Towards Ontological Cognitive System

Carles Fernandez[1], Jordi Gonzàlez[1], João Manuel R. S. Tavares[2] and F. Xavier Roca[1]

1 Dep. Computer Science & Computer Vision Centre, Edifici O. Universitat Autonoma de Barcelona, 08193, Bellaterra, Spain

2 Instituto de Engenharia Mecânica e Gestão Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, S/N - 4200-465 Porto, Portugal

Abstract The increasing ubiquitousness of digital information in our daily lives has positioned video as a favored information vehicle, and given rise to an astonishing generation of social media and surveillance footage. This raises a series of technological demands for automatic video understanding and management, which together with the compromising attentional limitations of human operators, have motivated the research community to guide its steps towards a better attainment of such capabilities. As a result, current trends on cognitive vision promise to recognize complex events and self-adapt to different environments, while managing and integrating several types of knowledge. Future directions suggest to reinforce the multi-modal fusion of information sources and the communication with end-users.

Introduction

The revolution of information experienced by the world in the last century, especially emphasized by the household use of computers after the 1970s, has led to what is known today as the society of knowledge. Digital technologies have converted post-modern society into an entity in which networked communication and information management have become crucial for social, political, and economic practices. The major expansion in this sense has been rendered by the global effect of the Internet: since its birth, it has grown into a medium that is uniquely capable of integrating modes of communication and forms of content.

In this context, the assessment of interactive and broadcasting services has spread and generalized in the last decade –e.g., residential access to the Inter-net, video-on-demand technologies–, posing video as the privileged information vehicle of our time, and promising a wide variety of applications that aim at its efficient exploitation. Today, the automated analysis of video resources is not tomorrow's duty anymore. The world produces a massive amount of digital video files

every passing minute, particularly in the fields of multimedia and surveillance, which open windows of opportunity for smart systems as vast archives of recordings constantly grow.

Automatic content-based video indexing has been requested for digital multimedia databases for the last two decades (Foresti et al. 2002). This task consists of extracting high-level descriptors that help us to automatically annotate the semantic content in video sequences; the generation of reasonable semantic indexes makes it possible to create powerful engines to search and retrieve video content, which finds immediate applications in many areas: from the efficient access to digital libraries to the preservation and maintenance of digital heritage. Other usages in the multimedia domain would also include virtual commentators, which could describe, analyze, and summarize the development of sport events, for instance.

More recently, the same requirements have applied also to the field of video surveillance. Human operators have attentional limitations that discourage their involvement in a series of tasks that could compromise security or safety. In addition, surveillance systems have strong storage and computer power requirements, deal with continuous 24/7 monitoring, and manage a type of content that is susceptible to be highly compressed. Furthermore, the number of security cameras increases exponentially worldwide on a daily basis, producing huge amounts of video recordings that may require further supervision. The conjunction of these facts establishes a need to automatize the visual recognition of events and content-based forensic analysis on video footage.

We find a wide range of applications coming from the surveillance domain that point to real-life, daily problems: for example, a smart monitoring of elder or disabled people makes it possible to recognize alarming situations, and speed up reactions towards early assistance; road traffic surveillance can be useful to send alerts of congestion or automatically detects accidents or abnormal occurrences; similar usage can be directed to urban planning, optimization of resources for transportation allocations, or detection of abnormality in crowded locations such in airports or lobbies.

Such a vast spectrum of social, cultural, commercial, and technological demands have repeatedly motivated the research community to direct their steps towards a better attainment of video understanding capabilities.

1.1 Collaborative efforts on video event understanding

A notable amount of European Union (EU) research projects have been recently devoted to the unsupervised analysis of video contents, in order to automatically extract events and behaviors of interest, and interpret them in selected contexts. These projects measure the pulse of the research in this field demonstrate previous success on particular initiatives, and propose a series of interesting applications to

such techniques. And, last but not least, they motivate the continuation of this line of work. Some of them are briefly described next, and depicted in Figures 1-2.

ADVISOR (IST-11287, 2000–2002): It addresses the development of management systems for networks of metro operators. It uses Closed Circuit Television (CCTV) for computer assisted automatic incident detection, content based annotation of video recordings, behavior pattern analysis of crowds and individuals, and ergonomic human computer inter-faces.

ICONS (DTI/EPSRC LINK, 2001–2003): Its aim is to advance towards (i) zero motion detection, detection of medium- to long-term visual changes in a scene, e.g., deployment of a parcel bomb, theft of a precious item, and (ii) behavior recognition –characterize and detect undesirable behavior in video data, such as thefts or violence, only from the appearance of pixels.

AVITRACK (AST-CT-502818, 2004–2006): It develops a framework for automatically supervision of commercial aircraft servicing operations from the arrival to the departure on an airport's apron. A prototype for scene understanding and simulation of the apron's activity was going to be implemented during the project on Toulouse airport.

ETISEO (Techno-Vision, 2005–2007): It seeks to work out a new structure contributing to an increase in the evaluation of video scene understanding. ETISEO focuses on the treatment and interpretation of videos involving pedestrians and (or) vehicles, indoors or outdoors, obtained from fixed cameras.

CARETAKER 5 (IST-027231, 2006–2008): This project aims at studying, developing and assessing multimedia knowledge-based content analysis, knowledge extraction components, and metadata management sub-systems in the context of automat-ed situation awareness, diagnosis and decision sup-port.

SHARE 6 (IST-027694, 2006–2008): It offers an information and communication system to support emergency teams during large-scale rescue operations and disaster management, by exploiting multimodal data, as audio, video, texts, graphics, location. It in-corporates domain dependent ontology modules, and allows for video/voice analysis, indexing/retrieval, and multimodal dialogues.

HERMES (IST-027110, 2006–2009): Extraction of descriptions of people's behavior from videos in restricted discourse domains, such as inter-city roads, train stations, or lobbies. The project studies human movements and behaviors at several scales, addressing agents, bodies and faces, and the final communication of meaningful contents to end-users.

BEWARE (EP/E028594/1, 2007–2010): The project aims to analyze and combine data from alarm panels and systems, fence detectors, security cameras, public sources and even police files, to unravel patterns and signal anomalies, e.g., by making comparisons with historical data. BEWARE is self-learning and suggests improvements to optimize security.

VIDI-Video (IST-045547, 2007–2010): Implementation of an audio-visual semantic search engine to enhance access to video, by developing a 1000 element thesaurus to index video content. Several applications have been suggested in sur-

veillance, conferencing, event reconstruction, diaries, and cultural heritage documentaries.

SAMURAI (IST-217899, 2008–2011): It develops an intelligent surveillance system for monitoring of critical public infrastructure sites. It is to fuse data from networked heterogeneous sensors rather than just using CCTV; to develop real-adaptative behavior profiling and abnormality detection, instead of using predefined hard rules; and to take command input from human operators and mobile sensory in-put from patrols, for hybrid context-aware behavior recognition.

SCOVIS (IST-216465, 2007–2013): It aims at automatic behavior detection and visual learning of procedures, in manufacturing and public infrastructures. Its synergistic approach based on complex camera networks also achieves model adaptation and camera network coordination. User's interaction improves behavior detection and guides the modeling process, through high-level feedback mechanisms.

ViCoMo (ITEA2-08009, 2009–2012): This project concerns advanced video interpretation algorithms on video data that are typically acquired with multiple cameras. It is focusing on the construction of realistic context models to improve the decision making of complex vision systems and to produce a faithful and meaningful behavior.

As it can be realized from the aforementioned projects, many efforts have been taken in the last decade, and are still increasing nowadays, in order to tackle the problem of video interpretation and intelligent video content management. It is clear from this selection that current trends on the field suggest a tendency to focus on the multi-modal fusion of different sources of information, and on more powerful communication with end-users. From the large amount of projects existing in the field we derive another conclusion: such a task is not trivial at all, and requires research efforts from many different areas to be joint into collaborative approaches, which success where individual efforts fail.

Fig. 1. Snapshots of the referred projects: (a) AVITRACK, (b) ADVISOR, (c) BEWARE, (d) VIDI-Video, (e) CARE-TAKER, (f) ICONS, (g) ETISEO and (h) HERMES.

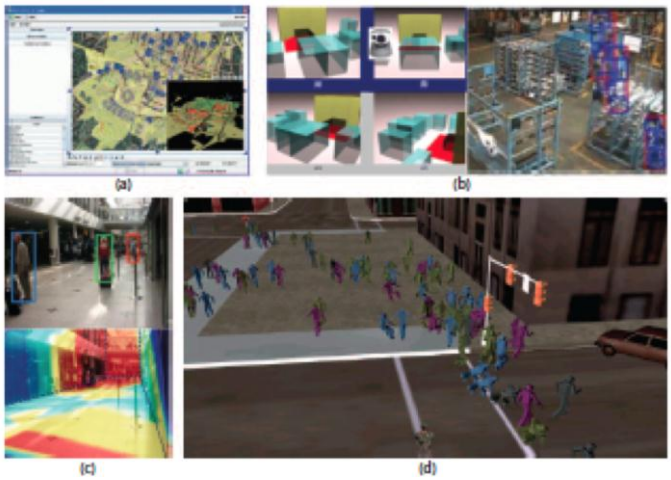


Fig. 2. Figure 2. Snapshots of the most recent projects in the field: (a) SHARE, (b) SCOVIS, (c) SAMURAI and (d) ViCoMo.

Pas, present and future of video surveillance.

The field of video surveillance has experienced a remarkable evolution in the last decades, which can help us think of the future characteristics that would be desirable for it. In the traditional video surveillance scheme, the primary goal of the camera system was to present to human operators more and more visual information about monitored environments, see Figure 3. First-generation systems were completely passive, thus having this information entirely processed by human operators. Nevertheless, a saturation effect appears as the information availability increases, causing a decrease in the level of attention of the operator, who is ultimately responsible of deciding about the surveilled situations.

The following generation of video surveillance systems used digital computing and communications technologies to change the design of the original architecture, customizing it according to the requirements of the end-users. A series of technical advantages allowed them to better satisfy the demands from industry, i.e., higher-resolution cameras, longer retention of recorded video, Digital Video Recorders (DVRs) replaced Video Cassette Recorders (VCRs) and video encoding standards appeared, reduction of costs and size, remote monitoring capabilities provided by network cameras, or more built-in intelligence, among others (Nilsson 2009).

The continued increase of machine intelligence has derived into a new generation of smart surveillance systems lately. Recent trends on computer vision and artificial intelligence have deepened into the study of cognitive vision systems, which use visual information to facilitate a series of tasks on sensing, understanding, reaction, and communication, see Figure 3(b). Such systems enable traditional surveillance applications to greatly enhance their functionalities by incorporating methods for:

1. Recognition and categorization of objects, structures, and events.
2. Learning and adaptation to different environments.
3. Representation, memorization, and fusion of various types of knowledge.
4. Automatic control and attention.

As a consequence, the relation of the system with the world and the end-users is enriched by a series of sensors and actuators, e.g., distributions of static and active cameras, enhanced user interfaces; thus, establishing a bidirectional communication flow, and closing loops at a sensing and semantic level. The resulting systems provide a series of novel applications with respect to traditional systems, like automated video commentary and annotation, or image-based search engines. In the last years, European projects, like CogVis or CogViSys, have investigated the-

se and other potential applications of cognitive vision systems, especially concerning video surveillance.

Recently, a paradigm has been specifically proposed for the design of cognitive vision systems aiming to analyze human developments recorded in image sequences. This is known as Human Sequence Evaluation (HSE) (Gonzàlez et al. 2009). An HSE system is built upon a linear multilevel architecture, in which each module tackles a specific abstraction level. Two consecutive modules hold a bidirectional communication scheme, in order to:

1. generate higher-level descriptions based on lower-level analysis, i.e., bottom-up inference, and
2. support low-level processing with high-level guidance, i.e., top-down reactions. HSE follows as well the aforementioned characteristics of cognitive vision systems.

Nonetheless, although cognitive vision systems conduct a large number of tasks and success in a wide range of applications; in most cases, the resulting prototypes are tailored to specific needs or restricted to definite domains. Hence, current research aims to increase aspects like extensibility, personalization, adaptability, interactivity, and multi-purpose of these systems. In particular, it is becoming of especial importance to stress the role of communication with end-users in the global context, both for the fields of surveillance and multimedia: end-users should be allowed to automatize a series of tasks requiring content-mining, and should be presented the analyzed information in a suitable and efficient manner, see Figure 3(c).

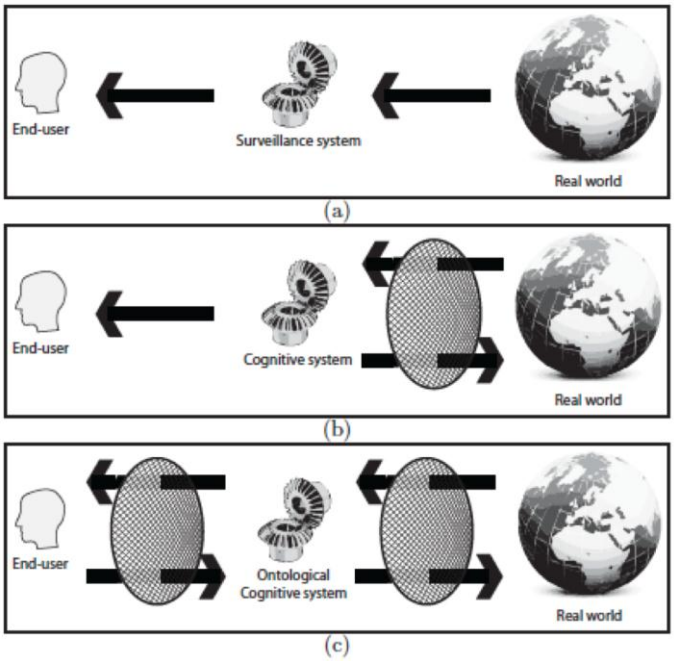


Fig. 3. Evolution of video surveillance systems, since its initial passive architecture (a) to the reactive, bidirectional communication scheme offered by cognitive vision systems (b), which highlight relevant footage contents. By incorporating ontological and interactive capabilities to this framework (c), the system performs like a semantic filter also to the end-users, governing the interactions with them in order to adapt to their interests and maximize the efficiency of the communication

As a result of these considerations, the list of objectives to be tackled and solved by a cognitive vision system has elaborated on the original approach, which aimed at the single – although still ambitious today – task of transducing images to semantics. Nowadays, the user itself has become a piece of the puzzle, and therefore has to be considered a part of the problem.

Mind the gaps.

The search and extraction of meaningful information from video sequences are dominated by 5 major challenges, all of them defined by gaps (Smeulders et al. 2000). These gaps are disagreements between the real data and that one expected, intended, or retrieved by any computer-based process involved in the information flow con-ducted between the acquisition of data from the real world, and until its final presentation to the end-users. The 5 gaps are described next; see Figure 4(a).

1. Sensory gap: The gap between an object in the world and the in-formation in an image recording of that scene. All these recordings will be different due to variations in viewpoint, lighting, and other circumstantial conditions.
2. Semantic gap: The lack of coincidence between the information that one can extract from the sensory data and the interpretation that same data has for a user in a given situation. It can be understood as the difference be-tween a visual concept and its linguistic representation.
3. Model gap: The impossibility to theoretically account the amount of notions in the world, due to the limited capacity to learn them
4. Query/context gap: The gap between the specific need for information of an end-user and the possible retrieval solutions manageable by the system.
5. Interface gap: The limited scope of information that a system inter-face offers compared to the amount of data actually intended to transmit.

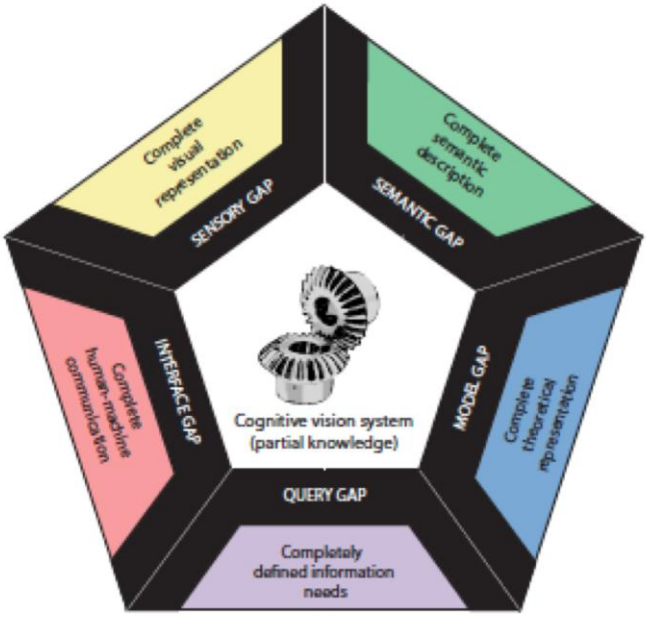
Although each of these challenges becomes certainly difficult to overcome by its own, a proper centralization of information sources and the wise reutilization of knowledge derived from them facilitates the overwhelming task of bridging each of these gaps. There exist multiple examples of how the multiple resources of the system can be redirected to solve problems in a different domain, let us consider three of them:

- From semantic to sensory gap: tracking errors or occlusions at a visual level can be identified by high-level modules that imply semantics oriented to that end. This way, the system can be aware of where and when a target is occluded, and predict its repartition.
- From sensory to interface gap: the reports or responses in user interfaces can become more expressive by adding selected, semantically relevant key-frames from the sensed data.
- From interface to query gap: in case of syntactic ambiguities in a query, e.g., “zoom in on any person in the group that is running”, end-users can be asked about their real interests via a dialogue interface: “Did you mean ‘the group that is running’, or ‘the person that is running?’.

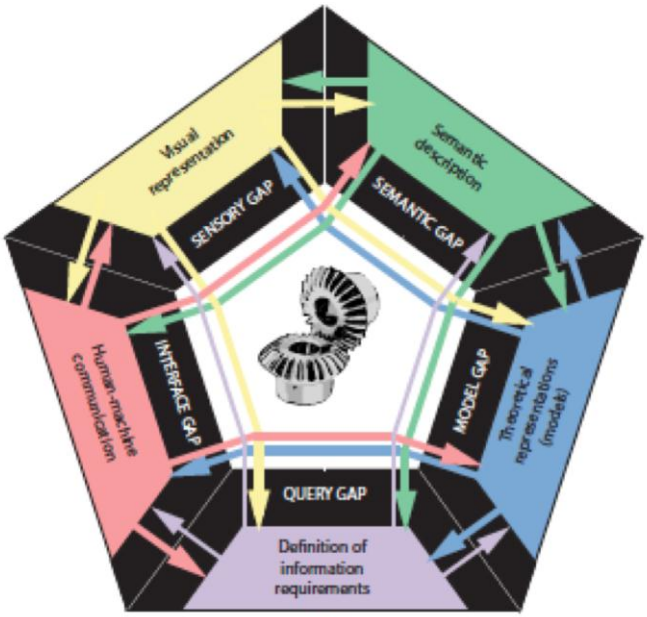
Given the varied nature of types of knowledge involved in our intended system, an ontological framework becomes a sensible choice of design: such a framework integrates different sources of in-formation by means of temporal and multi-modal fusion, i.e., horizontal integration, using bottom-up or top-down approaches, i.e., vertical integration, and incorporating prior hierarchical knowledge by means of an extensible ontology.

We propose the use of ontologies to help us integrate, centralize, and relate the different knowledge representations, such as visual, semantic, linguistic, etc., implied by the different modules of the cognitive system. By doing so, the relevant knowledge or capabilities in a specific area can be used to enhance the performance of the system in other distinct areas, as represented in Figure 4(b). Ontologies will enable us to formalize, account, and redirect the semantic assets of the

system in a given situation, and exploit them to empower the aforementioned capabilities, especially targeting the possibilities of interaction with end-users.



(a)



(b)

Fig. 4. (a) The five gaps that need to be bridged for the successful analysis, extraction, search, retrieval, and presentation of video content. (b) In some cases, a collaborative and integrative use of different knowledge sources allows us to achieve or enrich the accomplishment of these tasks. (Arrows stand for reusing ontological knowledge to enhance analyses in other areas.)

Ontologies to enhance video understanding.

It has been repeatedly stated how ontologies can be used effectively for relating semantic descriptors to image or video content, or at least use them to represent and fuse structured prior information from different sources towards that end (Town 2006). Sever-al classical methods from artificial intelligence to represent or match ontological knowledge–e.g., Description Logics (DL), frame-based representations, semantic networks– are becoming popular again since the start of the Semantic Web initiative (Kompatsiaris and Hobson 2008) (Baader et al. 2003). Nevertheless, the challenge today is how to apply these approaches to highly ambiguous or uncertain information, like that coming from language and vision, respectively. For this reason, the incorporation of ontologies into cognitive vision systems has also awakened the interests of many researchers in the field (Maillot et al. 2004) (Staab and Studer 2004). The use of DL to model uncertainty has been long discussed; an overview of the research in this field is presented in (Baader et al. 2003).

In the case of video surveillance, ontologies have been used to assist to the recognition of video events. Several authors have engaged initiatives to standardize taxonomies of video events, e.g., (Nevatia et al. 2004) proposed a formal language to describe event ontologies, VERL, and a markup language, VEML, to annotate instances of ontological events. The use of this language is exemplified in videos from the security and meeting domains. In (Ma and McKevitt 2004) the authors present an ontology of eventive verbs for multimodal storytelling system including visual and linguistic concepts.

Regarding the field of multimedia, the automatic processing of multimedia content has been enhanced by the apparition of new multimedia standards, such as MPEG-7, which provide basic functionalities in order to manipulate and transmit objects and metadata, and measure similarity in images or video based on visual criteria. However, most of the se-mantic content of video data is out of the scope of these standards. In these cases, ontologies are often used to extend standardized multimedia annotation by means of concept hierarchies (Troncy et al. 2007) (Jaimes and Chang 2000), and also to provide meaningful query languages –e.g., RDQL or SWRL– as tools to build, annotate, integrate, and learn ontological information. An overview of such languages is presented in (Zhang and Miller 2005). There have been efforts towards the generation of textual representations and summaries from ontologies (Bontcheva 2005) (Wilcock 2003). In fact, these approaches are general-purpose ontology verbalizers, agnostic of the class types and their properties, which result in outputs that are in general too verbose and redundant. Our contribution adapts the textual descriptions and summaries to the

type of content described, regarding its organization into the modeled domain ontology.

Ontology-based approaches are also suitable for designing processes to query, report, or mine data from distributed and heterogeneous sources. These capabilities derive a series of tasks that are usually requested in the domain of multimedia semantics, such as automatic video annotation to enable query-based video retrieval. In (Bertini et al. 2008) the authors have recently presented an ontology-based framework for semantic video annotation based on the learning of spatio-temporal rules. First Order Inductive Learner (FOIL) is adapted to learn rule patterns that have been then validated on some TRECVID video events. Similarly, other approaches emphasize the use of ontologies to enable forensic applications in video surveillance (Vadakkeveedu et al. 2007).

The understanding of linguistic events has also been approached with ontologies. For instance, Cimiano et al. (Cimiano et al. 2005) presented an ontology-driven approach that, based on Dis-course Representation Theory from linguistics, computes conceptual relations between events extracted form a text and a referring expression representing some other event, a state or an entity. Recent large-scale endeavors like the Virtual Human Project (Hartholt et al. 2008) propose a complete architecture for virtual humans, including NL capabilities for generation and under-standing, speech recognition and text-to-speech synthesis, task reasoning, behavior blending, and virtual environment generation. An ontological design was chosen for flexibility and extensibility, and to deal with the many multimodal representations of knowledge considered. This work stresses the importance of ontologies especially when relating language and concepts.

Conclusions

Human Sequence Evaluation (HSE) concentrates on how to extract descriptions of human behaviour from videos in a restricted discourse domain, such as (i) pedestrians crossing inner-city roads where pedestrians appear approaching or waiting at stops of busses or trams, and (ii) humans in indoor worlds like an airport hall, a train station, or a lobby. These discourse domains allow exploring a coherent evaluation of human movements and facial expressions across a wide variation of scale. This general approach lends itself to various cognitive surveillance scenarios at varying degrees of resolution: from wide-field-of-view multiple-agent scenes, through to more specific inferences of emotional state that could be elicited from high resolution imagery of faces. The true challenge of the HERMES project will consist in the development of a system facility, which starts with basic knowledge about pedestrian behaviour in the chosen discourse domain, but could cluster evaluation results into semantically meaningful subsets of behaviours. The envisaged system will comprise an internal logic-based representation, which enables it to comment each individual subset, giving natural language explanations of why the system has created the subset in question.

Multiple issues will be contemplated to perform HSE, such as detection and localization; tracking; classification; prediction; concept formation and visualization; communication and expression, etc. And this is reflected in the literature: a huge number of papers confront some of the levels, but rarely all of them. Summarizing, agent motion will allow HSE to infer behavior descriptions. The term behaviour will refer to one or several actions, which acquire a meaning in a particular context.

Body motion will allow HSE to describe action descriptions. We define an action as a motion pattern, which represents the style of variation of a body posture during a predefined interval of time. Therefore, body motion will be used to recognize style parameters, such as age, gender, handicapped and identification.

Lastly, face motion will lead to emotion descriptions. The emotional characteristics of facial expressions will allow HSE to confront personality modeling, which would enable us to carry out multiple studies and researches on advanced human-computer interfaces.

So these issues will require, additionally, assessing how, and by which means, the knowledge of context and a plausible hypothesis about he internal state of the agent may influence and support the interpretation processes.

Acknowledgments The authors wish to acknowledge the support of the Spanish Research Programs Consolider-Ingenio 2010: MIPRCV (CSD200700018); Avanza I+D Di-CoMa (TSI-020400-2011-55); along with the Span-ish projects TIN2009-14501-C02-01 and TIN2009-14501-C02-02; MICIN the A.I. PT2009-0023.

Foresti G.L., Marcenaro L., Regazzoni C.S. (2009) Automatic detection and indexing of video-event shots for surveillance applications. IEEE Transactions on Multimedia, 4(4):459–471, 2002.

Gonzàlez J., Rowe D., Varona J. Roca X. (2009) Understanding dynamic scenes based on human sequence evaluation. Image and Vision Computing, 27(10):1433–1444,.

Nilsson F. Intelligent network video: understanding modern video surveillance systems. CRC Press, 2009.

Smeulders A.W.M., Worring M., Santini S., Gupta A., Jain R. (2000) Contentbased image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349–1380.

Town C. (2006) Ontological inference for image and video analysis. Machine Vision and Applications, 17(2):94–115.

Kompatsiaris Y. Hobson P. (2008) Semantic multimedia and ontologies: theory and applications. Springer.

Baader, F., Calvanese D., McGuinness D.L, Patel-Schneider P., Nardi D. (2003) The description logic handbook: theory, implementation, and applications. Cambridge Univ Pr.

Maillot N., Thonnat M., Boucher A. (2004) Towards ontology-based cognitive vision. Machine Vision and Applications, 16(1):33–40.

Staab S., Studer (2004) R.. Handbook on ontologies. Springer.

Baader F., Calvanese D., McGuiness D., Nardi D., Patel-Schneider P. (2003). The Description Logic handbook. Cambridge University Press, Cambridge, UK.

Nevatia R., Hobbs J., Bolles B. (2004) An ontology for video event representation. In Proceedings of the international workshop on Detection and Recognition of Events in Video.

Ma M. Mc Kevitt P.. Visual semantics and ontology of eventive verbs. (2004) In Proc. of the 1st International Joint Conference on Natural Language Processing, pages 278–285.

Troncy R., Celma O., Little S., Garcıa R., Tsinaraki C. (2007) Mpeg-7 based multimedia ontologies: Interoperability support or interoperability issue. In 1st International Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies, pages 2–15.

Jaimes A. Chang S. (2000) A conceptual framework for indexing visual information at multiple levels. In Proceedings of the IS&T SPIE Internet Imaging.

Zhang Z. Miller J.A.. (2005) Ontology query languages for the semantic web: A performance evaluation. Journal of Web Semantics.

Bontcheva K. (2005) Generating tailored textual summaries from ontologies. In Proc. of the Extended Semantic Web Conference.

Wilcock G. Talking (2003) OWLs: towards an ontology verbalizer. In Proc. of the International Semantic Web Conference.

Bertini M., Del Bimbo A., Serra G. (2008) Learning rules for semantic video event annotation. In Proceedings of the international conference on Visual Information Systems (VISUAL).

Vadakkeveedu K., Xu P., Fernandes R., Mayer R.J. (2007) A content based video retrieval method for surveillance and forensic applications. In Proceedings of SPIE, volume 6560, page 656004.

Cimiano P., Reyle U., Saric J. (2005) Ontology driven discourse analysis for information extraction. Data and Knowledge Engineering Journal, 55:59–83.

Hartholt A., Russ T., Traum D., Hovy E., Robinson S. (2008) A common ground for virtual humans: using an ontology in a natural language oriented virtual human architecture. In Language Resources and Evaluation Conference (LREC).