High Level Multimodal Fusion and Semantics Extraction

Thanassis Perperis[**]

National and Kapodistrian University of Athens
Department of Informatics and Telecomunications
a.perperis@di.uoa.gr

Abstract. This thesis on multimodal fusion and semantics extraction, focuses on automated detection and annotation of harmful content in video data. The aim is not only to reason out the existence of violence or not (i.e. the binary problem), but also to determine the type of violence (e.g. fight, explosion, murder). Acknowledging the lack of knowledge representation and reasoning approaches for the problem at hand, we propose a semantic fusion approach that combines low to mid level modality specific semantics through ontological and rule reasoning. A major part of the proposed framework is the movie segmentation into meaningful and easy to handle units. We evaluate a set of shot boundary detection approaches combined through a majority voting scheme. In the sequel, state of the art classification methods are employed to extract audio and visual mid level semantics. The segmentation and modality specific analysis algorithms instantiate the corresponding video structure and modality specific ontologies developed in the context of the knowledge engineering framework. A set of consecutive and interleaved ontological and SWRL rule reasoning steps map sets and sequences of extracted low to mid level semantics into higher level concepts represented in the harmful content domain ontology. We present the involved ontologies, the corresponding SWRL rule sets and the reasoning mechanism in detail. Finally we present the evaluation of the proposed approach in a preanotated movie dataset, compare its results with the single modality approaches and a kNN late fusion meta classifier. We comment on the higher level semantics extraction ability and evaluate a set of extensions employed in the basic structure of the framework. The extensions concern the development of a scene detection module that combines markov clustering with SQWRL queries, the incorporation of existing rating and movie genre metadata in the violence identification procedure and the detection of pornography.

Keywords: Harmful Content Detection, Semantic Video Analysis, Knowledge Engineering, Ontologies, Rules, Reasoning

[**] Dissertation Advisor: Professor Sergios Theodoridis

1 Introduction

Detecting Harmful Content in video data, which are complex in nature, multimodal and of significant size is not an easy task and requires extensive and efficient analysis. Although quite different, in terms of the exploited modalities and methodology specific details, most of the proposed approaches in the literature fall in the *pattern recognition* discipline, following either or single- or a multi modal- approach. In the multimodal approach, the modality fusion is performed either at the feature level (i.e. early fusion) or at the decision level (i.e. late fusion).Examining the reported research on harmful content detection, we conclude that the extracted semantics are not at the desired level for higher level harmful content filtering applications. Most approaches tackle either the binary or constrained instances of the multiclass problem and are rather tailor made. In addition, there is a lack of interest on incorporating automatic annotation processes in terms either of MPEG-7[1] or of ontological representations. In this thesis we tackle violence and pornography detection by means of a unified approach exporting various levels of semantic abstractions, employing state of the art audio, visual and textual mid-level concept detectors, using domain knowledge, ontologies and reasoning for higher level semantics inferencing, annotation and filtering. Deploying ontologies and semantic web trends in the context of video analysis for automatic annotation and filtering is a very challenging task. Current ontological approaches appear mostly to enhance semantic searching applications in simplified and constrained domains like medical [1, 2], sports [3, 4] and surveillance [5] videos whereas limited work appears in the domain of movies and TV series for harmful content detection. An ad hoc ontological approach, exploiting Video Event Representation Language (VERL) [6] along with Video Event Markup Language (VEML), for surveillance, physical security and meeting video event detection applications, emerged as a result of the ARDA event taxonomy challenge project [7]. In addition, an interesting approach aiming towards extracting and capturing the hierarchical nature of actions and interactions, in terms of *Context Free Grammars (CFG)* rather than ontologies, from raw image sequences appears in [8]. To the best of our knowledge, the only approaches, directly employing multimedia and domain ontologies for multimodal video analysis, in the context of harmful content identification, are our works for violence detection presented in [9, 10] and [11, 12].

2 Knowledge-based framework for violence identification in movies

The presented ontological (Fig. 1) approach[2] is a direct application of the proposed framework in [9], further elaborating, extending and implementing

[1] http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm

[2] This work has been supported by the Greek Secretariat for Research and Technology, in the framework of the PENED program, grant number TP698.

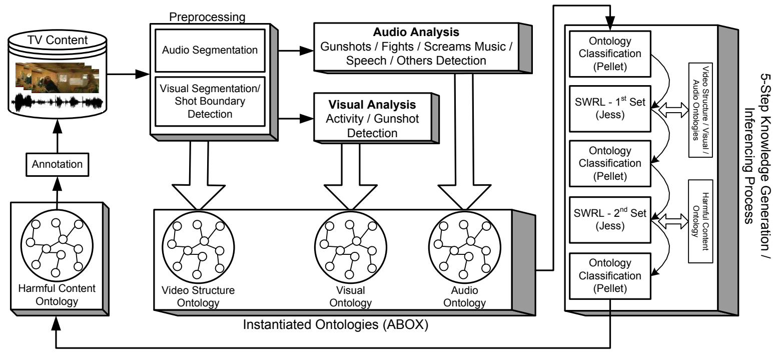
Preprocessing
TV Content
Audio Segmentation
Visual Segmentation/ Shot Boundary Detection
Audio Analysis
Gunshots / Fights / Screams Music / Speech / Others Detection
Visual Analysis
Activity / Gunshot Detection
Annotation
Harmful Content Ontology
Video Structure Ontology
Visual Ontology
Audio Ontology
Instantiated Ontologies (ABOX)
Ontology Classification (Pellet)
SWRL - 1st Set (Jess)
Ontology Classification (Pellet)
SWRL - 2nd Set (Jess)
Ontology Classification (Pellet)
Video Structure / Visual / Audio Ontologies
Harmful Content Ontology
5-Step Knowledge Generation / Inferencing Process

Fig. 1. Proposed Methodology For Violence Detection

each of the involved modules. Our aim is not to devise high quality low level analysis processes, but to propose an open and extendable framework, combining existing single modality low to mid-level concept detectors with ontologies and SWRL[3] rules, to identify and formally annotate an *extensive range of complex violent actions* in video data. We identify as major processes of the system, a preprocessing/segmentation step, a visual analysis step, an audio analysis step, each one interconnected with the corresponding data ontology, an inferencing procedure and a harmful content domain ontology. Speaking in terms of knowledge engineering, low level analysis extracts *basic facts* - *basic truth* that holds for the corresponding TV/video content data, while ontologies define *complex* and *terminological* facts that hold for the examined domain in general. Thus an explicit knowledge base is formed and the scope of the 5-step inferencing procedure is to draw new implicit conclusions/knowledge during each step. In the following, we further elaborate on the implementation decisions adopted for each of the aforementioned processes.

2.1 Preprocessing - Segmentation Semantics

Preprocessing tackles the task of temporal audio visual segmentation and feeds the corresponding low level analysis algorithms. The role of segmentation is three-fold: 1) Define the temporal annotation units, 2) Feed low level analysis with temporal segments of predefined duration, 3) Preserve a common time reference and extract sequence and overlapping relations, among visual and audio segments, for the temporal reasoning procedures. Thus, both modalities are initially segmented into fixed duration segments (1-sec as the minimum event duration), as defined from the low level modality specific algorithms, and then grouped into the corresponding shots to preserve the synchronicity and time reference, further forming the annotation units. The task of shot

[3] Employed Ontologies and SWRL rules are available through http://hermes.ait.gr/PenedHCD/

boundary detection is performed by means of a majority voting local content adaptive thresholding approach. In short, the grayscale pixel difference, the RGB pixel difference, the RGB and HSV histogram differences, the RGB and HSV histogram similarities based on a combined measure, the edge change ratio and the average motion through motion vectors are computed for each pair of consecutive frames. Further the mean values of the afforomentioned measures are computed on a 5-D temporal window. If the current difference is maximum in the examined window and is greater than twice the mean window value, a shot cut is detected for the examined measure. Finally a shot cut is detected when the majority of approaches satisfies the adaptive thresholding condition. Experimentation of the shot boundary detection module on a violent movie dataset achieves 95.18% for precision and 91.01% for recall.

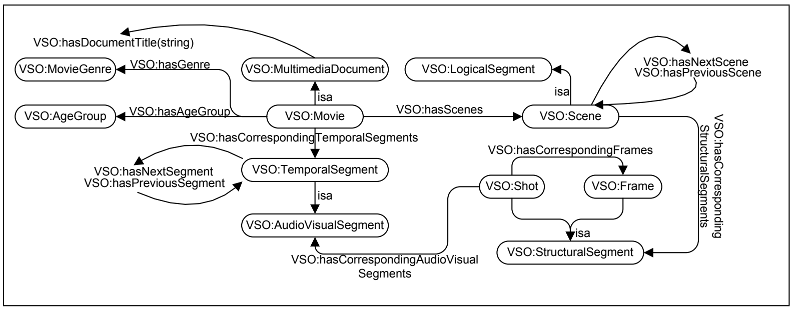

Fig. 2. Video Structure Ontology

For interoperability of the segmentation process reasons with modality specific and domain ontologies, the overall methodology demands for a Video Structure Ontology (Fig. 2), capturing temporal and structural semantics along with authoring information. In short, an example of the core classes is *VSO:MultimediaDocument*[4] that defines the class of documents containing combinations of audio, visual and textual information. Although our final target is to provide complete ontological representation for a multitude of multimedia document (e.g. streaming videos, web pages, images, audio recordings) currently, the only implemented subclass is the one representing movie/TV series content, defined to include individuals instantiating the corresponding property axioms with at least one temporal and at least one structural segment. In addition, Datatype properties (i.e. VSO:hasDocumentTitle, VSO:hasTotalFrameNumber, VSO:hasFrameRate) capture authoring information. For further interoperation with existing metadata annotations (e.g. TV-Anytime) providing overall content description, in terms of intended age groups or genre classification, corresponding axioms *VSO:hasAgeGroup*, *VSO:hasGenre* were included in the ontological

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

definition. *VSO:TemporalSegment* categorizes segments (in our case of 1-sec duration) exploiting either the auditory or visual modality or both to convey meaning associated with a temporal duration. Structurally every video is a sequence of shots (i.e. single camera capturing) and every shot is a sequence of frames. Logically, every video is a sequence of semantically meaningful scenes, each one composed of consecutive shots. Thus *VSO:StructuralSegment* defines content's elementary structural and *VSO:LogicalSegment* subsumed logical segments. The last two classes and their property axioms serve as the main interconnection point with the low level analysis algorithms and are the first to be instantiated, initiating thus the inferencing procedure. Obviously there is a strong interconnection of the video structure ontology with an MPEG-7 one. Thus we have incorporated, with property axioms, an extended with the MPEG-7 audio part version of Jane Hunter's ontology [13], aiming further for automated MPEG-7 based annotation.

2.2 Audio Visual Semantics

To optimally combine multimedia descriptions with the violence domain ontology, the knowledge representation process has further defined modality violence ontologies (audio, visual) that essentially map low-level analysis results to simple violent events and objects (medium-level semantics). Adopting a top down approach, the modality specific ontologies comprise an important "guide" for low level analysis algorithms. Namely, they define what to search or try to extract from raw data. Although concept detectors using statistical, probabilistic and machine learning approaches, are extensively studied in the literature, it is not yet possible to extract whatever a knowledge engineering expert prescribes. Consequently, the corresponding ontologies contain a broader set of potential for extraction concepts, further prioritizing low level analysis research towards developing novel concept detectors. Taking under consideration the intrinsic difficulties of low level analysis, providing erroneous results in some cases, the long-term target is to connect a broad set of concepts/events/object detectors with the modality specific ontologies, providing thus a cooperative mechanism towards increasing the detection accuracy.

Visual Semantics The Visual Ontology (Fig. 3) defines, in a hierarchical way, the set of objects/events - possibly related with harmful content - identified during visual analysis. Although a much broader set of low to mid level semantics is already defined, for interoperability with the employed visual analysis algorithms reasons, we focus our attention on the actually extracted visual clues. In particular, each fixed duration segment is classified in one of three activity (low, normal, high) and two gunshot (gunshot, no-gunshot) classes. Thus, the simple human events and visual objects ontological classes of

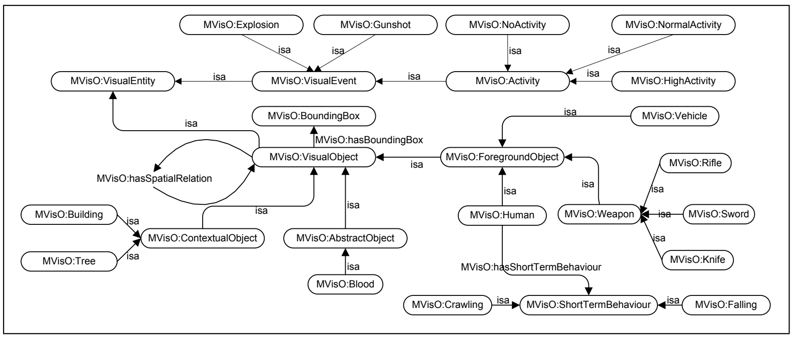

Fig. 3. Visual Ontology - Objects And Events

interest are *MVisO:HighActivity*[5], *MVisO:NormalActivity*, *MVisO:NoActivity*, *MVisO:Gunshot*, *MVisO:noGunshot* and *MVisO:Face.*

The activity detection process initially computes, in a per segment basis, the *average overall motion*, *variance of motion* and *average degree of overlap of the detected people (i.e. faces)* features. The former ones are computed by the corresponding *motion vectors* while the latter by the face detection and tracking component. The face detection component employs a set of Haar-like features and a boosted classifier to detect people in the scene while the tracking component employs the hierarchical fusion algorithm devised in [14]. To increase face detection robustness, a histogram based skin detection algorithm filters-out objects with minimum skin content. Finally, the classification process is performed by means of a weighted kNN (k-Nearest Neighbor) classifier that computes the segment likelihood to attain high, normal or no activity content. Measuring abrupt changes in the illumination intensity can provide valuable hints on the existence of fire, explosions or gunshots. Thus, the *maximum luminance difference* and *maximum luminance interval* features are used to train a distinct weighted kNN classifier discriminating between gunshot and no-gunshot segments. For further details on the visual analysis components the reader is referred to [11].

Audio Semantics Additional clues increasing the harmful content detection accuracy exist in the auditory modality. Contrary to visual, in audio semantics the ontological definition covers the full set of extracted mid level semantics. Thus the audio classes (Fig. 4) of interest are *MSO:Screams*[6], *MSO:Speech*, *MSO:Gunshot*, *MSO:Fights*, *MSO:SharpEnviromentalSound* and *MSO:SmoothEnviromentalSound* . The audio analysis process involves a variant of the “One-vs-All” (OVA) classification scheme presented in [15], on a segment

[5] MVisO: Stands for *Movie Visual Ontology* and is used as prefix for every element of the Visual Ontology

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

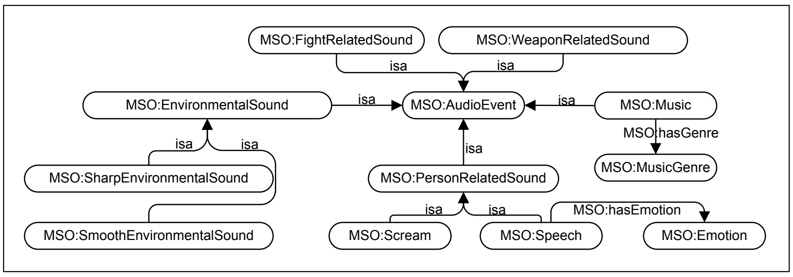

Fig. 4. Audio Ontology

basis, in order to generate a sequence of three violent (*Gunshots*, *Fights* and *Screams*) and four non - violent (*Music*, *Speech*, *Others1*: environmental sounds of low energy and almost stable signal level like silence, background noise, etc., and *Others2*: environmental sounds with abrupt signal changes like thunders or closing doors) audio class probabilities for each segment.

2.3 Domain Ontology Definition

An effective formal representation of the harmful content domain knowledge in all its complexity, abstractness and hierarchy depth, to drive corresponding acts detection, has never been attempted before. We have made a step forward towards this direction. The ontology definition has resulted from an extended investigation through close observation of mostly violent and some pornographic acts in video data collected from TV programs, movies, streaming videos, news and security camera captures. The Harmful Content Domain Ontology[7] (Fig. 5) includes the definition of numerous high level concepts as a set of related spatiotemporal entities (i.e. actions, events, objects) irrespective of low level analysis algorithms. Taking under consideration the inability of employed audio and visual analysis algorithms to extract the broad set of violence related modality specific semantics (i.e. guns, swords, vehicles, contextual objects, body parts, emotional state, simple actions, events, etc.) we are forced to limit our description and experimentation on the set of attained mid-level semantics. In addition due to *open world reasoning* in OWL, we cannot identify non-violence directly using simple *negation as failure* reasoning. Therefore the domain ontology further aims on the identification of a number of non harmful classes like dialogue, actions/activity and scenery.

2.4 Inferencing Procedure

Having the extracted low- to mid- level semantics, and the corresponding loosely coupled, using common terms, equivalent classes and object property

[7] HCO: Stands for *Harmful Content Ontology* and used as prefix for every element of the Domain Ontology.

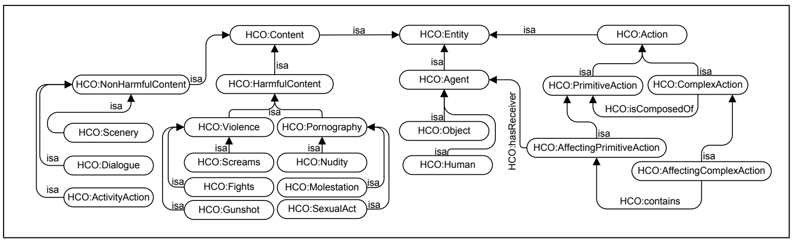

Fig. 5. Harmful Content Ontology

axioms, ontological descriptions, we have to tackle with 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 there is a need for a cross-modality ontological association mechanism between modality specific and domain ontologies, further increasing the semantic extraction capabilities of low level analysis. Towards this direction we investigate the usage of SWRL rules[8] combined in a 5-step inferencing procedure with ontological reasoning. SWRL (a Horn-like rule language combining OWL with RuleML), on the one hand, can reason about OWL instances (individuals) in terms of OWL classes and properties and is mostly used to increase the OWL expressivity. Ontological reasoning, on the other hand, implements consistency checking, classification and instance checking services, which is usually achieved using one of the existing reasoners (in our case Pellet[9]). Although manual SWRL rule construction is a tough procedure, especially for such a complicated domain like harmful content detection, we explore the potential for cross-modality representation of violent spatio-temporal relations and behavioral patterns. Since SWRL and OWL do not yet formally support reasoning under uncertainty, we attempt to capture such semantics by simply thresholding the corresponding datatype relations with SWRL built-in axioms.

Speaking in terms of Description Logic (DL), the involved ontologies define the conceptual representation of the domain in question, forming thus the *Terminological Box (TBox)* of our knowledge base. Taking into consideration that the inferencing procedure attempts to reclassify modality individuals to the violence domain ontology, the definition of the *Assertional Box (ABox)* - the set of individual definitions in terms of TBox's concepts, roles and axioms - is mandatory. In the examined field of application, the ABox derives directly from the segmentation and audio/visual analysis algorithms, forming the basic facts in terms of individuals representing segments, frames, events and objects along with the corresponding low level numerical values and occurrence/accuracy probability results. Essentially, the instantiated model captures existing and extracted knowledge for the TV content in question. Based on this explicit knowledge, the 5-step inferencing procedure is applied. Each

[8] http://www.w3.org/Submission/SWRL

[9] http://clarkparsia.com/pellet/

ontological reasoning step draws implications that could possibly trigger the execution of a new rule set. Similarly, each rule execution implies fresh knowledge exploited by subsequent ontological reasoning steps. Namely, the process is as follows:

Step-1: During ontology instantiation, individuals are subclassed under *owl:Thing*, thus the main purpose of this step is to check the consistency of the instantiated VSO model and assert each individual's initial class, according to TBox's necessary and sufficient conditions.

Step-2: The first SWRL rule set, composed of 35 rules, is applied. Datatype property axioms expressing arithmetic probabilities and numerical values are translated to object property axioms, instantiated with classes and instances from the modality specific ontologies (e.g. assign the winner class). In addition, instantiation of the extracted video structure is performed (i.e. shots are related with the corresponding segments and segments with the corresponding frames).

Step-3: Consistency checking and classification services on the implied modality specific models are applied.

Step-4: The second SWRL rule set composed of 39 rules is applied. Implied audio and visual mid level semantics are combined using common sense logic, spatiotemporal relations and simplified conclusions drawn from low level analysis modules (i.e. confusion matrices) for cross-modality reasoning, further mapping video segments in one of the domain ontology classes.

Step-5: In this final step, consistency checking and classification services on HCO are applied to infer violent and non violent segments (instances reclassify from children to parents) on the one hand and extended semantics (instances reclassify from parents to children) on the other hand. Since the first classification case is straightforward (e.g. every fight segment is also a violent one) we will further describe the second case using a simple example.

2.5 Implementation and Experimentation

The presented ontological approach was developed using Matlab and the OpenCV library[10] for audio/visual feature extraction and classification, Protégé[11] for the definition of ontologies and SWRL rules, Pellet and Jess[12] for ontology reasoning services and rules execution and finally Jena semantic web framework[13] for ontologies instantiation and synchronization of the knowledge generation lifecycle. For evaluation purposes, 50 videos have been extracted from 10 different movies. The total duration of the test data is 2.5 hours. The video streams have been manually annotated by three persons, using the Anvil[14] video annotation research tool to generate ground truth data for performance evaluation. According to manual annotations 19.4% of the data was of violent content.

[10] http://sourceforge.net/projects/opencvlibrary/

[11] http://protege.stanford.edu

[12] http://www.jessrules.com/

[13] http://jena.sourceforge.net/

[14] http://www.anvil-software.de/

	Recall	Precision	F_1	Mean Accuracy
Audio-binary	82.9%	38.9%	53%	61%
Visual-binary	75.6%	34%	46.9%	54.8%
Violence Inference	91.2%	34.2%	50%	62.7%
Fights Inference	61.6%	68.2%	64.8%	64.9%
Screams Inference	41.4%	33.5%	37.1%	37.4%
Shots-Explosions Inference	63.3%	38.2%	47.6%	50.7%

Table 1. Segment based Binary and Multiclass Detection Performance Measures

Tables 1 and 2 demonstrate the achieved experimental results for the binary and multiclass violence detection problem in a per segment (1-sec duration) and in a per shot basis. For comparison with the single modality approaches purposes in table 1, we further demonstrate the achieved performance of low level audio and visual analysis algorithms for the binary problem. The performance measures are computed by means of ***Precision*** (i.e. the number of correctly detected violence segments, divided by the total number of detected violence segments), ***Recall*** (i.e. the number of correctly detected violence segments divided by the total number of **true** violence segments), *Mean Accuracy* and F_1 *measure* ($F_1 = \frac{2 \cdot P \cdot R}{P+R}$).

	Recall	Precision	F_1	Mean Accuracy
Violence Inference	61.07%	68.80%	64.7%	64.93%
Fights Inference	68.72%	89.9%	77.89%	79.31%
Screams Inference	25.0%	41.17%	31.10%	33.08%
Shots-Explosions Inference	89.39%	40.68%	55.91%	65.03%

Table 2. Shot based Binary and Multiclass Detection Performance Measures.

3 Enhanced Knowledge-based Framework

In this section we briefly sketch a set of elementary extensions deployed in the initial ontological framework. In order to increase the semantics detection ability from the one hand and the detection accuracy from the other hand we feed the semantic framework with the results of three pairs of Support Vector Machines and Multilayer Perceptron audio classifiers, trained using Principal Components Analysis on 160-D feature vectors extracted from the initial 7 class dataset, a 9 class music genre dataset and a 4 class pornography related dataset. Further epxloiting the shot boundary information along with log average luminance and a trained gaussian skin classifier for explosions/fire/gunshots components and large skin areas for pornography detection. Importing those mid level demands for new SWRL rule construction and results in increased performance for the violece detection case. Experimentation on a pornographic movie dataset results in 87.01% and 92.44% recall rate for the 1-sec and the shot case respectively. Finally we employ unsupervised markov clustering along with SQWRL queries to tackle scene detection. We define scenes as consecutive shots with audiovisual and semantic coherence. Audio visual coherence is achieved through markov unsupervised clustering applied on a complete weighted shot graph. Vertex distance is defined as $w_{i,j} = dist(\bar{x_i}, \bar{x_j}) = (1.0 - e^{\gamma \cdot \|\bar{x_i} - \bar{x_j}\|^2}) \cdot e^{-\frac{(t_i - t_j)}{k}}$ where $\bar{x}$ is a 38-D multimodal feature vector. The MCL algorithm retrieves graph node clusters of maximum similarity and instatiates the coresponding video structure ontology class. At the end of the knowledge generation lifecycle a set of SQWRL queries is applied to retrieve the set of consecutive semantically equivalent shot instances that belong in the same cluster and define a scene.

4 Conclusions

In this thesis we have proposed a complete ontological framework for harmful content identification and experimentally evaluated the violence detection case. The presented approach performs better than the employed audio and visual ones, for the 1-sec segment based binary violence detection problem, by means of recall rate, and achieves a small boosting in terms of mean accuracy. The performance is significantly increased both for the binary and the multiclass problem in the shot based approach. We notice that for the *fight* class we achieve the best results whereas for the *screams* class the worst. This happens on the one hand because audio analysis algorithms produce the most accurate hints for fights identification and on the other hand because visual analysis does not actually aid screams identification. Summarizing the low value of attained results is greatly affected by the following facts: *i*) Extracted visual analysis clues are not at the desired level, since activity classification is rather generic for specific concepts identification and detection of violence related objects (i.e. guns, swords, knifes) and human actions (i.e. punch, kick) remains unfeasible. *ii*) Extracted audio and visual mid level clues are biased towards non-violence (i.e. five violent classes: audio-gunshot, screams, fights, high activity, visual-gunshot and seven non-violent: music, speech, smooth environmental sound, sharp environmental sound, no activity, normal activity, no gunshot). *iii*) Uncertain single modality results are treated as certain. We conclude that the attained results are really promising both for the binary and multiclass violence detection problem and that the main advantage of using such an ontological approach still remains the higher level semantics extraction ability, using an unsupervised procedure and common sense reasoning.

References

1. Jie Bao, Yu Cao, Wallapak Tavanapong, and Vasant Honavar. Integration of Domain-Specific and Domain-Independent Ontologies for Colonoscopy Video Database Annotation. In *Proceedings of the International Conference on Information and Knowledge Engineering*, pages 82–90, Las Vegas, Nevada, USA, 21-24 Jun. 2004. CSREA Press.
2. Jianping Fan, Hangzai Luo, Yuli Gao, and Ramesh Jain. Incorporating Concept Ontology for Hierarchical Video Classification, Annotation, and Visualization. *IEEE Transactions on Multimedia*, 9(5):939–957, 2007.

3. Marco Bertini, Alberto Del Bimbo, and Carlo Torniai. Automatic Video Annotation Using Ontologies Extended with Visual Information. In *Proceedings of the 13th ACM International Conference on Multimedia*, ACM Multimedia, Singapore, 6-11 Nov. 2005. ACM.
4. Liang Bai, Songyang Lao, Gareth J. F. Jones, and Alan F. Smeaton. Video semantic content analysis based on ontology. In *Proceedings of the 11th International Machine Vision and Image Processing Conference*, pages 117–124, Maynooth, Ireland, 5-7 Sept. 2007. IEEE Computer Society.
5. Lauro Snidaro, Massimo Belluz, and Gian Luca Foresti. Domain Knowledge for Surveillance Applications. In *Proceedings of the 10th International Conference on Information Fusion*, pages 1–6, 2007.
6. Alexandre R. J. Francois, Ram Nevatia, Jerry Hobbs, and Robert C. Bolles. VERL: An Ontology Framework for Representing and Annotating Video Events. *IEEE MultiMedia*, 12(4):76–86, 2005.
7. Bob Bolles and Ram Nevatia. A Hierarchical Video Event Ontology in OWL, ARDA Challenge Workshop Report, 2004.
8. Michael S. Ryoo and J. K. Aggarwal. Semantic Understanding of Continued and Recursive Human Activities. In *Proceedings of the 18th International Conference on Pattern Recognition*, pages 379–382, Hong Kong, 20-24 Aug. 2006. IEEE Computer Society.
9. Thanassis Perperis, Sofia Tsekeridou, and Sergios Theodoridis. An Ontological Approach to Semantic Video Analysis for Violence Identification. In *Proceedings of I-Media'07 and I-Semantics'07. 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*, pages 139–146, Graz, Austria, 5-7 Sept. 2007.
10. Thanassis Perperis and Sofia Tsekeridou. A knowledge engineering approach for complex violence identification in movies. In *Artificial Intelligence and Innovations 2007: from Theory to Applications, Proceedings of the 4th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI 2007)*, volume 247, pages 357–364, Peania, Athens, Greece, 19-21 Sep. 2007. Springer.
11. Thanassis Perperis, Theodoros Giannakopoulos, Alexandros Makris, Dimitrios I. Kosmopoulos, Sofia Tsekeridou, Stavros J. Perantonis, and Sergios Theodoridis. Multimodal and Ontology-based Fusion Approaches of Audio and Visual Processing for Violence Detection in Movies. *Expert Systems with Applications*, 38(11):14102 – 14116, October 2011.
12. Thanassis Perperis and Sofia Tsekeridou. *TV Content Analysis*, chapter TV Content Analysis and Annotation for Parental Control. CRC Press, Taylor Francis, 2011.
13. Jane Hunter. Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In *Proceedings of the First Semantic Web Working Symposium*, pages 261–281, Stanford, USA, 2001.
14. Alexandros Makris, Dimitris Kosmopoulos, Stavros S. Perantonis, and Sergios Theodoridis. Hierarchical Feature Fusion for Visual Tracking. In *IEEE International Conference on Image Processing*, volume 6, pages 289 – 292, San Antonio, Texas, USA, 16 Sept. - 19 Oct. 2007.
15. Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. A Multi-Class Audio Classification Method With Respect To Violent Content In Movies, Using Bayesian Networks. In *Proceedings of IEEE International Workshop on Multimedia Signal Processing*, 2007.