Ontology-Driven Semantic Video Analysis Using Visual Information Objects

Georgios Th. Papadopoulos[1,2], Vasileios Mezaris[2], Ioannis Kompatsiaris[2], and Michael G. Strintzis[1,2]

[1] Information Processing Laboratory
Electrical and Computer Engineering Department
Aristotle University of Thessaloniki, Greece
[2] Informatics and Telematics Institute/CERTH
1st Km Thermi-Panorama Road
Thessaloniki, GR-57001 Greece
{papad,bmezaris,ikom}@iti.gr, strintzi@eng.auth.gr

Abstract. In this paper, an ontology-driven approach for the semantic analysis of video is proposed. This approach builds on an ontology infrastructure and in particular a multimedia ontology that is based on the notions of Visual Information Object (VIO) and Multimedia Information Object (MMIO). The latter constitute extensions of the Information Object (IO) design pattern, previously proposed for refining and extending the DOLCE core ontology. This multimedia ontology, along with the more domain-specific parts of the developed knowledge infrastructure, supports the analysis of video material, models the content layer of video, and defines generic as well as domain-specific concepts whose detection is important for the analysis and description of video of the specified domain. The signal-level video processing that is necessary for linking the developed ontology infrastructure with the signal domain includes the combined use of a temporal and a spatial segmentation algorithm, a layered structure of Support Vector Machines (SVMs)-based classifiers and a classifier fusion mechanism. A Genetic Algorithm (GA) is introduced for optimizing the performed information fusion step. These processing methods support the decomposition of visual information, as specified by the multimedia ontology, and the detection of the defined domain-specific concepts that each piece of video signal, treated as a VIO, is related to. Experimental results in the domain of disaster news video demonstrate the efficiency of the proposed approach.

1 Introduction

Over the past decades, access to multimedia content has become the cornerstone of several everyday activities, as well as a key enabling factor at professional level. However, due to the fact that literally vast amounts of multimedia data are generated, stored and distributed from multiple information sources, new needs arise regarding their effective and efficient manipulation. This has triggered intense research efforts towards the development of intelligent systems capable of automatically locating, organizing, accessing and presenting such huge and heterogeneous

B. Falcidieno et al. (Eds.): SAMT 2007, LNCS 4816, pp. 56–69, 2007.

amounts of multimedia information in an intuitive way, while attempting to understand the underlying semantics of the multimedia content [1].

Among the proposed solutions for the problem of semantic analysis of multimedia content, i.e. bridging the so called *semantic gap* between the low-level numerical audio-visual data and the higher-level human perceivable concepts and entities [2], the exploitation of *a priori* knowledge emerges as a very promising one. Approaches belonging to this category require the specification of appropriate knowledge structures for defining a representation of the prior knowledge necessary for analyzing multimedia content and providing support for learning possible links between low-level audiovisual information and semantically meaningful concepts [3].

Regarding the possible domain knowledge representation formalisms, ontologies have been particularly favored due to the significant advantages they present. In particular, they achieve to exhibit a coherent domain knowledge representation model, provide machine-processable semantics definitions and allow automatic analysis and further processing of the extracted semantic descriptions [4]. Concerning the process of semantic video analysis, ontologies have been broadly used in a wide range of approaches. In [5], an ontology framework is proposed for detecting events in video sequences, based on the notion that complex events are constructed from simpler ones by operations such as sequencing, iteration and alternation. A large-scale concept ontology for multimedia (LSCOM) is designed in [6] to simultaneously cover a large semantic space and increase observability in diverse broadcast news video data sets. Additionally, in [7], a pictorially enriched ontology is used both to directly assign multimedia objects to concepts and to extend the initial knowledge for the soccer video domain.

In this paper, an ontology-driven approach for the semantic analysis of video is proposed. The approach builds on an ontology infrastructure and principally a multimedia ontology, whose design is based on the notion of the MMIO. The developed infrastructure is accompanied with signal-level video processing techniques, that are necessary for associating the developed ontology infrastructure with the signal domain. The proposed system supports the decomposition of the visual information and the detection of the defined ontological concepts, thus resulting in a higher-level semantic representation of the video content. Experimental results in the domain of disaster news video demonstrate the efficiency of the proposed approach. The remainder of the paper is organized as follows: Section 2 presents the overall system architecture. Sections 3 and 4 describe the employed high-level knowledge and the low-level information processing, respectively. Sections 5 and 6 detail individual system components. Experimental results are presented in Section 7 and conclusions are drawn in Section 8.

2 System Overview

The first step to the development of the proposed ontology-driven semantic video analysis architecture is the definition of an appropriate knowledge infrastructure that will model the knowledge components that need to be explicitly defined

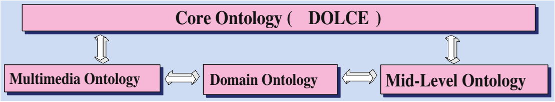

Fig. 1. Knowledge Infrastructure

for the analysis process. For that purpose, ontologies were used, due to the advantageous characteristics that they present and were discussed in the previous section. The developed knowledge architecture, which is depicted in Fig. 1, consists of four individual modules: the Core Ontology (DOLCE), the Mid-Level Ontology, the Domain Ontology and the Multimedia Ontology.

The Core Ontology, which is based on the DOLCE core ontology [8], contains specifications of domain independent concepts and relations based on formal principles derived from philosophy, mathematics, linguistics and psychology. In the proposed framework, it is introduced in order to facilitate the integration and alignment of the individual ontological modules. The Mid-Level Ontology aims to include additional concepts that are generic and not included in the core ontology, thus attempting to ease the alignment of the abstract philosophy of the Core Ontology and the concrete philosophy of the Domain Ontology. Moreover, the Domain Ontology provides a conceptualization of the domain of interest by defining a taxonomy of domain concepts, which are in turn separated into global and local ones. The latter can be used to further characterize parts of a video signal that can be associated with a global one. Furthermore, the Multimedia Ontology, which models the content of multimedia data, serves as an intermediate layer between the Domain Ontology and the audiovisual features, through which the associations of the domain concepts are realized, and includes algorithms for processing the content.

The design of the Multimedia Ontology is based on the notion of the MMIO, and in particular the VIO. The latter constitute extensions and adaptations of the IO design pattern, previously proposed for refining and extending the DOLCE core ontology. Each VIO represents a piece of the video signal to be analyzed, defines its relations and interactions with other VIOs and encompasses the means and methods for its semantic interpretation. The aforementioned ontological modules are suitably aligned and are used to drive the semantic video analysis process. Regarding the particular domain of experimentation, the disaster news video domain was selected and an appropriate Domain ontology was developed.

At the signal level, the video processing procedure, that is necessary for associating the developed ontology infrastructure with the visual domain, is initiated with the application of a temporal segmentation algorithm for segmenting the video sequence into shots and is followed by a keyframe extraction step. More specifically, for every shot a single keyframe is extracted. Subsequently, low-level global frame descriptors are estimated for every keyframe and form a *frame feature vector*. This is utilized for associating the keyframe with the global concepts defined in the domain ontology based on global-level descriptors, serving as

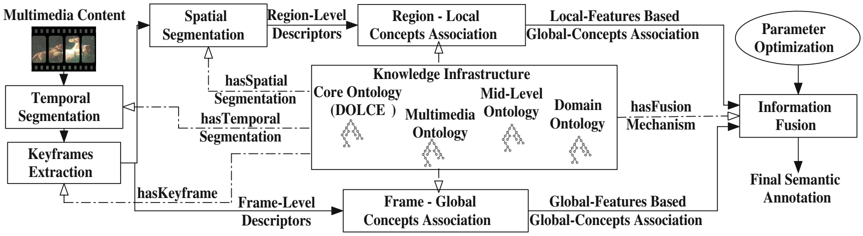

Fig. 2. System Architecture: The dashed arcs denote properties of the developed ontologies that correspond to specific multimedia content processing algorithms

input to a set of SVMs, where each SVM has been trained to detect instances of a particular concept. Every SVM returns a numerical value which denotes the degree of confidence to which the corresponding frame is assigned to the ontology global concept associated with the particular SVM.

In parallel to this process, spatial segmentation is performed for every keyframe and low-level descriptions are estimated for every resulting segment. These are employed for generating hypotheses regarding the region's association to an ontology concept. This is realized by evaluating the low-level *region feature vector* and using a second set of SVMs, where each SVM is trained this time to identify instances of a local concept defined in the domain ontology. SVMs were selected for the aforementioned tasks due to their reported generalization ability [9]. The computed region-level hypothesis sets are subsequently introduced to a *decision function* that is defined in the Multimedia Ontology and which realizes keyframe-global concept association based on local-level information.

Then, a fusion mechanism is introduced, which implements the fusion of the computed keyframe-global concept association based on global- and local-features, in order to make a final keyframe semantic annotation decision. A GA is employed for optimizing the parameters of the fusion mechanism. The choice of a GA for this task is based on its extensive use in a wide variety of global optimization problems [10], where they have been shown to outperform other traditional methods.

Since the final semantic annotation decision is made for every keyframe, it is in turn used to indicate the respective video shot semantic interpretation. Thus, the output of the proposed semantic video analysis framework is a set of shots, to which the input video sequence is decomposed to, and a global concept, defined in the domain ontology, associated with each shot. The overall architecture of the proposed system is illustrated in Fig. 2.

3 Multimedia Ontology

As was described in the previous section, the Multimedia Ontology generally models the content of the multimedia data, serves as an intermediate layer between the Domain Ontology and the audiovisual features, through which the associations of the domain concepts are realized, and includes algorithms for processing the content. Because of its crucial role in the overall semantic video analysis approach, it is described in detail in this section.

Under the proposed approach, the role of the multimedia ontology is to provide the adequate amount of knowledge so that the semantic video analysis procedure is tailored to the specific requirements of a particular application case. More specifically, the multimedia ontology aims to suitably model the content layer of video, define a mapping between low-level audio-visual features or video processing techniques and high-level domain concepts, and generally drive the overall semantic video analysis procedure.

Since the multimedia ontology objective is to guide the semantic video analysis process, its structure should be designed in a way so that both the multimedia properties for specific domain concepts can be described in an arbitrary way and the actual multimedia material is appropriately modeled. For that purpose, the IO design pattern, previously proposed for refining and extending the DOLCE core ontology [11], was adapted and suitably extended. In particular, the DOLCE IO was enriched with two additional properties, namely the 'hasDecomposition' and the 'refersTo' properties, and the resulting information object is denoted with the term MMIO [12]. The MMIO model combines the DOLCE IO pattern with the MPEG-7 standard for the representation of media content and multimedia features [13][14].

Regarding the IO extensions, the 'hasDecomposition' property, the range of which is 'Decomposition', is introduced for describing the decomposition of multimedia objects. Every piece of multimedia information is considered as a multimedia object. 'Decomposition' will provide the MMIO with the needed concepts for the structural description of the multimedia content, in accordance to the respective MPEG-7 description scheme. For that purpose, a variation of the MPEG-7 subpart of the SWIntO [15] ontology, which is in turn an MPEG-7 based ontology for semantic annotation of multimedia content, was adopted. This part of the SWIntO ontology focusses on the MPEG-7 Content Description and Content Management Description Scheme that suffice to model concepts describing storage features (e.g. format, encoding), spatial, temporal and spatio-temporal components (e.g. scene cuts, region segmentation, motion tracking), and low-level features (e.g. color, shape, texture, timbre, melody) of multimedia content. Additionally, the 'refersTo' property is introduced for realizing the connection between MMIOs that are expressed in different modalities but refer to the same content unit. Thus, objects that derive through decomposition of the multimedia content can constitute independent MMIOs with the capability of one referring to another.

The remaining MMIO properties that were inherited from the prototypical DOLCE IO are: the 'about' property, which is used for refering to domain ontology classes or properties; the 'realizedBy' property, which is used as the link to the 'MultimediaFile' class, which in turn describes the physical realization of the MMIO; the 'orderedBy' property, which denotes the format that expresses the MMIO (it can represent a multimedia standards' format, e.g. JPG, UTF-8,

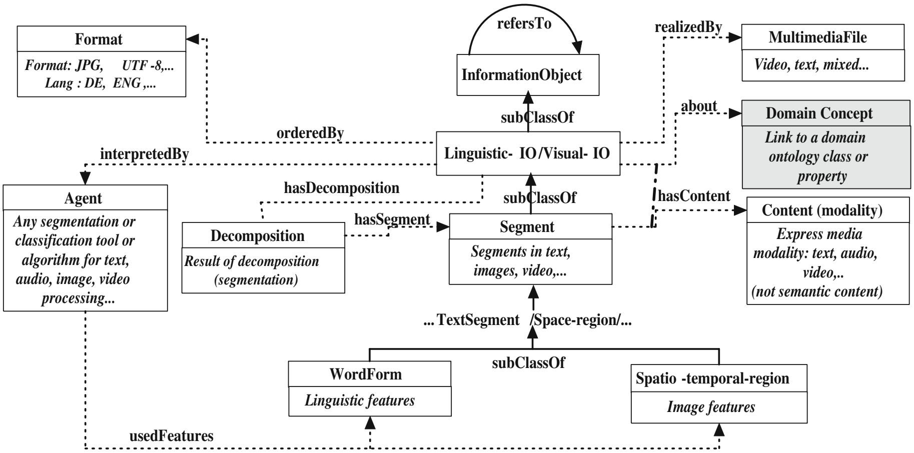

Fig. 3. Multimedia Information Object (MMIO) Design Pattern

etc., or a language, e.g. DE, EN, etc., the latter for the case of textual multimedia information); the ‘interpretedBy’ property, which is used to represent any segmentation or classification tool or multimedia algorithm that can be used for processing the MMIO. A schematic description of a MMIO is depicted in Fig. 3.

The Multimedia Ontology was also enriched with the ‘hasMMAnalysisContextProperty’ property, which was defined in order to provide the appropriate information for reinforcing the semantically driven video analysis process. More specifically, this property associates individual domain ontology concepts in terms of multimedia evidence. Thus, it covers the analysis context of a particular concept in terms of its relationship with other concepts defined in the domain ontology. According to its usage, it covers every modality that multimedia includes (image, text, sound), while it can be further analyzed in a list of properties in order to represent more specific contextual information. For example, the ‘hasMMAnalysisContextProperty’ property can be particularized to the property ‘isLocalConceptOf’ for denoting the relation of concept *debris* to the concept *earthquake* in a possible domain ontology, i.e. for denoting the relation of a local to a global concept defined in the domain ontology, as already mentioned. Another example is the ‘hasFrequencyOfAppearance’ property, which is introduced for indicating the degree of association of a specific local concept to a particular global concept of the domain ontology, i.e. a quantitative interpretation of the aforementioned ‘isLocalConceptOf’ property.

In the developed framework, the MMIO can be sub-divided into three subclasses, namely the Visual Information Object (VIO), the Linguistic Information Object (LIO) and the Audio Information Object (AIO), that each bears all the information about the corresponding distinct modality (Visual, Textual and Audio). In the presented work, only the visual medium is considered, i.e. only the VIO notion is utilized in the semantic video analysis process. A VIO, as being sub-class of the MMIO, uses the same model of object relations that connect it to other concepts and data-type relations, which add to the visual information it carries.

4 Low-Level Visual Information Processing

As already described in Section 2, the video processing procedure, that is necessary for associating the developed ontology infrastructure with the visual domain, is initiated with the application of a temporal segmentation algorithm for segmenting the video sequence into shots and followed by a keyframe extraction step, as denoted by the 'hasDecomposition' property of the Multimedia ontology. The segmentation algorithm proposed in [16] is adopted for that purpose, due to its low computational complexity. Additionally, for every shot a single keyframe is extracted and specifically the median frame is selected.

The association of every extracted keyframe with global concepts of the domain ontology based on global-level information, as will be described in detail in the sequel, requires that appropriate low-level descriptions are extracted at the frame level and form a *frame feature vector*. The frame feature vector employed in this work comprises three different descriptors of the MPEG-7 standard, namely the *Scalable Color*, *Homogeneous Texture* and *Edge Histogram* descriptors. Following their extraction, the frame feature vector is produced by stacking all extracted descriptors in a single vector. This vector constitutes the input to the SVMs structure which realizes the association of every keyframe with global concepts of the domain ontology using global-level information, as described in Section 5.1.

Moreover, in order to perform the association of frame regions with local concepts of the domain ontology, every keyframe has to be spatially segmented into regions, as denoted again by the 'hasDecomposition' property of the Multimedia ontology, and suitable low-level descriptions have to be extracted for every resulting segment. In the current implementation, a modified K-Means-with-connectivity-constraint pixel classification algorithm has been used for segmenting the keyframes [17]. Output of this segmentation algorithm is a segmentation mask S, $S = \{s_i\ ,\ i = 1, ..., \ N\}$, where s_i, $i = 1, ...N$ are the created spatial regions. For every generated frame segment, the following MPEG-7 descriptors are extracted: *Scalable Color*, *Homogeneous Texture*, *Region Shape* and *Edge Histogram*. The above descriptors are then combined to form a single *region feature vector*. This vector constitutes the input to the SVMs structure which computes the hypothesis sets regarding the association of every frame region with the local concepts of the domain ontology (Section 5.2).

5 Keyframe-Concept Association

5.1 Keyframe-Concept Association Using Global Features

In order to perform the association of every extracted keyframe with the global concepts defined in the domain ontology using global level descriptions, a *frame* *feature vector* is initially formed, as described in Section 4. Then, a SVMs structure is utilized to associate each keyframe with the appropriate global concept. This comprises R SVMs, one for every defined global concept C_r^G, each trained under the '*one-against-all'* approach. For the purpose of training the SVMs, a set of keyframes belonging to the domain of interest is assembled, $\mathcal{Q}_{tr}$, as described in Section 7, and is used as training set. The aforementioned frame feature vector constitutes the input to each SVM, which at the evaluation stage returns for each keyframe of unknown global concept association a numerical value in the range $[0, 1]$ denoting the degree of confidence to which the corresponding frame is assigned to the global concept associated with the particular SVM. The metric adopted is defined as follows: For every input feature vector the distance z_r from the corresponding SVM's separating hyperplane is initially calculated. This distance is positive in case of correct classification and negative otherwise. Then, a sigmoid function is employed to compute the respective degree of confidence, h_r^G, as follows:

$$h_r^G = \frac{1}{1 + e^{-t \cdot z_r}} \quad , \tag{1}$$

where the slope parameter t is experimentally set. For each keyframe, the maximum of the R calculated degrees of association indicates its global concept assignment based on global-level information, whereas all degrees of confidence, h_r^G, constitute its respective global concept hypotheses set H^G, where $H^G = \{h_r^G,\ r = 1, ...R\}$.

5.2 Keyframe-Concept Association Using Local Features

As already described in Section 2, the SVMs structure, used in the previous section for global concept assignment using global features, is also utilized to compute the association of every keyframe region with local concepts of the domain ontology. Similarly to the global case, an individual SVM is introduced for every local concept C_j^L, to detect the corresponding association.

For that purpose, a training process similar to the one performed for the global concepts is followed. The differences are that now the region feature vector, as defined in Section 4, is utilized and that each SVM returns a numerical value in the range $[0, 1]$ which in this case denotes the degree of confidence to which the corresponding segment is assigned to the local concept associated with the particular SVM. The respective metric adopted for expressing this degree is defined as follows: Let $h_{ij}^L = I_M(g_{ij})$ denote the degree to which the visual descriptors extracted for segment s_i match the ones of local concept C_j^L, where g_{ij} represents the particular assignment of C_j^L to s_i. Then, $I_M(g_{ij})$ is defined as

$$I_M(g_{ij}) = \frac{1}{1 + e^{-t \cdot z_{ij}}} \quad , \tag{2}$$

where z_{ij} is the distance from the corresponding SVM's separating hyperplane for the input feature vector used for evaluating the g_{ij} assignment. The pairs of all supported local concepts of the domain ontology and their respective degree of confidence h_{ij}^L computed for segment s_i comprise the segment's local concept hypotheses set H_i^L, where $H_i^L = \{h_{ij}^L,\ j = 1, ...J\}$.

After the local concept hypotheses sets, H_i^L, are generated for every keyframe region s_i, a decision function, which is defined in the multimedia ontology, is introduced for realizing the global concept association based on local features, i.e. estimating the global concept assignment for every keyframe on the basis of the local concept hypotheses sets of its constituent regions:

$$d(C_r^G) = \sum_{s_i,\ where\ C_j^L \subset C_r^G} I_M(g_{ij}) \cdot (a_r \cdot freq(C_j^L, C_r^G) + (1 - a_r) \cdot area(s_i)) \quad (3)$$

where $\subset$ denotes the 'isLocalConceptOf' property (already defined in the multimedia ontology), $area(s_i)$ is the percentage of the frame area captured by region s_i and $freq(C_j^L, C_r^G)$ is the frequency of appearance of local concept C_j^L with respect to the global concept C_r^G of the domain ontology. The latter is denoted by the 'hasFrequencyOfAppearance' property of the Multimedia ontology, as already described in Section 3. Regarding the computation of its value, the keyframe set, $\mathcal{Q}_{tr}$, assembled as described in Section 7, is utilized. The reported frequency of appearance of each local concept C_j^L with respect to the global concept C_r^G, $freq(C_j^L, C_r^G)$, is defined as the percentage of the keyframes associated with the global concept C_r^G where the local concept C_j^L appears. The computed values are stored in the multimedia ontology. Parameters a_r are introduced for adjusting the importance of the aforementioned frequencies against the regions' areas for every defined global concept. Their values are estimated according to the procedure described in Section 6.

5.3 Information Fusion for Final Keyframe-Concept Association

After global concept association has been performed using global-, h_r^G, and local-level, $d(C_r^G)$, information, a fusion mechanism is introduced for deciding upon the final global concept association for every keyframe. This has the form of a weighted summation, based on the following equation:

$$D(C_r^G) = \mu_r \cdot d(C_r^G) + (1 - \mu_r) \cdot h_r^G \quad (4)$$

where μ_r, $r = 1, ..., R$ are global-concept-specific normalization parameters, which adjust the magnitude of the global features against the local ones upon the final outcome and their values are estimated according to the procedure described in Section 6. The global concept with the highest $D(C_r^G)$ value constitutes the final global concept association for every keyframe and consequently the semantic annotation of the respective video shot.

6 Optimizing Information Fusion

In Sections 5.2 and 5.3, variables a_r and μ_r are introduced for adjusting the importance of the frequency of appearance against the frame region's area and the global- against the local-level information on the final global concept association decision, respectively. A GA is employed for estimating their values (Section 2).

Initially, the keyframe set $\mathcal{Q}_{tr}$, that was assembled as described in Section 7, is divided into two equal in terms of amount subsets, namely a sub-training $\mathcal{Q}_{tr}^2$ and a validation $\mathcal{Q}_v^2$ set. $\mathcal{Q}_{tr}^2$ is used for training the employed SVMs framework and $\mathcal{Q}_v^2$ for validating the overall system global concept association performance during the optimization process.

Subject to the problem of concern is to compute the values of parameters a_r and μ_r that lead to the highest correct global concept association rate. For that purpose, *Global Concept Association Accuracy*, *GCAA*, is used as a quantitative performance measure and is defined as the fraction of the number of the keyframes that are associated with the correct global concept to the total number of keyframes to be examined.

Under the proposed approach, each chromosome F represents a possible solution, i.e. a candidate set of values for parameters a_r and μ_r. In the current implementation, the number of genes of each chromosome is predefined and set equal to $2 \cdot r \cdot 2 = 4 \cdot r$. The genes represent the decimal coded values of parameters a_r and μ_r assigned to the respective chromosome, according to the following equation:

$$F \equiv [\ f_1\ f_2\ ...f_{4 \cdot r}\ ] = [\mu_1^1\ \mu_1^2 ... \mu_r^1\ \mu_r^2 a_1^1\ a_1^2 ... a_r^1\ a_r^2] \tag{5}$$

where $f_k\ \epsilon\ \{0, 1, ...9\}$ represents the value of gene k and μ_p^q, a_p^q represent the q^{th} decimal digit of variable μ_p, a_p, respectively. The genetic algorithm is provided with an appropriate *fitness function*, which denotes the suitability of each solution. More specifically, the fitness function $W(F)$ is defined as equal to the $GCAA$ metric already defined, $W(F) \equiv GCAA(F)$, where $GCAA(F)$ is calculated over all keyframes that comprise the validation set $\mathcal{Q}_v^2$, after applying the fusion mechanism (Section 5.3) with parameter values for a_r and μ_r denoted by the genes of chromosome F.

Regarding the GA's implementation details, an initial population of 50 randomly generated chromosomes is employed. New generations are iteratively produced until the optimal solution is reached. Each generation results from the current one through the application of the following operators:

- Selection: a pair of chromosomes from the current generation are selected to serve as parents for the next generation. In the proposed framework, the Tournament Selection Operator [10], with replacement, is used.
- Crossover: two selected chromosomes serve as parents for the computation of two new offsprings. Uniform crossover with probability of 0.2 is used.
- Mutation: every gene of the processed offspring chromosome is likely to be mutated with probability of 0.4.

To ensure that chromosomes with high fitness will contribute to the next generation, the overlapping populations approach was adopted. More specifically, assuming a population of m chromosomes, m_s chromosomes are selected according to the employed selection method, and by application of the crossover and mutation operators, m_s new chromosomes are produced. Upon the resulting $m + m_s$ chromosomes, the selection operator is applied once again in order to select the m chromosomes that will comprise the new generation. After experimentation, it was shown that choosing $m_s = 0.4m$ resulted in higher performance and faster convergence. The above iterative procedure continues until the diversity of the current generation is equal to/less than 0.001 or the number of generations exceeds 30. The final outcome of this optimization procedure are the optimal values of parameters a_r and μ_r, used in Eq. 3 and 4, and which are stored in the Multimedia ontology.

7 Experimental Results

In this section, experimental results of the application of the proposed approach to videos belonging to the disaster news domain are presented. The first step to the application of the presented approach for semantic video analysis is the development of the appropriate knowledge infrastructure for representing the knowledge components that need to be explicitly defined, as described in detail in Section 3. Regarding the particular domain of experimentation, an individual domain ontology was developed. This defines the domain concepts of concern, their separation into global and local ones, and the relations among them. The taxonomy of these concepts for the selected domain is depicted in Fig. 4.

Regarding the tasks of SVMs training, parameter optimization and evaluation of the proposed system performance, a number of keyframe sets needs to be formed. More specifically, a set of 400 keyframes, $\mathcal{Q}$, that were extracted from respective disaster news videos according to the procedure described in Section 4 and include global concepts of the developed domain ontology, was assembled. Each keyframe was manually annotated (i.e. assigned to a global concept and, after segmentation is applied, each of the resulting frame regions associated with a local concept of the domain ontology). This set was divided into two equal in terms of amount sub-sets, $\mathcal{Q}_{tr}$ and $\mathcal{Q}_{te}$. $\mathcal{Q}_{tr}$ was used for training purposes, while $\mathcal{Q}_{te}$ served as a test set for the evaluation of the proposed system performance.

According to the SVMs training process (Section 5), a polynomial function was used as a kernel function by each SVM for both global and local concept association cases. The respective low-level *frame* and *region feature vector* are composed of 405 and 445 values respectively, normalized in the interval $[-1, 1]$.

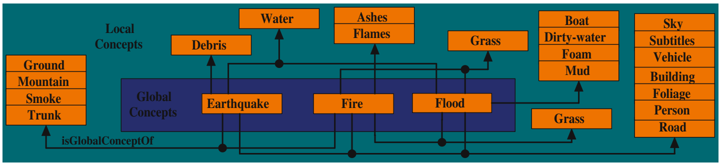

Fig. 4. Taxonomy of Domain Concepts: The arcs denote the 'isGlobalConceptOf' property, which connects the domain concepts in terms of multimedia evidence

Extracted Keyframe			
Keyframe-Concept Association Using Global-Level Information	Earthquake :**0.84** Fire :0.22 Flood :0.43	Earthquake :0.21 Fire :**0.87** Flood :0.11	Earthquake :**0.54** Fire :0.21 Flood :0.31
Keyframe-Concept Association Using Local-Level Information	Earthquake :0.51 Fire :0.18 Flood :**0.52**	Earthquake :0.24 Fire :**0.69** Flood :0.31	Earthquake :0.22 Fire :0.14 Flood :**0.62**
Keyframe-Concept Association Using Information Fusion	**Earthquake**	**Fire**	**Flood**

Fig. 5. Indicative Keyframe-Concept Association Results

The disaster news videos to be analyzed, were initially temporally segmented and corresponding keyframes were extracted, following the procedure described in Section 4. Then, based on the trained SVMs structure, keyframe-concept association based on global level features is performed, as described in Section 5.1. In parallel, after spatial segmentation is applied to the extracted keyframes, local concept hypotheses are generated for each frame segment and a decision function realizes keyframe-concept association based on local features (Section 5.2). Afterwards, the approach described in Section 5.3 is employed for implementing the fusion of the global and the local features based keyframe-concept association information and computing the final keyframe-concept assignment for every keyframe, which in turn constitutes the semantic interpretation of the respective video shot. The values of the fusion mechanism parameters are estimated according to a GA-based optimizer (Section 6).

In Fig. 5 indicative keyframe-concept association results are presented, showing the extracted keyframe, the keyframe-concept association using only global (row 2) and only local (row 3) information and the final keyframe-concept assignment after the implementation of the fusion mechanism. Additionally, in Fig. 6 exemplary region-concept association results are presented, showing the extracted keyframe (row 1) and, after spatial segmentation is applied, the association of the local concepts of the domain ontology (row 2). Furthermore, in Table 1, the respective quantitative performance measures for every individual algorithm are given in terms of accuracy for each global concept and overall. Accuracy is defined as the percentage of the keyframes that are associated with the correct global concept. From the results presented in Table 1, it can be verified

Table 1. Keyframe-Concept Association Accuracy

	Accuracy			
Method	Earthquake	Fire	Flood	Overall
Keyframe-Concept Association Using Global-Level Information	**93.75%**	**98.08%**	72.13%	86.96%
Keyframe-Concept Association Using Local-Level Information	83.33%	75.00%	59.02%	71.43%
Keyframe-Concept Association Using Information Fusion	**93.75%**	94.23%	**91.80%**	**93.17%**

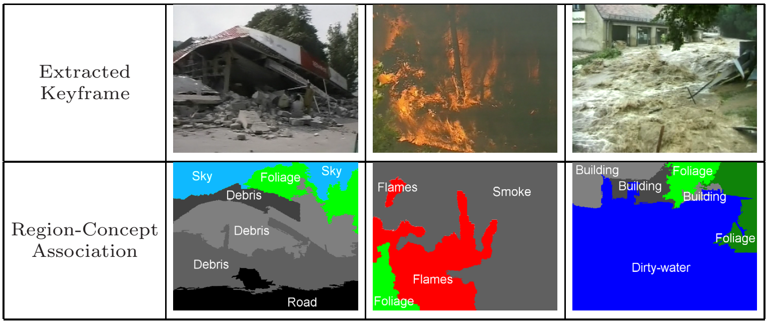

Fig. 6. Indicative Region-Concept Association Results

that the keyframe-concept association based only on global information generally outperforms the respective association based only on local information. Furthermore, it must be noted that the proposed global and local features information fusion approach leads to a significant performance improvement, compared to the keyframe-concept association based solely on global or local features.

8 Conclusions

In this paper, an ontology-driven approach to semantic video analysis that is based on the notion of the Visual Information Object was presented. The proposed framework can easily be extended or applied to additional domains, provided that the employed knowledge infrastructure is appropriately modified and that the utilized training set is enriched with suitable training samples. Future plans include the introduction of audio signal processing tools and text analysis algorithms, so that the entire capabilities of the developed framework can be fully exploited and multi-modal semantic multimedia analysis is realized.

Acknowledgements

The work presented in this paper was supported by the European Commission under contracts FP6-001765 aceMedia, FP6-027685 MESH and FP6-027026 K-Space, and by the GSRT under project DELTIO.

References

1. Al-Khatib, W., Day, Y.F., Ghafoor, A., Berra, P.B.: Semantic modeling and knowledge representation in multimedia databases. IEEE Trans. on Knowledge and Data Engineering 11, 64–80 (1999)
2. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)

3. Zlatoff, N., Tellez, B., Baskurt, A.: Image understanding and scene models: a generic framework integrating domain knowledge and Gestalt theory. In: ICIP 2004. Int. Conf. on Image Processing, vol. 4, pp. 2355–2358 (2004)
4. Staab, S., Studer, R.: Handbook on ontologies. In: Int. Handbooks on Information Systems, Springer, Berlin (2004)
5. Francois, A.R.J., et al.: VERL: an ontology framework for representing and annotating video events. IEEE Multimedia 12(4), 76–86 (2005)
6. Naphade, M., et al.: Large-scale concept ontology for multimedia. IEEE Multimedia 13(3), 86–91 (2006)
7. Bertini, M., Cucchiara, R., del Bimbo, A., Torniai, C.: Video Annotation with Pictorially Enriched Ontologies. In: Proc. of ICME, pp. 1428–1431 (July 2005)
8. Laboratory for Applied Ontology, `http://www.loa-cnr.it/DOLCE.html`
9. Kim, K.I., Jung, K., Park, S.H., Kim, H.J.: Support vector machines for texture classification. IEEE Trans. on Pattern Analysis and Machine Intelligence 24, 1542–1550 (2002)
10. Mitchell, M.: An introduction to genetic algorithms. MIT Press, Cambridge (1995)
11. Gangemi, A., et al.: Sweetening ontologies with DOLCE. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, Springer, Heidelberg (2002)
12. SmartWeb project: `http://www.smartweb-projekt.de/`
13. Buitelaar, P., Sintek, M., Kiesel, M.: A Lexicon Model for Multilingual/Multimedia Ontologies. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, Springer, Heidelberg (2006)
14. Romanelli, M., Buitelaar, P., Sintek, M.: Modeling Linguistic Facets of Multimedia Content for Semantic Annotation. In: SAMT 2007. LNCS, vol. 4816, pp. 240–251. Springer, Heidelberg (2007)
15. SWIntO: SmartWeb Integrated Ontology, `http://smartweb.dfki.de/ontology_en.html`
16. Kobla, V., Doermann, D., Lin, K.: Archiving, indexing, and retrieval of video in the compressed domain. In: Proc. SPIE Conf. on Multimedia Storage and Archiving Systems, vol. 2916, pp. 78–89 (1996)
17. Mezaris, V., Kompatsiaris, I., Strintzis, M.G.: Still Image Segmentation Tools for Object-based Multimedia Applications. Int. Journal of Pattern Recognition and Artificial Intelligence 18(4), 701–725 (2004)