RNPS No. 2142
ISSN 2072-6287
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

State-of-the-art Approaches for Automatic Video Annotation and Retrieval Using High Level Features

Annette Morales-González and Edel García-Reyes

RT_061 **abril 2014**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

REPORTE TÉCNICO
Reconocimiento de Patrones

State-of-the-art Approaches for Automatic Video Annotation and Retrieval Using High Level Features

Annette Morales-González and Edel García-Reyes

RT_061 **abril 2014**

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Table of Contents

List of Figures

List of Tables

State-of-the-art Approaches for Automatic Video Annotation and Retrieval Using High Level Features

Annette Morales-González and Edel García-Reyes

Pattern Recognition Research Team, Advanced Technologies Application Center (CENATAV),
Havana, Cuba
{amorales,egarcia}@cenatav.co.cu

RT_061, Serie Azul, CENATAV
Aceptado: 17 de marzo de 2014

Abstract. The booming of mobile capture devices like digital cameras and cell phones, the development of sophisticated video-surveillance systems and the increasing growth of social networks and online storage systems for gathering multimedia archives, have stimulated the Computer Vision community to make a bigger emphasis in achieving higher accuracy in video content analysis research. In this technical report, a group of video annotation and retrieval approaches are surveyed. These approaches bear in common the use of high-level features as a way to complement and to add semantics to low-level based methods for concept detection in videos. We also revisit works that use ontologies to represent knowledge for semantic video annotation tasks and especial attention is given to the development of ontologies for video-surveillance domain. At the end of the report, the surveyed methods are summarized, highlighting their main characteristics and comparing their results and the way they were obtained. After that, we discuss the main deficiencies that derive so far in this research area.

Keywords: video content analysis, content-based video retrieval, video concept detection, semantic video annotation.

Resumen. El auge de los dispositivos de captura móviles, como cámaras digitales y celulares, el desarrollo de sofisticados sistemas de videovigilancia y la creciente expansión de las redes sociales y los sistemas de almacenamiento online de archivos multimedia, han propiciado que la comunidad de Visión por Computadora haga un mayor énfasis en lograr mayor eficacia en los métodos relacionados con el análisis del contenido de los videos. En este reporte se relacionan una serie de trabajos en el área de anotación y recuperación de videos utilizando rasgos de alto nivel, como forma de complementar y añadir semántica a los métodos basados en rasgos de bajo nivel. También se estudian trabajos en los cuales se utilizan ontologías como formas de representación del conocimiento para la anotación semántica de videos. Especial atención se hace en el desarrollo de ontologías para el dominio de video vigilancia. Al final del reporte se resumen los métodos analizados, resaltando su principales características y comparando sus resultados y la forma de obtenerlos, para luego discutir las principales deficiencias que se derivan hasta el momento en esta área de investigación.

Palabras clave: análisis del contenido de video, recuperación de video por contenido, detección de conceptos en video, anotación semántica de videos.

1. Introduction

The increasing growth in video content generation by people worldwide (boosted by the ease of use of digital cameras and the availability of online storage services like Youtube[1] and Vimeo[2], as well as the

[1] http://www.youtube.com
[2] https://vimeo.com

deployment of surveillance cameras in public locations, etc.), have stimulated the research and development of video search engines. Most commercial video search engines provide access to video based on text, since this is the easiest way for a user to describe what he wants to find. The indices are often based on filenames, social tagging, closed captions, or speech transcripts. It is almost evident to notice that the retrieval performance will be low when the visual content is not mentioned, or properly reflected in the associated text. Also, if the textual information provided with the videos is not in English language (ex. contents generated in China, Russia, etc.), querying the content becomes even harder [6].

The problem of automatic identification of video content has led to a new Computer Sciences field known as Video Content Analysis. Currently, this is split into two main parts: (1) low-level features extraction and (2) high-level feature extraction. The low-level feature extraction line deals with raw video properties such as color and resolution, as well as detection of shot[3] delimiters. The high-level feature extraction line aims to mine and describe concepts, events, scenes and objects present in the video. An important issue to solve towards this direction is related to the way of assigning low-level descriptors to high-level concepts. The current inability to accurately connect low-level features extracted from raw image and video data, with high concepts present in human minds, is known as the semantic gap.

The objective of performing content-based video analysis and description is to improve the accuracy of nowadays systems for searching and retrieving videos. Video indexing is the process of assigning links, labels or access points to a video, based on its content. Unfortunately, due to the semantic gap problem, concept-based video indexing remains a critical obstacle to the success of the query-by-concept search approach. Generally, semantic concepts are still difficult to detect accurately, so their detection in video remains a challenging problem. According to [7], the accuracy of state-of-the-art concept detection in videos can range from less than 0.1 (measured by average precision) for semantic concepts such as “people marching” or “fire weapon” to above 0.6 for a concept such as “face”.

Exploiting the semantic relationships between concepts has received a large amount of attention from the scientific community since they can improve the detection of concepts and obtain a richer semantic annotation of a video. Also, temporal information can provide important cues for disambiguating the detected concepts. To this end, ontologies are expected to improve the capability of computer systems to automatically detect even complex concepts and events from visual data with higher reliability.

In recent years, there have been an increasing interest towards automatic video annotation and retrieval in large repositories. In this sense, several research projects have emerged proposing algorithms, tools and benchmarks to develop the field. The most prominent among them are the following:

1. ViPER-GT and ViPER-PE [8][9]. They constitute an interoperable platform to select concepts and events manually in a video. They generate test data and annotate video files in XML format.
2. I-Lids (Imagery Library for Intelligent Detection Systems) [10]. This is an iniciative from the United Kingdom to facilitate the development and benchmarking of vision-based detection systems that comply the government requirements.
3. ETISEO (Evaluation du Traitment et de l’Interpretation de Sequences Video) [11]. This is a platform for benchmarking video processing algorithms regarding a specific task (for instance, object identification, classification, tracking and event recognition), a specific scene (ex. a road), and a global difficulty level (ex. shadow contrast).
4. VACE (Video Analysis and Content Extraction) [12]. This is a project for the development of algorithms for automatic object and event recognition in a scene, in a robust and scalable way. The events, objects and relationships among them define the key components of video content.

[3] A shot is a continuous frame subsequence that shows a story.

5. VISOR (Video Surveillance Online Repository) [13]. This is a project designed for creating an open platform to collect, annotate, retrieve and share surveillance videos. They also evaluate the accuracy of automatic video surveillance. They propose an open ontology structured as a simple list of video surveillance concepts.

Since 2003, the National Institute for Standardization and Technologie (NIST) from the United States of America has been organizing the TRECVID contest [14][15] related to video retrieval. In this competition, several video repositories have been employed, depicting different domains such as news, video surveillance, etc.

The main purpose of this technical report is to provide insight of the semantic video analysis area, as well as to describe the more relevant state-of-the-art methods that use high-level information to improve content-based video annotation and retrieval accuracies. The remainder of this paper is as follow. In Section 2 we enclose all topics related to semantic video search principles. This section is divided in four subsection: subsection 2.1 will focus on methods to assign concepts to low-level data, subsection 2.2 will describe several approaches for introducing multi-concept and temporal relations, subsection 2.3 will expose how ontologies have been used for adding high-level knowledge to video content annotation, and subsection 2.4 will show some approaches especially developed for video surveillance purposes. Section 3 presents a summary of the analyzed methods, comparing their results and pointing the main deficiencies in this research area. At the end of the report, conclusions are provided.

2. Semantic Search in Video

The semantic search of videos aims to improve the accuracy of the search task, trying to understand the user intentions and the contextual meaning of the terms, in order to obtain highly relevant results. Adding semantics to video representations have been done in different ways, namely, adding textual descriptions (captions usually provided by users), extracting text from the video using optical character recognition (OCR) or automatic speech recognition (ASR), automatically annotating videos using machine learning techniques (concept detection), enhancing existing annotations using rule learning and relations to infer high-level concepts and finally, employing ontologies to represent the knowledge in the video domain. In this section we aim to revisit approaches from concept detection to ontology development, in order to provide a panoramic view of the main state-of-the-art trends in this field.

2.1. Concept Detection

The problem of concept detection in video can be stated as a pattern classification problem, where several classifiers based on visual, auditive and/or textual features are trained using information coming from the raw video data and a set of annotations. In other words, the task focuses in learning a correspondence among low-level features extracted from videos and high level semantic annotations [16][17].

In [17] they proposed a set of 374 semantic concept detectors, called “Columbia374”. The 374 concepts were selected from the LSCOM ontology [18] (See section 2.3). They construct three Support Vector Machines (SVMs) based on local edge (i.e. SIFT), color moment and wavelet texture features. Finally, for each shot, the recognition score of the object is computed as the average of outputs of the above SVMs.

In the literature, the problem of concept detection is addressed from the division of the video in shots (for news videos, documentaries, etc, that differ from surveillance videos) and the detection of keyframes from each shot. In order to explore the spatial content of keyframes, low-level feature extraction is per-

formed after the image is modelled as a grid. In this way, colors and textures are analyzed locally. In other proposals, low-level features are extracted from image regions obtained after the unsupervised segmentation of the image. A more elaborated process focuses in the obtention of a thesaurus of regions. For this, low-level features extracted from a significative large number of keyframes are clustered and the centroids are considered visual words. After this, each keyframe is represented by a vector containing the distances of each region to each visual word (for each visual word, the smallest distance among all the regions of the image and the visual word is stored). Finally, a support vector machine (SVM) is trained with the resulting vectors to detect each of the high-level concepts [19]. In the MediaMill system [6], they proposed a similar approach, with the difference that they defend the idea that using a multiframe sampling strategy is better than selecting keyframes from a video. They state that taking more frames into account during analysis, makes possible to recognize concepts that are visible during the shot, but not necessarily in a single key frame. These approaches are extrapolated from the field of image analysis into the video analysis field, which is straightforward by the main definition of a video as a sequence of images. Nevertheless, they apply their method to each image in the sequence, as if they were independent from each other, disregarding the temporal information implicitly present in the sequence. Also, this approaches do not take into account the contextual dependence among concepts.

Another type of promising approach involves refining the scores of concept-specific detectors for better final indexing accuracy, often by exploring contextual correlation and temporal coherence. An example of contextual correlation is the co-occurrence between semantic concepts in a shot; temporal coherence relates to a single concept that occurs in multiple neighboring shots.

2.2. Multi-concept and Temporal Relations

Previous experiences have shown that semantic concepts are not independent from each other. Therefore, the use of relations among multiple semantic concepts in video may be an effective approach to improve the accuracy in concept detection, since this provides important contextual information. According to [17], when context-based reranking is applied after single concept detection, the average precision in search tasks can improve from 15 to 25 %.

In [20], they state that current re-indexing methods that exploit contextual and temporal relations to refine the initial scores can be classified into three categories, according to the extra knowledge involved:

1. Self-refining methods (unsupervised learning). They use only initial scores to explore informative cues to refine indexing performance [21].
2. Example-based refining methods (supervised learning). They discover relations from user information (examples and annotations) to improve the initial results [22][1][23].
3. Crowd-refining methods. They use external knowledge (like WordNet or Wikipedia), heterogeneous resources (like social media), or search engines (like Google and Yahoo!) for better performance [24].

As discussed in [21], example-based refining methods display much better results compared to self-refining methods, nevertheless, supervised example based methods heavily rely on manual annotations in order to collect reliable training data. Crowd-refining methods (as well as self-refining methods) do not require expensive user information, but they present cross-domain problems, when the data distribution of the external sources do not match those of the target domain.

In [21], the authors propose a self-refining method based on the idea of collaborative filtering, taking into account the concept-to-concept correlation and shot-to-shot (temporal) similarity embedded within the score matrix. Collaborative filtering utilizes user-user similarity and item-item correlation to predict

missing preferences. The authors extrapolate this idea to video analysis, by treating video shots as users and concepts in the lexicon as items. Although in this work they defend the advantages of unsupervised learning approaches for this task, in order to avoid using user annotations, in fact, their own approach relies on supervised concept detectors to obtain the initial score matrix. Therefore, it is unclear the contribution towards this matter.

Even though manual user annotations are hard to obtain, most of the state-of-the-art methods follow the example-based refining approach, since human information is usually more reliable when it comes to gaining semantic insight of a problem or domain.

The co-occurrence of several semantic concepts may indicate the presence of other concepts. For example, the presence of the concept “building” might implicate the presence of the concept “exterior scene”. In this way, we could discover associations among concepts from existing annotations, and use them to improve the accuracy in concept detection. This kind of association can be present at a semantic concept level or at a visual concept level. Also, the discovery of temporal dependencies among concepts may be helpful to infer the occurrence of semantic concepts. See an example in Figure 1.

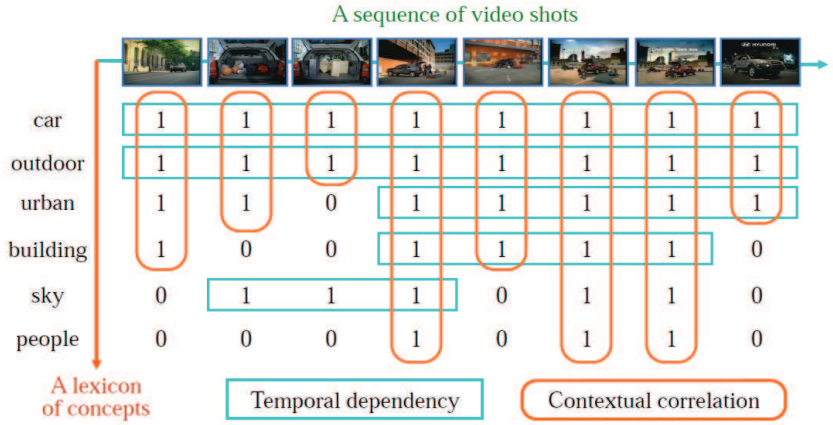

Fig. 1. An example of an annotation system that takes into account multi-concept and temporal relations among shots. In the annotation matrix, 1 indicates the presence of a concept in the shot, and 0 the opposite. This image was taken from [1].

In example-based refining approaches, during training stage, contextual and temporary cues for each concept must be extracted as high order relations, from videos manually labeled. In test stage, the contextual and temporal relations discovered are combined with the prediction values obtained by the concept classifiers. Thus, the classification results are refined, exploiting not only the detection scores, but also the contextual and temporal relations.

On the other hand, it is possible to detect compound concepts patterns, that are defined from temporary and semantic relations among concepts of an ontology. These are assumed to be characteristic of the application domain, and can be included in the ontology to facilitate the annotation of long video sequences and to express complex queries.

One approach to combine these types of information, is the creation of a non-directed graph model to represent relations among concepts and subsequences. This model is called Multiple Discriminative Random Field (MDRF) [2]. In Figure 2, the relation between pairs of concepts can be seen as edges of a graph, representing the interaction potential in the MDRF model. Dotted lines, depicting the presence of a concept in the shot, stand as the association potential in the model.

In other works, association rules for concepts and algorithms to detect frequent items (such as *Apriori* algorithm) have been used [22]. From training data, they discover concept association rules that capture

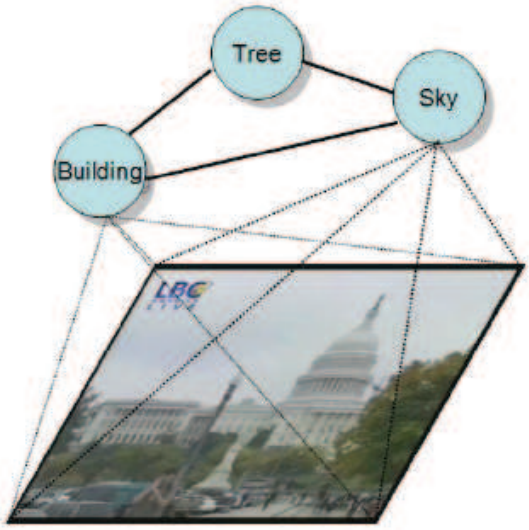

Fig. 2. Example of multi-concept relations in a shot, represented as an undirected graph. The edges between concepts represent relations among them. The dotted lines from concepts to the video shot represent the classification of each concept in the shot. This image was taken from [2].

inter-concept relationships among multiple concepts. After a concept detection stage, where only prediction values for classifying shots are obtained, the rule-based post-filtering module uses the learned association and temporal rules to re-rank the test shots. In [1][23] they model multi-concept and temporal relations as depicted in Figure 1. They compute inter-concept correlation values and also correlation values between neighboring shots (temporal information). They use a graphical model approach, where they employ the information shown in Figure 1 as observations and the correlation values previously computed, in order to improve the concept detection in every shot. In [25], they propose a multiple hypergraph approach in order to combine different types of information. They build three hypergraphs, one for visual features, another for textual features (coming from automatic speech recognition transcript as source text in video) and the last one for multi-concept relations. Temporal information in this approach is taken into account through the hyperedges, which are formed by all the shots containing the same visual, texture or concept feature in each hypergraph. In the work presented by [26], they propose a concept fusion algorithm called Temporal-Spatial Node Balance algorithm (TSNB), using a representation very similar to the one presented in [1]. They call "spatial relations" to the co-ocurrence of concepts in a shot space (what is called "Contextual correlation" in Figure 1), instead of defining spatial relations among concepts in image space. Temporal relations are modeled as in [1]. Instead of having binary values for defining the presence or absence of concepts in a shot, they employ the detection scores obtained by a classifier. They employ a graphical model to perform the concept fusion, defining potential functions for spatial and temporal relations among shots.

In the aforementioned approaches ([22][1][23][25][2]), they consider a shot to be the basic unit for classification and annotation, therefore, at concept detection stage, classification is performed on a whole shot (as a sequence of frames), disregarding local analysis of concept relations on each frame, and the temporal relationship of neighboring frames within a single shot. This is illustrated in Figure 2, where concepts are associated to the entire shot, instead of taking advantage of their spatial distribution within image space.

2.3. Multimedia Ontologies

Ontologies are commonly used for knowledge representation of different domains. Ontologies consist of concepts, concept properties, and relationships between concepts. They organize semantic heterogeneity of information, using a formal representation, and provide a common vocabulary that encodes semantics and supports reasoning. Research activities for video domain ontologies have focus on ontology definition methods, ontology standards and languages and methodologies to connect knowledge extracted from data to the concepts of the ontology. Several multimedia ontologies have been proposed recently as suitable knowledge models to narrow the semantic gap and to enable the semantic interpretation of images.

Traditional ontologies are based on linguistic concepts. Nevertheless, several authors [4][27] sustain that the idea behind multimedia ontologies presupposes that traditional linguistic ontologies are not able to describe the diversity of visual events and elements present in a video. Also, they cannot support annotation and retrieval at the level of a specific pattern of events or entities (like those that are represented in a video). Multimedia ontologies are expressed in the OWL standard. The linguistic part is formed by a set of classes, expressing the domain main concepts (ex. objects, actions, etc.) and their relationships.

Multimedia ontologies research activities are split into two major groups: those who create, adapt or expand existing domain ontologies and ontology languages, in order to match the requirements underlying the semantic representation of media objects; and those that develop new methods to link low-level multimedia information with high-level concepts represented in the ontologies.

As an example of the first case, the work presented by [28], attempts to expand the OWL ontology language in order to support temporal and spatial relations while modeling multimedia data using a multimedia ontology. Their approach focuses more on the tools to adapt the OWL language than on methods to properly associate low-level features to high-level concepts, and to extract those relations from row video data. Another example is the one presented by [29], where the main goal is to analyse the requirements of multimedia object representations, which, according to them, are not fulfilled by most semantic multimedia ontologies. They present a new Core Ontology Multimedia, named COMM, following the principles they exposed in their work. COMM does not represent high-level entities of the scene, such as people or events. Instead, it identifies the components of a MPEG-7 video sequence in order to link them to (Semantic) Web resources. Another popular approach towards knowledge representation in this area is the LSCOM ontology [18], which includes more than 834 visual concepts jointly defined by researchers, information analysts, and ontology specialists according to the criteria of usefulness, feasibility, and observability. These concepts are related to events, objects, locations, people, and programs that can be found in general broadcast news videos. One crucial problem of LSCOM is that it just provides a list of concepts. That is, to utilize LSCOM in video retrieval, it would be necessary to organize LSCOM concepts into a meaningful structure. An attempt to do this was performed by [3], where they explain how to organize 374 of the LSCOM concepts into a video ontology and how to select concepts related to a query. Those 374 concepts were the ones used by [17] to create the “Columbia374” set of concept detectors. Since the objective of [3] was to construct a video ontology which utilizes object recognition scores, they did not describe their computation, and instead, they employed the scores already obtained by [17]. An example of the video ontology can be seen in Figure 3.

Since the main focus of this technical report relies on analysing existing approaches for connecting video data to perceptual concepts, we prefer to make a deeper analysis of works following the second branch of research activities on the multimedia ontology field: those that develop new methods to link low-level multimedia information with high-level concepts represented in the ontologies. Although several approaches have been presented, practical implementations of (visual) data fusion systems based on formal knowledge representations still scarce.

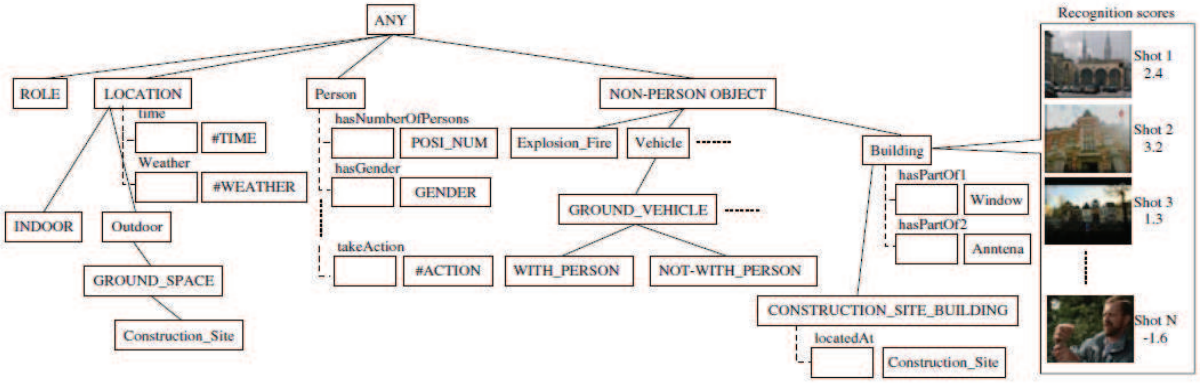

Fig. 3. Example of the video ontology proposed in [3]. Concepts belonging to LSCOM appear in lower-case letters and concepts introduced by authors appear in capital letters.

The multimedia ontology developed in [4] and [27] includes conceptual and perceptual items. It is created by linking video sequences as instances of the linguistic ontology concepts and performing an unsupervised clustering of video instances. The visual features used for clustering are both generic attributes (ex. trajectories, motion fields, color and edge histograms extracted from image raw data, etc.) and domain-specific descriptors (ex. spatio-temporal feature combinations) that represent especial events. The group centroids are considered visual concepts, each one representing an specific pattern that describe an action or an event. An especial class of un-detected events is also created, containing all video sequences that are not classified as instances of a concept within a predefined confidence. Video annotation is performed at two different levels. At video subsequence level, the subsequences are annotated by checking the similarity with the visual concepts in the multimedia ontology. When the similarity with a particular visual concept is confirmed, the high-level concept linked to it in the ontology is immediately associated with the subsequence. New annotated subsequences are associated with the existing groups and the centroids are updated, therefore, subsequences linked to the group of undetected events are re-analyzed for a possible association with the new groups. According to this mechanism, the ontologies have a static linguistic part and a dynamic visual part (visual concepts change when new information is presented to the system). At sequence level, the compound concept patterns are annotated. The system must allow to check whether a video sequence contains a sequence of subsequences, thus verifying the compound concept patterns predefined. An illustration of this kind of ontology can be seen in Figure 4.

A knowledge infrastructure to describe video content has been proposed in [30]. They connect four types of ontologies: a Core Ontology (base on the Dolce foundation ontology), which contains generic concepts derived from philosophy, mathematics, linguistics and psychology; a Mid-Level Ontology, to include additional concepts that are generic and not included in the core ontology; a Domain Ontology which provides a conceptualization of the domain of interest by defining a taxonomy of domain concepts; and finally, a Multimedia Ontology, which models the content of multimedia data and serves as an intermediate layer between the Domain Ontology and the audiovisual features. The association between low-level features and the ontology concepts is performed after the video is segmented in shots and a single keyframe is extracted for each one of them. They use global features to classify a frame with global concepts in the ontology, and they segment each keyframe to extract local features that will be mapped to local concepts of the ontology.

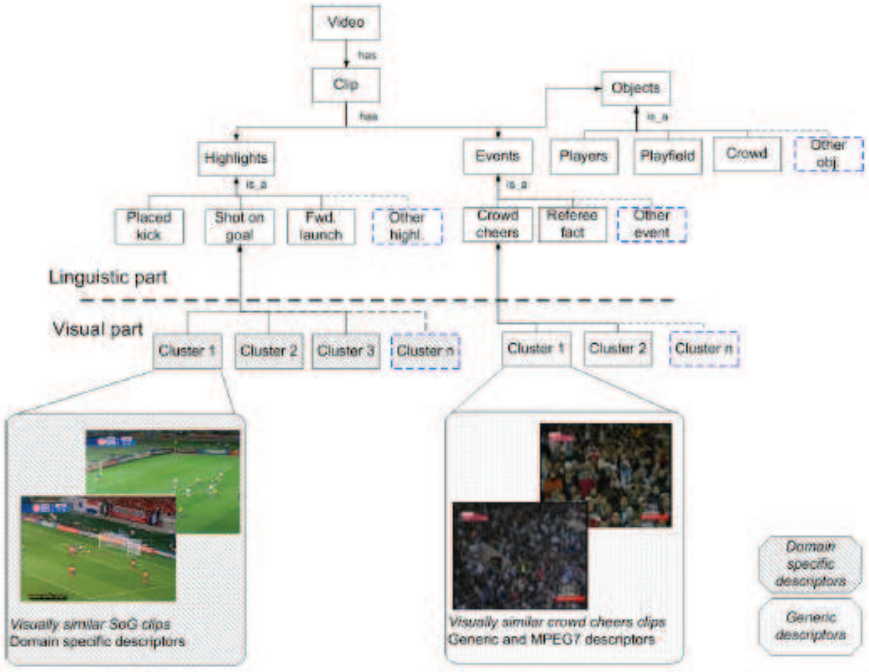

Fig. 4. Example multimedia ontology of soccer domain using clusters of visual instances. This image was taken from [4].

In order to bridge the semantic gap between low-level features and concepts in an ontology, in [31], they define a Perception Concept and a Semantic Concept. The perception concept is the abstraction of feature patterns that have similar low-level features and occur frequently. The semantic concept is related to high-level concepts that users perceive when they watch a video. It can be represented as a combination of several perception concepts and their relations, and can be described by a linguistic term. The detection of semantic concepts depends on the detected perception concepts and contextual information (in the form of textual information extracted by using VOCR and automatic speech recognition).

In [30], although explicit relations among concepts exist in the developed ontology framework, they disregard this information in the low-level feature classification step. Both [30] and [31] dismiss temporal information, as well as the spatial relations of concepts in image space.

The work presented in [7] proposes the automatic determination of semantic linguistic relations between concepts. Concept detectors are linked to the corresponding concepts in an ontology and they propose a rule-based method for automatic semantic annotation of composite concepts and events (such as “a person enters in a secured area”) in videos. The concept relationship of co-occurrence and the temporal consistency of video data are used to improve the performance of individual concept detectors.

Object detection in video images captured in vehicular traffic situations is performed in [32]. Detected objects (by using stereo vision the detection outputs cuboidal objects) are mapped to one of four concepts (Automobile, UtilityPole, Person or Obstacle for the unclassified objects) present in the OpenCyc ontology [33]. After low-level features are extracted, object recognition is performed by using a cascade of classifiers. Spatio-temporal rules are used for improving object classification.

In order to exploit context knowledge, in [34] they propose a context-based layer that will manage the ontological representation of a scene (including context and perceived knowledge), and below this layer, a general tracking layer, that manages a classical object tracking procedure. The current state of a

given scene is represented with instances of the concepts modeled in the context layer and its relations. An interface between the layers guarantees interoperability and independence between them. Updated track information triggers reasoning processes in the context layer that, supported by context knowledge, update the whole interpretation of the scene. Recommendations are then created by reasoning with the current scene model and a priori knowledge in order to improve the tracking process. The use of a tracking system implicitly provides temporal information in the model. The context layer model encompasses various ontologies that represent tracking data, scene objects, activities, impacts, and tracking layer recommended actions.

The work presented in [35] is one of the most complete approaches when it comes to involve high-level relations to low-level representations. They propose an automatic semantic content extraction framework in terms of object, event, concept, spatial and temporal relation extraction. They developed a metaontology (named VISCOM) for domain ontologies that provides a domain-independent rule construction standard. VISCOM is utilized as a metamodel to construct domain ontologies, and domain specific semantic contents are defined as individuals of VISCOM classes and properties. Both the ontology model and the semantic content extraction process are developed considering uncertainty issues. Spatial relations (currently distance, topological, and positional relations are used) are fuzzy relations and, for each relation type, membership values can be calculated according to the positions of objects relative to each other. In the concept extraction process, extracted object, event, and concept instances are used. When an object or event that is used in the definition of a concept is extracted, the related concept instance is automatically extracted with the relevance degree given in its definition. Concepts are extracted with a membership value between 0 and 1, which represents the possibility of the concept realization in the extracted concept period and the roles of objects taking part in the concept.

2.4. Video Surveillance Domain Ontologies

For the specific case of video surveillance, it is important to notice that, in order to label and to retrieve frame sequences, specific approaches must be developed. This is because the structural composition of these frames is different from edited materials (namely, news videos, documentaries and films).

For the case of video surveillance, ontologies have been used to assist the recognition of video events. Several authors have engaged initiatives to standardize taxonomies of video events. Such is the case of [36], which proposes a formal language to describe event ontologies (VERL, which defines the concepts to describe processes) and a markup language (VEML) to annotate instances of ontological events. In [37] they defined a meeting ontology that is determined by the knowledge base of various meeting sequences and an ontology for describing bank attack scenarios was proposed by [38]. More general ontology design principles were developed by [39] and the authors adapted them to the specific domains of human activity, bank and airport surveillance. In [40] they developed a verb ontology to enhance the description of relations between events. This ontology is used to classify events that may help the comprehension of other events (for instance, when an event is a precondition of another one). More recently, in [41] they defined a formal model of events that allows interchange of event information between different event-based systems, causal relationships between events, and interpretations of the same event by different humans. The proposal of [34] (briefly described in section 2.3), presents example results in a video surveillance application, using some scenes of the PETS2002 benchmark. Also, the approaches presented by [32] and [35] (see Section 2.3)) are tested in a video surveillance environment. In [42] they present a semantic-based wireless video surveillance system, and, although they present the technical aspects of the system requirements and deployment (hardware platform), they also propose to incorporate semantics by means of object recognition, object description and object classification algorithms. A very simple ontology is constructed

with the purpose of hierarchically ordering the objects to be classified. No multi-concept relations nor spatial relations are established among objects in the scenes.

Authors have also contributed to event sharing repositories based on ontologies, with the aim of establishing open platforms for collecting annotating, retrieving and sharing surveillance videos. An example is the VISOR project [13][5], which is based on a reference ontology for video surveillance applications. It integrates hundreds of concepts, some of them taken from LSCOM [4] [18] (which has created a specialised vocabulary for news video) and MediaMill [6] ontologies. VISOR allows to export the video surveillance ontology and its video annotation using MPEG-7 and OWL standards. It is possible to perform queries based on keywords and to retrieve videos by concept, which involves to search for the desired concept in the annotation database and retrieve a list of annotations and related videos linked to that concept.

A generic concept might be represented by the video content or by its context. The content may be physical objects that appear in the scene or event/actions that happen. A video annotation is a set of class instances represented by a list of textual concepts. Some of them describe directly the nature of the instance (i.e. they are connected to an element using the relation “is-a”). Other concepts describe characteristics or properties of the instance using the relation “has-a”. A list of more than 200 video surveillance concepts can be downloaded from the VISOR project web site [5]. Some sample video frames from the VISOR project can be seen in Figure 5.

Fig. 5. Example of scenarios for video surveillance present in the VISOR project. This image was taken from [5]

[4] http://www.lscom.org
[5] http://imagelab.ing.unimore.it/visor

3. State-of-the-art Results

In order to provide a comparison frame (if possible) among the works reviewed in this report, we present a summary of works dealing with concept detection in videos, highlighting their main features (i.e. high-level features employed to improve the task), as well as their results. The summary for concept detection approaches can be seen in Table 1 and a similar summary can be seen in Table 2, for approaches incorporating semantics through ontologies.

In this tables we can see in the first column the reference of the paper in question, as well as a descriptive short text for recognizing each approach (some names are given by the authors themselves, while we defined a brief description to those unnamed). In the second column, the year when the approach was published is presented. Columns 2–6 for Table 1 and 2–5 for Table 2 depict high level features that can be employed to improve concept detection results. These are multi-concept, temporal and spatial relations, as well as textual features. In this case, when we talk about spatial relations, we are referring to the spatial configuration of objects or concepts in image space (and not to the co-occurrence of them in a shot, as many works referred to). The sixth column in Table 1 refers to the discovery of those high-level features in an unsupervised way (i.e. not using training information for discovering relations and co-ocurrences). The column named "Unit of analysis" displays whether the information extracted from video data is analyzed at frame, keyframe or shot level. Column "Dataset" shows the dataset employed for experiments and the next column show the number of concepts tested in the corresponding dataset. Results are shown in the last two columns. Column "MAP" depicts the mean average precision value and column "CR" displays the classification rate. Empty values in the table indicate that the current feature or result was not taken into account, or not mentioned in the paper.

Tabla 1. Results reported on state-of-the-art works for concept detection.

Paper/algorithm	Year	High-level information					Unit of analysis	Dataset	No. of concepts	Results	
		M-concept	Temporal	Spatial	Textual	Unsuperv.				MAP	CR
[19]/Region thesaurus	2007	X					Keyframe	TRECVID 2005	6		80.3 %
[22]/Association rules	2008	X	X				Shot	TRECVID 2005	101	0.319	
[25]/Multi-hypergraphs	2010	X	X		X		Shot	TRECVID 2005	35	0.35	
[26]/TSNB	2012	X	X				Shot	TRECVID 2005	10	0.271	
[2]/MDRF	2007	X			X		Shot	TRECVID 2006	39	0.159	
[1]/Multi-cue fusion	2008	X	X				Shot	TRECVID 2006	20	0.196	
[21]/Unsup. matrix fact.	2012	X	X			X	Shot	TRECVID 2006	20	0.187	
[26]/TSNB	2012	X	X				Shot	TRECVID 2006	20	0.197	
[1]/Multi-cue fusion	2008	X	X				Shot	TRECVID 2007	20	0.132	
[21]/Unsup. matrix fact.	2012	X	X			X	Shot	TRECVID 2007	20	0.124	
[26]/TSNB	2012	X	X				Shot	TRECVID 2007	20	0.073	
[1]/Multi-cue fusion	2008	X	X				Shot	TRECVID 2008	19	0.161	
[21]/Unsup. matrix fact.	2012	X	X			X	Shot	TRECVID 2008	19	0.151	
[26]/TSNB	2012	X	X				Shot	TRECVID 2008	19	0.058	
[26]/TSNB	2012	X	X				Shot	TRECVID 2009	8	0.061	
[26]/TSNB	2012	X	X				Shot	TRECVID 2010	30	0.208	
[6]/MediaMill search engine.	2011			X			Keyframe	TRECVID 2011	346	0.172	

Tabla 2. State-of-the-art ontology-based results reported in the literature.

Paper/algorithm	Year	High-level information				Unit of analysis	Dataset	No. of concepts	Results	
		M-concept	Temporal	Spatial	Textual				MAP	CR
[27]/Soccer ontology	2007					Shot	Soccer collection	19		63.9 %
[30]/Disasters ontology	2007	X			X	Keyframe	Disaster news videos	3		93.2 %
[7]/Composite concepts	2010	X	X			Shot	TRECVID 2005	101	[a]	
[31]/Diplomatic Policy	2011	X			X	Shot	Broadcast videos from CNN, NBC, CCTV, etc.	41	0.52	
[3]/LSCOM concepts	2011					Shot	TRECVID 2009	[b]	-	-
[34]/Object tracking-Context	2011	X	X			Frame	PETS 2002		[c]	
[32]/Stereo vision approach	2011	X	X			Shot	Traffic surveillance videos			86.5 %
[35]/VISCOM	2013	X	X	X		Keyframe	Office surveillance	12		90.0 %[d]
							Basketball videos	4		87.5 %[5]
							Football videos	3		69.0 %[5]
[42]/Semantic wireless surveillance	2013		X			Frame	Road scene	2		

[a] Authors do not provide an absolute value, but relative to a base line.
[b] Results are given only for four queries defined by the authors
[c] Experimental validation is very poor, they only show examples of results.
[d] Authors present their results using Precision/Recall values. For the sake of comparison, we mentioned only Recall in this table, which is comparable with CR

3.1. Datasets

As can be seen in Table 1, TRECVID datasets are widely used to test this kind of approaches. The TRECVID evaluation meetings [14][15] are a series of workshops with the purpose of encouraging research in the areas of content-based retrieval and exploitation of digital video. They provide a large testbed, uniform scoring procedures, and a forum for organizations interested in comparing their results. Some activities relative to the TRECVID evaluation campaign are the analysis, indexing and retrieval of video shots. From 2003 to 2012 this project has released official test collections. There are a few differences among these datasets. For example, TRECVID 2003 – 2006 were collected from multilingual news videos in American, Arabic, and Chinese broadcast channels, TRECVID 2007 – 2009 provided participants with cultural, news magazine, documentary, and education programming supplied by the Netherlands Institute for Sound and Vision and surveillance event detection was evaluated using airport surveillance video provided by the UK Home Office. TRECVID 2010 – 2012 maintained the airport surveillance video used in TRECVID 2009 and included a a new set of challenging videos (named IACC.1), resembling “web videos” with large variations in creator, content, style, production qualities, original collection device/encoding, language, etc. In general, the unit of testing and performance assessment of search tasks in TRECVID datasets is the video shot. The evaluation measure is the mean average precision which correspond to the area under an ideal (non-interpolated) recall/precision curve.

On the other hand, in Table 2, the first thing to notice is the heterogeneous nature of the datasets employed. Most works dealing with ontologies restrict their research scope to a limited domain and create their own knowledge representations and test data. This is a problem when we try to compare the performance of different approaches, since most results are not comparable.

3.2. Evaluation Measures

In the results surveyed in this report, three main evaluation measures are used: mean average precision (MAP), classification rate (CR) and Precision-recall values.

Precision and recall are single-value metrics based on the whole list of documents returned by the system. Precision is the fraction of the relevant retrieved documents over all the retrieved documents. Recall is the fraction of relevant retrieved documents over all relevant documents. In this sense, recall can be also seen as the classification rate, since it measures how many of the classified documents are correct, versus the real class of the documents.

Mean Average precision is a very popular performance measure in information retrieval in general. Average precision (AP) is used to score document retrieval and it captures the importance of the ranking or ordering of the retrieved documents. For systems that return a ranked sequence of documents, it is desirable to also consider the order in which the returned documents are presented. When precision-recall values are computed at every position in the ranked sequence of documents, a precision-recall curve is obtained (precision vs. recall). AP computes the average value of precision in the recall interval $[0,1]$, which is actually the area under the precision-recall curve. AP, when averaged over all queries and reported as a single score, is called mean average precision.

3.3. Discussion

Tables 1 and 2 comprise a considerable amount of approaches related to content-based video search and give a global vision regarding most prominent aspects in each research. Several major concerns arise when analyzing this information:

- Regarding the high-level information used, it can be seen that very few approaches incorporate several high-level features at the same time. Many approaches dealing with content-based image retrieval and object recognition in still images have proposed solutions for dealing with high-level information at image level, but this knowledge is rarely extrapolated to video content analysis. Most works related to semantic video retrieval just assume a naive representation/classification of images, and focus more in adding information regarding temporal or motion aspects.
- In many works, the shot is the basic unit of analysis, which means that concept detection is performed at shot level and the relations are establish among shots (ex. temporal relations are established among consecutive shots). This is highly related to the previous item, since using the shot as basic unit means that each shot (sequence of frames) will behave as a bag of unordered features, and multi-concept detection will be performed on them, disregarding then the spatial information among features or concepts in image space, as well as the temporal cues that connected frames may provide to recognize concepts.
- Benchmarks employed for displaying results are heterogeneous. Even though most of them in Table 1 use TRECVID datasets, for different editions of the competition, data is different and systems are different. Only metrics, in most cases, are the same. Also, the evaluation protocol in each work is different, even when using the same dataset. Researchers evaluate their algorithm performance for different amounts of concepts (from 6 to 346), therefore, results are not comparable. Also, some works use different evaluation metrics in the same dataset, which is another problem when trying to establish advantages among them. More concerning in this sense is the information provided in Table 2, where no single work employs the same dataset or knowledge representation of any other work. Furthermore, all domain specific approaches presented collect and structure their own test data.

- One major concern is related to the results. Even though they are not comparable, it can be notice that in Table 1 MAP scores range from 0.05 to 0.31, which is still a very low score for these tasks. In Table 2, most results are given in terms of classification rate, which ranges from 63.9 % to 93.2 %. It is important to notice, nevertheless, that even getting some scores above 90 %, the testing domains are rather limited and with very few concepts.

This indicates that there is still much work to be done in this field, in several aspects: improving low-level representation of videos, taking advantages of methods developed for still images; finding new ways to represent and combine high-level information that can be used to add semantics to video analysis tasks and trying to unify results in order to make them comparable with state-of-the-art methods.

4. Conclusions

Despite the fact that performance improvements have been reported in the last years in the field of semantic video search, and that many approaches have been created to extend the number of different concept classifiers, to add meaningful relations and to provide knowledge representations for video content, there is still much work to be done.

Several approaches reviewed in this report have proposed approaches for robust detection and representation of high-level concepts, for mining multi-concept and temporal relations, for modeling of events and approaches to represent domain knowledge and contextual information of activities and actions. These methods have been applied to several different domains, from sport to surveillance videos, showing promising results, but clearly they are still not good enough. Last section of this report summarized the main aspects and results of this works, and the main concerning issues were discussed.

Poor results, in general, means that more deeper evaluation of each aspect should be made, ranging from concept detection to ontology development areas. The advances made so far need to be consolidated, in terms of their robustness to real-world conditions and, especially for surveillance applications, there is need of reaching realtime performance.

References

1. Weng, M.F., Chuang, Y.Y.: Multi-cue fusion for semantic video indexing. In El-Saddik, A., Vuong, S., Griwodz, C., Bimbo, A.D., Candan, K.S., Jaimes, A., eds.: ACM Multimedia, ACM (2008) 71–80
2. Hauptmann, A.G., yu Chen, M., Christel, M.G., Lin, W.H., 0003, J.Y.: A hybrid approach to improving semantic extraction of news video. In: ICSC, IEEE Computer Society (2007) 79–86
3. Shirahama, K., Uehara, K.: Effectiveness of video ontology in query by example approach. In Zhong, N., Callaghan, V., Ghorbani, A.A., Hu, B., eds.: AMT. Volume 6890 of Lecture Notes in Computer Science., Springer (2011) 49–58
4. Bertini, M., Del Bimbo, A., Torniai, C., Cucchiara, R., Grana, C.: Mom: multimedia ontology manager. a framework for automatic annotation and semantic retrieval of video sequences. In: Proceedings of the 14th annual ACM international conference on Multimedia. MULTIMEDIA '06, New York, NY, USA, ACM (2006) 787–788
5. Vezzani, R., Cucchiara, R.: Video surveillance online repository (visor). In: Proceedings of the ACM Multimedia Systems 2013, Oslo, Norway (February 2013)
6. Snoek, C.G.M., van de Sande, K.E.A., Li, X., Mazloom, M., Jiang, Y.G., Koelma, D.C., Smeulders, A.W.M.: The MediaMill TRECVID 2011 semantic video search engine. In: Proceedings of the 9th TRECVID Workshop, Gaithersburg, USA (December 2011)
7. Ballan, L., Bertini, M., Bimbo, A.D., Serra, G.: Video annotation and retrieval using ontologies and rule learning. IEEE MultiMedia **17**(4) (2010) 80–88
8. Mariano, V.Y., Min, J., Park, J.H., Kasturi, R., Mihalcik, D., Li, H., Doermann, D.S., Drayer, T.: Performance evaluation of object detection algorithms. In: ICPR (3). (2002) 965–969

9. lamp: Viper-gt, available at `http://viper-toolkit.sourceforge.net/products/gt/`. url: http://viper-toolkit.sourceforge.net/products/gt/ (2007)
10. Home Office Scientific Development Branch: Imagery library for intelligent detection systems (i-LIDS). url: http://homeoffice.gov.uk (2007)
11. Nghiem, A.T., BrÃ©mond, F., Thonnat, M., Valentin, V.: Etiseo, performance evaluation for video surveillance systems. In: AVSS, IEEE Computer Society (2007) 476–481
12. Disruptive Technology Office: Video analysis and content extraction (VACE). url: http://www.informedia.cs.cmu.edu/arda/index.html (2003)
13. Vezzani, R., Cucchiara, R.: Video surveillance online repository (ViSOR): An integrated framework. Multimedia Tools and Applications **50**(2) (2010) 359–380
14. Smeaton, A., Over, P., Kraaij, W.: TRECVID: Evaluating the effectiveness of information retrieval tasks on digital video. In: Proceedings of the ACM MM'04, ACM (October 2004) 652–655
15. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: Proceedings of the 8th ACM international workshop on Multimedia information retrieval. MIR '06, New York, NY, USA, ACM (2006) 321–330
16. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proceedings of the 14th annual ACM international conference on Multimedia. MULTIMEDIA '06, New York, NY, USA, ACM (2006) 421–430
17. Yanagawa, A., Chang, S.F., Kennedy, L., Hsu, W.: Columbia university's baseline detectors for 374 lscom semantic visual concepts. Technical report, Columbia University (March 2007)
18. Naphade, M., Smith, J.R., Tesic, J., Chang, S.F., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J.: Large-scale concept ontology for multimedia. IEEE MultiMedia **13**(3) (July 2006) 86–91
19. Spyrou, E., Avrithis, Y.S.: High-level concept detection in video using a region thesaurus. In Maglogiannis, I., Karpouzis, K., Wallace, M., Soldatos, J., eds.: Emerging Artificial Intelligence Applications in Computer Engineering. Volume 160 of Frontiers in Artificial Intelligence and Applications. IOS Press (2007) 143–153
20. Liu, Y., Mei, T., Hua, X.S.: Crowdreranking: exploring multiple search engines for visual search reranking. In Allan, J., Aslam, J.A., Sanderson, M., Zhai, C., Zobel, J., eds.: SIGIR, ACM (2009) 500–507
21. Weng, M.F., Chuang, Y.Y.: Collaborative video re-indexing via matrix factorization. ACM Transactions on Multimedia Computing, Communications and Applications **8**(2) (May 2012) 23:1–23:20
22. Liu, K.H., Weng, M.F., Tseng, C.Y., Chuang, Y.Y., Chen, M.S.: Association and temporal rule mining for post-filtering of semantic concept detection in video. IEEE Transactions on Multimedia **10**(2) (2008) 240–251
23. Weng, M.F., Chuang, Y.Y.: Cross-domain multi-cue fusion for concept-based video indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(10) (October 2012) 1927–1941
24. Aytar, Y., Shah, M., Luo, J.: Utilizing semantic word similarity measures for video retrieval. In: CVPR, IEEE Computer Society (2008)
25. Han, Y., Shao, J., Wu, F., Wei, B.: Multiple hypergraph ranking for video concept detection. Journal of Zhejiang University - Science C **11**(7) (2010) 525–537
26. Geng, J., Miao, Z., Chi, H.: Temporal-spatial refinements for video concept fusion. (2012) III:547–559
27. Bertini, M., Bimbo, A.D., Torniai, C., Grana, C., Cucchiara, R.: Dynamic pictorial ontologies for video digital libraries annotation. In Fotouhi, F., Grosky, W.I., Stanchev, P.L., eds.: MS, ACM (2007) 47–56
28. Li, Q., Lu, Z., Yu, Y., Liang, L.: Multimedia ontology modeling: An approach based on MPEG-7. In: International Conference on Advanced Computer Control. (2011)
29. Arndt, R., Troncy, R., Staab, S., Hardman, L.: Comm: A Core Ontology For Multimedia Annotation. In Staab, S., Studer, R., eds.: Handbook on Ontologies. second edn. Springer Verlag (2009)
30. Papadopoulos, G.T., Mezaris, V., Kompatsiaris, I., Strintzis, M.G.: Ontology-driven semantic video analysis using visual information objects. In Falcidieno, B., Spagnuolo, M., Avrithis, Y.S., Kompatsiaris, I., Buitelaar, P., eds.: Semantic Multimedia, Second International Conference on Semantic and Digital Media Technologies, SAMT 2007, Genoa, Italy, December 5-7, 2007, Proceedings. Volume 4816 of Lecture Notes in Computer Science., Springer (2007) 56–69
31. Bai, L., Lao, S., Guo, J.: Video semantic concept detection using ontology. In: Proceedings of the Third International Conference on Internet Multimedia Computing and Service. ICIMCS '11, New York, NY, USA, ACM (2011) 158–163
32. Raluca Brehar, Carolina Fortuna, S.B.D.M.S.N.: Spatio-temporal reasoning for traffic scene understanding. In: IEEE International Conference on Intelligent Computer Communication and Processing - ICCP'2011. (2011)
33. Matuszek, C., Cabral, J., Witbrock, M., Deoliveira, J.: An introduction to the syntax and content of cyc. In: Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering. (2006) 44–49
34. GóMez-Romero, J., Patricio, M.A., GarcíA, J., Molina, J.M.: Ontology-based context representation and reasoning for object tracking and scene interpretation in video. Expert Syst. Appl. **38**(6) (June 2011) 7494–7510
35. Yildirim, Y., Yazici, A., Yilmaz, T.: Automatic semantic content extraction in videos using a fuzzy ontology and rule-based model. IEEE Trans. Knowl. Data Eng. **25**(1) (2013) 47–61

36. Nevatia, R., Hobbs, J., Bolles, B.: An ontology for video event representation. In: Computer Vision and Pattern Recognition, IEEE (2004)
37. Hakeem, A., Shah, M.: Ontology and taxonomy collaborated framework for meeting classification. In: ICPR (4). (2004) 219–222
38. Georis, B., Maziere, M., Bremond, F., Thonnat, M.: A video interpretation platform applied to bank agency monitoring. In: Proceedings of IDSS'04 - 2nd Workshop on Intelligent Distributed Surveillance Systems. (FEB 23 2004)
39. Akdemir, U., Turaga, P.K., Chellappa, R.: An ontology based approach for activity recognition from video. In El-Saddik, A., Vuong, S., Griwodz, C., Bimbo, A.D., Candan, K.S., Jaimes, A., eds.: ACM Multimedia, ACM (2008) 709–712
40. Pattanasri, N., Jatowt, A., Tanaka, K.: Enhancing comprehension of events in video through explanation-on-demand hypervideo. In Cham, T.J., Cai, J., Dorai, C., Rajan, D., Chua, T.S., Chia, L.T., eds.: MMM (1). Volume 4351 of Lecture Notes in Computer Science., Springer (2007) 535–544
41. Scherp, A., Franz, T., Saathoff, C., Staab, S.: F–a model of events based on the foundational ontology dolce+dns ultralight. In Gil, Y., Noy, N.F., eds.: K-CAP, ACM (2009) 137–144
42. Gu, H., Liu, W., Zhao, X.: Semantic-based wireless video surveillance system. In Du, Z., ed.: Intelligence Computation and Evolutionary Computation. Volume 180 of Advances in Intelligent Systems and Computing. Springer Berlin Heidelberg (2013) 965–973

RT_061, abril 2014
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba