action sequence detection in motion pictures. In *Knowledge-based media analysis for self-adaptive and agile multi-media, Proceedings of the European workshop for the integration of knowledge, Semantics and digital media technology, QMUL, London, UK* (pp. 25–26).

Leonardi, R., Migliorati, P., & Prandini, M. (2004). Semantic indexing of soccer audio–visual sequences: A multimodal approach based on controlled markov chains. *IEEE Transactions on Circuits and Systems for Video Technology, 14*, 634–643.

Li, Z., & Tan, Y.-P. (2005). Event detection using multimodal feature analysis. In *IEEE international symposium on circuits and systems* (Vol. 4, pp. 3845–3848). IEEE.

Lienhart, R., & Maydt, J. (2002). An extended set of Haar-like features for rapid object detection. In *Proceedings of the international conference on image processing, Rochester, New York, USA* (pp. 900–903).

Lin Huang, C., Chia Shih, H., & Yuan Chao, C. (2006). Semantic analysis of soccer video using dynamic bayesian network. *IEEE Transactions on Multimedia, 8*(4), 749–760.

Makris, A., Kosmopoulos, D., Perantonis, S. S., & Theodoridis, S. (2007). Hierarchical feature fusion for visual tracking. In *IEEE international conference on image processing (ICIP), San Antonio, Texas, USA* (Vol. 6, pp. 289–292).

Nam, J., Alghoniemy, M., & Tewfik, A. H. (1998). Audio–visual content-based violent scene characterization. *International conference on image processing* (Vol. 1, pp. 353–357). IEEE.

Nam, J., & Tewfik, A. H. (2002). Event-driven video abstraction and visualization. *Multimedia Tools and Applications, 16*, 55–77.

Neumann, B., Möller, R. (2008). Cognitive vision. *Special Issue on Image and Vision Computing, 26* (1), 82–101.

Perperis, T., Tsekeridou, S., & Theodoridis, S. (2007). An ontological approach to semantic video analysis for violence identification. In *Proceedings of I-media'07 and I-Semantics 2007. International conferences on new media technologies and semantic technologies (Triple-i: i-Know, i-Semantics, i-Media), best paper award in multimedia metadata applications (M3A) workshop* (pp. 139–146).

Rasheed, Z., & Shah, M. (2002). Movie genre classification by exploiting audio–visual features of previews. In *Proceedings of the 16th international conference on pattern recognition* (pp. 1086–1089).

Reidsma, D., Kuper, J., Declerck, T., Saggion, H., & Cunningham, H. (2003). Cross document ontology based information extraction for multimedia retrieval. In *Supplementary proceedings of the international conference on computational science, Dresden*.

Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research, 5*, 101–141.

Sirin, E., Parsia, B., Cuenca Grau, B., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web, 5*(2), 51–53.

Snidaro, L., Belluz, M., & Foresti, G. L. (2007). Domain knowledge for surveillance applications. In *Proceedings of the 10th international conference on information fusion* (pp. 1–6).

Theodoridis, S., & Koutroumbas, K. (2008). *Pattern recognition* (4th ed.). Orlando, FL, USA: Academic Press Inc.

Vasconcelos, N., & Lippman, A., (1997). Towards semantically meaningful feature spaces for the characterization of video content. In *Proceedings of the 1997 international conference on image processing (ICIP'97)* (3-Volume Set-Volume 1, pp. 25). Washington, DC, USA: IEEE Computer Society.

Zajdel, W., Krijnders, J. D., Andringa, T. C., & Gavrila, D. M. (2007). CASSANDRA: Audio–video sensor fusion for aggression detection. In *IEEE conference on advanced video and signal based surveillance* (pp. 200–205). London: IEEE Computer Society.