Semantic Annotation of Complex Human Scenes for Multimedia Surveillance

Carles Fernández[1], Pau Baiget[1], Xavier Roca[1], and Jordi Gonzàlez[2]

[1] Computer Vision Centre, Edifici O. Campus UAB, 08193, Bellaterra, Spain
[2] Institut de Robòtica i Informàtica Ind. UPC, 08028, Barcelona, Spain
{perno|pbaiget|xroca|poal}@cvc.uab.es

Abstract. A Multimedia Surveillance System (MSS) is considered for automatically retrieving semantic content from complex outdoor scenes, involving both human behavior and traffic domains. To characterize the dynamic information attached to detected objects, we consider a deterministic modeling of spatio-temporal features based on abstraction processes towards fuzzy logic formalism. A situational analysis over conceptualized information will not only allow us to describe human actions within a scene, but also to suggest possible interpretations of the behaviors perceived, such as situations involving thefts or dangers of running over. Towards this end, the different levels of semantic knowledge implied throughout the process are also classified into a proposed taxonomy.

1 Introduction

The idea of *Multimedia Surveillance Systems* (MSS) is growing importance in various application fields concerning the automatic extraction and management of contents from large databases of media streams [5]. The term *multimedia* is related to systems which provide human end-users for accessing to or communicating with applications dealing with certain type of multimedia content. Applications for sports and surveillance domains, on the other hand, generate continuous streams of abundant data from real-time monitoring over controlled environments, thus requiring proper techniques for the evaluation of video and audio sequences and efficient storing of their significant content. An effective management of these documents depends in great measure on the availability of indexes, as stated in [12], and since manual techniques are not feasible for large audiovisual collections, systems for automatic management of such databases in a real-time context are very valuable in order to facilitate real-time reactivity.

The detection, understanding, and description of human motion patterns in video sequences has been a subject of increased interest over the last years. Motion analysis techniques are used to address the semantic classification of visually perceived information. Motion detection is required for the solution of many other complex tasks, such as image sequence analysis and understanding. Moeslund et al. have extensively reviewed recent advances in human motion capture and analysis in [8]. The extraction of content from image sequences is currently becoming a central focus of attention for many researchers, too [6].

Most efforts in semantic annotation and video indexing have been devoted so far to classification techniques for particular, definite domains, such as sports, movies, or news. Assfalg et al. have discussed approaches for the retrieval of visual information and semantic annotation of its contents [2]. They encourage the use of low-level visual primitives to capture semantic content at a higher level of significance, in specific domains such as news or sports. Also, Smeulders et al. discussed the content extraction and analysis in image retrieval systems [11].

Here, we focus on the automatic, real-time extraction of semantic content from video recordings obtained in outdoor surveillance environments. We aim to evaluate complex behaviors involving both humans and vehicles, detecting the significant events developed, and providing semantic-oriented outputs which are more easily handleable for indexing, search, and retrieval purposes, such as annotated video streams, small sets of content-expressing images, or summaries of occurrences. An important objective is also the temporal characterization of the structure of a recording from this semantic perspective. Although a single video modality is implemented so far, the designed outline pursues the eventual integration of knowledge sources having intrinsically different natures. For this purpose, a taxonomy has been conceived for ensuring a future effective collaboration among knowledge derived from agent, body, and face motion.

This paper is organized as follows: first, our proposal of taxonomy for classifying the different types of knowledge involved in MSS is presented in Section 2. In Section 3, the automatic acquisition of visual features is introduced, which provides structural information over time. Section 4 discusses how a subsequent abstraction step converts this knowledge into a logic-deductive form, thus expressing it by means of semantic predicates. Section 5 presents the high-level analysis of the contextualized situations to generate a set of thematic descriptors. Experimental results in Section 6 describe the main relevant happenings in a particular outdoor scene, resulting in the division of the temporal structure of the recording into content-based intervals. Finally, Section 7 concludes the paper and presents future lines of work.

2 Human Motion Taxonomy

When considering systems for the annotation of content, it becomes laborious –and sometimes uncertain– to accurately classify the concepts related to the field of evaluation, and to establish proper relationships among them. The reason for such a statement comes from thinking of the great generality of the situations to be considered. Upon this, a question arises: which kind of information needs to be represented for properly evaluating human behavior?

Although a specific problem domain has been chosen in this paper, a MSS needs to cope with a wide range of occurrences. The temporal dimension of a video sequence determines the nature of the semantic content to be extracted, thus suggesting that the use of mechanisms for a proper characterization should be based on motion detection. Additionally, the most difficult issues to solve are obviously related to human behaviors, so the proposed classification scheme is

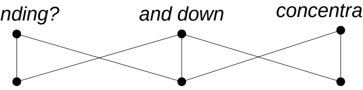
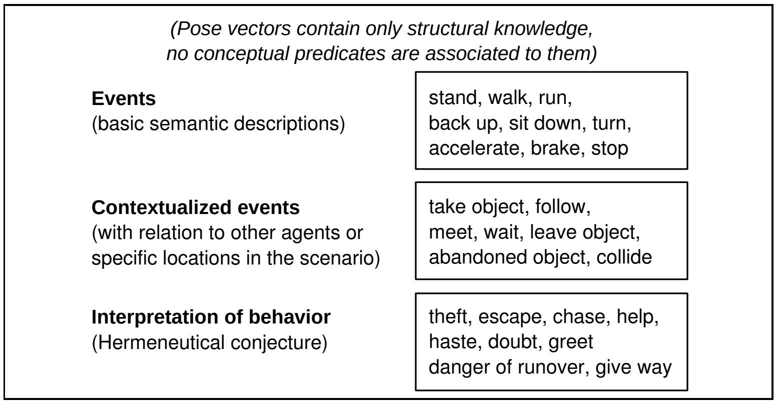
		Examples for:		
	Description	**Agent**	**Body**	**Face**
Pose Vector	Array of Static Configurations over time	trajectory of locations	sequence of postures, gestures	sequence of expressions
Event	Dynamic interpretaton of static elements	*stopped: sitting? / standing?*	*stand, move head up and down*	*serious expr.: worry? / concentration?*
				
Contextualized Event	Disambiguation using information from other levels	*standing in front of a scene object*	*nod? / search for something?*	*attention (staring at scene object)*
Interpreted Behavior	Hypothesis of interpretation using complete scene knowledge	*"A person is searching for information on the timetables of the subway station"*		

Table 1. A knowledge-based classification of human motion representations. This *human motion taxonomy*, which includes several levels of representation, will guide the process of interpretation.

conceived in terms of human motion. Literature already contains some suggested classification criteria based on different models of taxonomy. The current taxonomy enhances the proposal of Gonzàlez [6], which in turn unifies the taxonomies of Bobick [4] and Nagel [9]. Our classification proposes four hierarchical levels of semantics, see Table 1:

- First of all, *Status Vectors* are collections of detected static configurations for the tracked elements. This includes positions, orientations or velocities for the agents for each time-step, and other spatial information over time such as postures, gestures, or expressions, but no interpretation at all. Semantics is present in form of structural information.
- The analysis of these collections over time allow to detect *Events*, i.e. basic dynamic interpretations of a predefined set of static configurations for a local environment. Patterns which are shared in common with a sequence of temporal status are identified and classified according to predefined models. Some examples include to detect that a pedestrian is running or turning, and that a car is accelerating or braking.
- Next level is suggested to be a *Contextualized Event*, this is, a concretion or disambiguation of the possible interpretations for an event. An event will be analyzed with respect to other detected elements within the scene. The interrelation among results from different tracking domains plays an important role in the process of contextualizing an event, e.g. ‘sit down’–‘bus stop’, ‘wave hand’–‘open mouth’, or ‘tired expression’–‘climb stairs’.



(Pose vectors contain only structural knowledge, no conceptual predicates are associated to them)

Events (basic semantic descriptions)	stand, walk, run, back up, sit down, turn, accelerate, brake, stop
Contextualized events (with relation to other agents or specific locations in the scenario)	take object, follow, meet, wait, leave object, abandoned object, collide
Interpretation of behavior (Hermeneutical conjecture)	theft, escape, chase, help, haste, doubt, greet danger of runover, give way

Table 2. Classification of some possible semantic descriptions using the proposed taxonomy. The ontology of terms is formed by a set of conceptual predicates, which are associated to agents (human or vehicular), objects, and events related to the detected occurrences within the scene. Higher semantic levels also incorporate a higher level of uncertainty for the inferences.

- Finally, a conjecture for global understanding of a scene is suggested by an hypothesis of *Interpretation* for the detected behaviors. All the information sources are considered at this point, including some domain-dependent bases of knowledge from the world. Interpretations include an important part of uncertainty, thus giving an impression of subjectiveness, which is founded on visual evidence and given models. This last level constitutes an actual process of *video hermeneutics* over the scene.

Table 2 applies this taxonomy to classify some of the targeted occurrences to be detected. In order to infer a complete interpretation of the scene, both sensory and semantic gaps need to be bridged. The last level of 'guided' uncertainty is included into the taxonomy towards this goal. It can be seen that this proposal takes into account all the considered levels of extraction of visual information –i.e. agent, body, face, and relation with other detected objects, agents, and events–, and also suggests a proper way of managing the different stages of knowledge. This categorization takes into account the relevance of the retrieved information, some hierarchical degrees of perspective, and also the level of subjectiveness required for a scene interpretation.

3 Extraction of Visual Information

There are many different tasks involved in the processes of acquisition of visual information from the video recordings. In the first place, a process of segmentation locates the significant information within the image data and considers it as a part of the foreground region by means of a background model. Next,

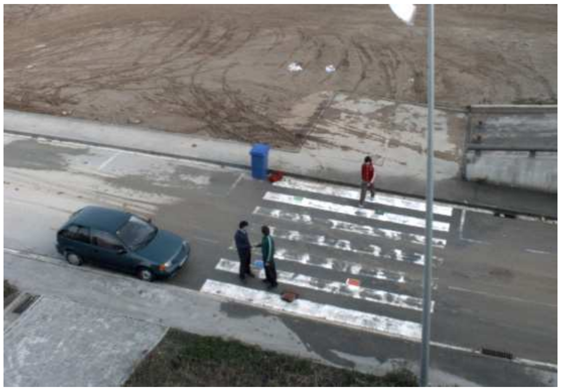


Fig. 1. The Theft Scene.

tracking processes maintain the identification of the targets and recover from possible segmentation errors, principally derived from occlusions and illumination conditions. In this paper we use the segmentation and tracking processes presented in [3], which are far from the scope of this paper. As a result, information about the geometrical position of the agent in ground-plane coordinates at each time-step, and also his instantaneous velocity and orientation are achieved. This knowledge is enclosed in the so-called *Agent Status Vectors*.

Experimental results have been focused to be specialized to a single type of scenario in order to study the problems in-depth, rather attempting to come up with a supposedly generally applicable solution. A particular scene has been considered, which contains complex situations resulting from the interaction of pedestrians and vehicles in an outdoor environment. It consists of a crosswalk scene, in which some agents (pedestrians and cars) and objects appear and interact. This scene includes several behaviors by the agents, e.g. displacements, meetings, crossings, accelerations, object disposals, and more complex situations such as an abandoned object, a danger of running over, and a theft. The recording has been obtained using a distributed system of static cameras, and the scenario has been modeled a priori, see Fig. 1.

An algorithm has been applied for automatic segmentation of the scenario. The trajectories of the agents are sampled, interpolated, and finally a set of segments are defined for the scenario. More details about this process can be seen in [3]. Once the scenario has been automatically segmented in geometrical terms, the resulting locations need to be categorized according to their semantic properties. This way, different regions are enclosed into separate categories like *sideway_segment*, *road_segment*, *crosswalk*, *exit*, or *waiting_zone*, depending on which behaviors can be expected for an agent at the considered location. This step will enable further contextualization for the events of an agent. This process of categorization is done in a supervised manner at present.

4 Abstraction and Conceptual Manipulation

The acquisition of visual information produces an extensive amount of geometric data, considering that computer vision algorithms are applied continuously over the recordings. Such a large collection of results turns out to be increasingly difficult to handle. Thus, a process of abstraction is needed in order to extract and manage the relevant knowledge derived from the tracking processes. The question arises how these spatiotemporal developments should be represented in terms of significance, also allowing further semantic interpretations. Several requirements have to be accomplished towards this end [7]:

1. Generally, the detected scene developments are only valid for a certain time interval: the produced statements must be updated and time-delimited.
2. There is an intrinsic *uncertainty* derived from the estimation of quantities in image sequences (i.e. the sensory gap), due to the stochastic properties of the input signal, artifacts during the acquisition processes, undetected events from the scene, or false detections.
3. An abstraction step is necessary to obtain a formal representation of the visual information retrieved from the scene.
4. This representation has to allow different domains of human knowledge, e.g. analysis of human or vehicular agents, posture recognition, or expression analysis, for an eventual semantic interpretation.

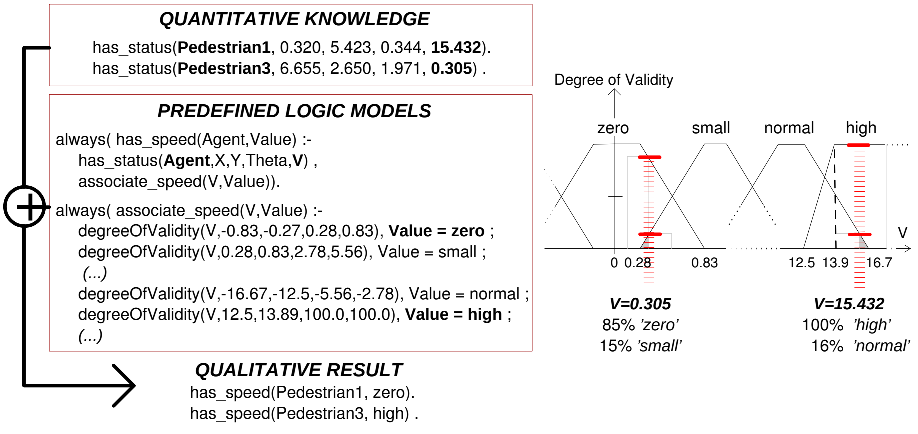
Fuzzy Metric Temporal *Horn* Logic (FMTHL) has been conceived as a suitable mechanism to solve each of the aforementioned demands [10]. It is a rule-based inference engine in which conventional logic formalisms are extended by a temporal and a fuzzy component. This last one enables to cope with uncertain or partial information, by allowing variables to have degrees of truth or falsehood. The temporal component permits to represent and reason about propositions qualified in terms of time. These propositions are represented by means of *conceptual predicates*, whose validity is evaluated at each time-step.

All sources of knowledge are translated into this logic predicate formalism for the subsequent reasoning and inference stages. In the first place, agent status vectors are converted into `has_status` conceptual predicates:

$$t \quad ! \quad \texttt{has\_status (Agent, X, Y, Theta, V)} \tag{1}$$

These predicates hold information for a global identification of the agent ($Agent$), his spatial location in a ground-plane representation of the scenario (X, Y), and his instantaneous orientation ($Theta$) and velocity (V). A *has_status* predicate is generated at each time-step for each detected agent. In addition, certain predicates are generated for identifying the category of the agent, e.g. `pedestrian(Agent)` or `vehicle(Agent)`. Similarly, the segmented regions from the scenario are also converted into logic descriptors holding spatial characteristics, and the semantic categories are also included:

```
point (14, 5, p42)
```



```
QUANTITATIVE KNOWLEDGE
    has_status(Pedestrian1, 0.320, 5.423, 0.344, 15.432).
    has_status(Pedestrian3, 6.655, 2.650, 1.971, 0.305) .

PREDEFINED LOGIC MODELS
always( has_speed(Agent,Value) :-
    has_status(Agent,X,Y,Theta,V) ,
    associate_speed(V,Value)).
always( associate_speed(V,Value) :-
    degreeOfValidity(V,-0.83,-0.27,0.28,0.83), Value = zero ;
    degreeOfValidity(V,0.28,0.83,2.78,5.56), Value = small ;
     (...)
    degreeOfValidity(V,-16.67,-12.5,-5.56,-2.78), Value = normal ;
    degreeOfValidity(V,12.5,13.89,100.0,100.0), Value = high ;
    (...)

QUALITATIVE RESULT
    has_speed(Pedestrian1, zero).
    has_speed(Pedestrian3, high) .
```

(a) **(b)**

Fig. 2. Conversion from quantitative to qualitative values. (a) The input status vectors contain information for the agents. As a result, qualitative descriptions are represented logically. (b) FMTHL includes fuzzy mechanisms accepting more than one single interpretation, since it confers *degrees of validity* to values on uncertain ranges.

$$\begin{array}{c}\texttt{line (p42, p43, l42)}\\ \texttt{segment (l31, l42, lseg\_31)}\\ \texttt{crosswalk\_segment (lseg\_31)}\end{array} \tag{2}$$

The abstraction process is thus applied over the information obtained both from the scenario and from the agents, i.e. the categorized segments from the considered location and the agent status vectors generated. Quantitative values are converted into qualitative descriptions in form of conceptual predicates, by adding fuzzy semantic parameters such as *close*, *far*, *high*, *small*, *left*, or *right*. The addition of fuzzy degrees allows to deal with the uncertainty associated to visual acquisition processes, also stating the goodness of the conceptualization. Fig. 2 gives an example for the evaluation of a `has_speed` predicate from an asserted `has_status` fact. The conversion from quantitative to qualitative knowledge is accomplished by incorporating domain-related models to the reasoning system. Hence, new inferences can be performed over an instantaneous collection of conceptual facts, enabling the derivation of logical conclusions from the assumed evidence. Higher-level inferences progressively incorporate more contextual information, i.e. relations with other detected entities in the scenario. This spatiotemporal universe of basic conceptual relations supplies the dynamic interpretations which are necessary for detecting *events* within the scene, as described in the taxonomy.

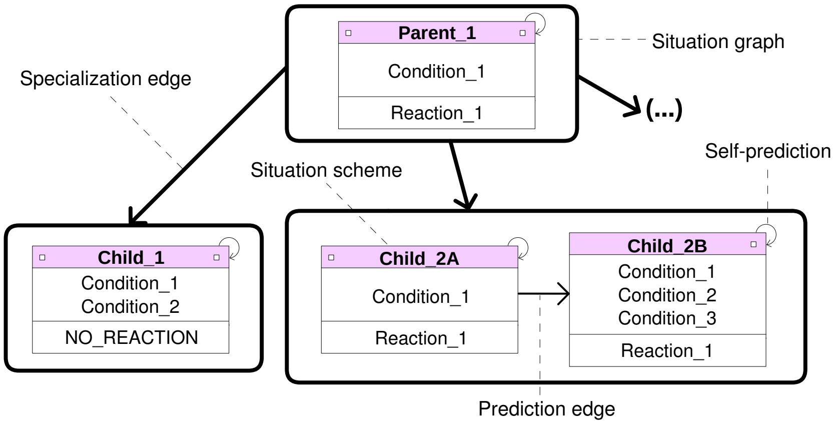


Fig. 3. General example of a Situation Graph Tree. Situation graphs are represented as rounded rectangles, situation schemes as normal ones. When the conditions of the parent situation scheme are asserted, the situation specializes in one of the possible situation graphs. Small rectangles to the left or to the right of the name of a situation scheme mark that scheme as a start- or end-situation.

5 Behavioral Analysis

An independent stage is implemented for achieving effective modeling of behaviors. The concurrence of hundreds of conceptual predicates makes necessary to think of a separate module for dealing with new semantic properties at a higher level: some guidelines are necessary to establish relations of cause, effect, precedence, grouping, interaction, and in general any reasoning performed with time-constrained information at multiple levels of analysis, i.e. the contextualization and interpretation levels as proposed in the taxonomy.

The tool which has been chosen to enable behavior modeling is the Situation Graph Tree (SGT), see [1, 6]. SGTs are hierarchical classification trees used to describe the behavior of the agents in terms of situations they can be in. These trees are based on deterministic models which explicitly represent and combine the specialization, temporal, and semantic relationships of the conceptual facts which have been asserted. A general example of a SGT is shown in Fig. 3.

The semantic knowledge related to any agent at a given point of time is contained in a *situation scheme*, which constitutes the basic component of a SGT. A situation scheme can be seen as a semantic function which evaluates an input consisting of a set of conditions –the so-called *state predicates*–, and generates logic outputs at a higher level –the *action predicates*– once all the conditions are asserted. Here, the action predicate is a `note` method which generates a semantic annotation in a language-oriented form, containing fields related to thematic roles such as *Agent*, *Patient*, *Object* or *Location*. Situations schemes representing different temporal episodes of the same situation are enclosed by situation graphs. The evolution over time is indicated by means of *prediction edges*, either

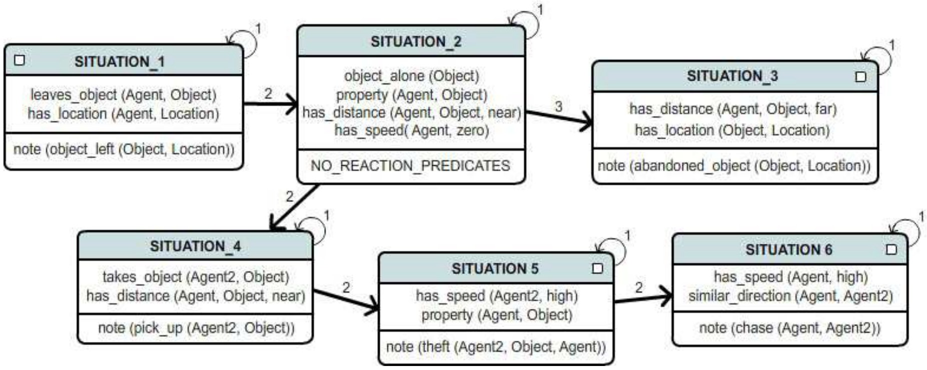

SITUATION_1
leaves_object (Agent, Object)
has_location (Agent, Location)
note (object_left (Object, Location))
SITUATION_2
object_alone (Object)
property (Agent, Object)
has_distance (Agent, Object, near)
has_speed(Agent, zero)
NO_REACTION_PREDICATES
SITUATION_3
has_distance (Agent, Object, far)
has_location (Object, Location)
note (abandoned_object (Object, Location))
SITUATION_4
takes_object (Agent2, Object)
has_distance (Agent, Object, near)
note (pick_up (Agent2, Object))
SITUATION 5
has_speed (Agent2, high)
property (Agent, Object)
note (theft (Agent2, Object, Agent))
SITUATION 6
has_speed (Agent, high)
similar_direction (Agent, Agent2)
note (chase (Agent, Agent2))
1
2
3

Fig. 4. Part of the SGT used for behavioral analysis of the Theft Scene. This situation graph is evaluated when the system detects that an object has been left by the pedestrian who owns it. The set of conditions are FMTHL predicates, the reaction predicate is a `note` command which generates a semantic tag.

between different situations or within a persistent state (self-prediction). On the other hand, SGTs also allow for conceptual or temporal particularizations of general situations, using *specialization edges*. Tree structures appear as a result of these specializations. A SGT has been designed for the Theft Scene: Fig. 4 shows a piece of it, which illustrates how the behavioral analysis is performed in order to identify situations such as an abandoned object or a theft. The whole SGT is transformed into FMTHL for automatic exploitation of its behavior schemes.

By selecting the conditions to incorporate into the set of state predicates, we both integrate asserted facts obtained from different sources and also establish certain attentional factors over the whole universe of occurrences. Thus, this first step accomplishes the *contextualization* stage included in the taxonomy. On the other hand, the generation of `note` action predicates can be understood as the *interpretation* of a line of behaviors, for a concrete domain and towards a concrete goal.

The results obtained from the behavioral level, i.e. the annotations generated by the situational analysis of an agent, can yet be considered as outputs of a process for *content detection*. From this point of view, a SGT would be the classified collection of all possible domain-related semantic tags to be assigned to a video sequence. In addition, the temporal segmentation of video is also achieved: since the semantic tags are temporally valid, a video sequence can be split in the time-intervals which define these tags. As a result, each video segment is associated to individual and cohesive conceptual information.

6 Experimental Results

Tests have been performed over recordings showing different behaviors. Fig. 5 gathers the collection of semantic annotations automatically generated for the *Theft Scene*. Experimental results show semantic shots generated at certain points of time, which allows to perform content-based time segmentation over the whole video sequence, by identifying remarkable occurrences and relating them to specific spaces of time[3]. The video sequence is thus split into significant time intervals, tagged with conceptual annotations, so the relevant content of a recording can be expressed as a collection of individual images and tags, such as the '*pick up*', '*theft*', and '*chase*' interpreted behaviors.

Since the system operates differently depending on the concrete needs of a certain user, the level of detail for the descriptions can be established, concerning the amount of semantic shots generated. This adjustment is achieved by confronting a more extensive number of spatiotemporal descriptions with a more reduced number of logical inferences and interpretations (i.e. higher semantics involving greater uncertainty). Thus, subsequent search and retrieval algorithms can operate at different levels.

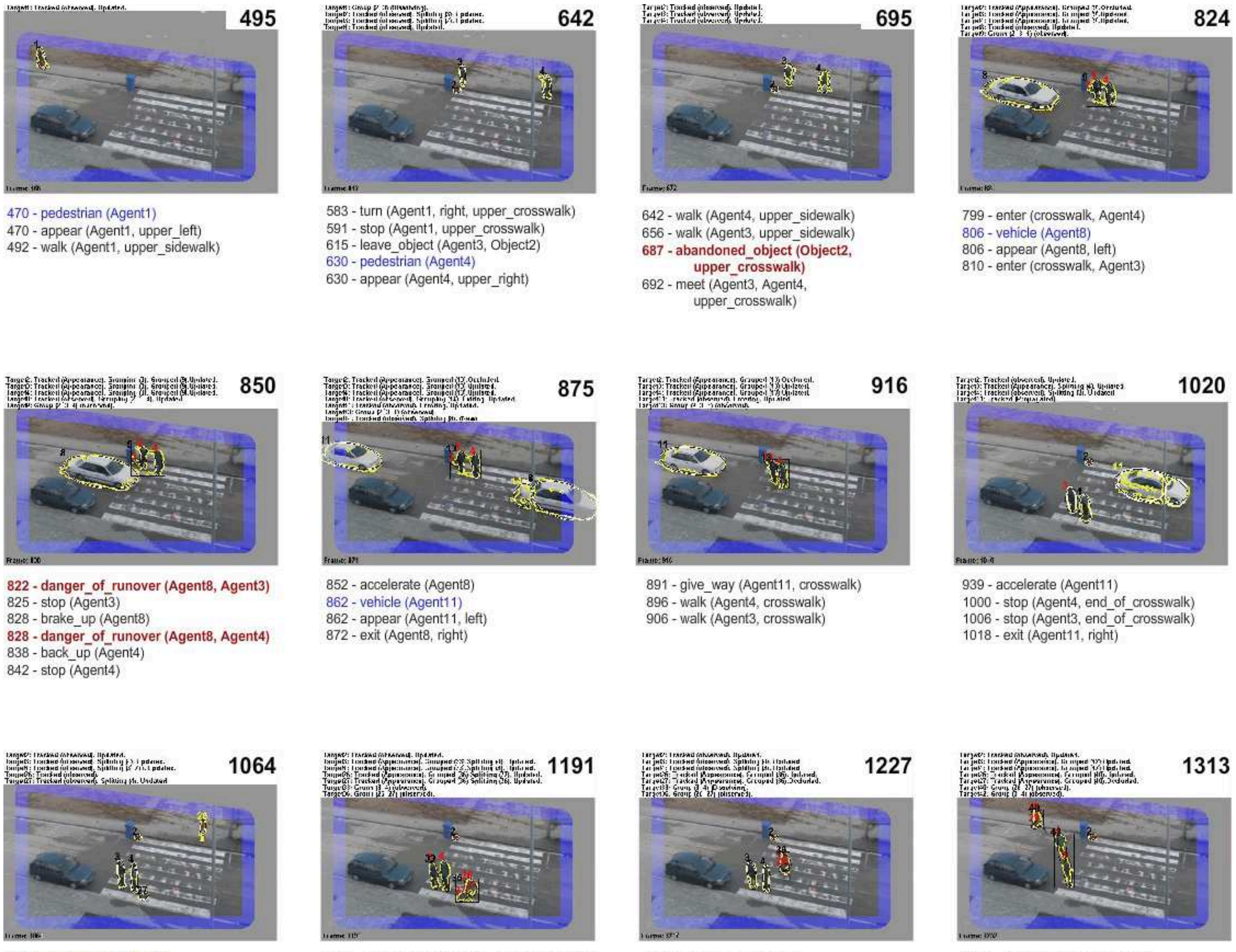
7 Conclusions and Future Work

The system presented in this paper (i) provides distilled video sequences from the monitored area, (ii) partitions these sequences into definite time-intervals depending on the associated content, and (iii) attaches semantic annotations in a real-time environment. It has allowed to automatically infer high-level reasonings and interpretations from geometrical results extracted by vision processes, and to generate semantic annotations abstracting these inferences. It identifies the targeted behaviors, by analyzing interactions among the different agents and objects involved in a scene. It also makes possible to ring alarms in the presence of certain defined complex situations, e.g. abandoned objects in the scenario, danger of imminent running over pedestrians or collision between vehicles, and detection of possible thefts. Since the conceptual stages of the system are deterministic, the accuracy of the results at these levels depend on the contextual and subjective models used to interpret occurrences. In those situations in which human agreement over the interpretation of events is low, no inferences are made.

New domains (e.g. sports) will be included for evaluation. The system needs to be extended regarding its multimedia capabilities. Further steps include processing of audio streams to detect sound and speech, and the evolution towards a multitracking system, which performs not only agent tracking, but also face expression and body analysis; towards this end, a set of PTZ (Pan, Tilt, Zoom) active cameras will be incorporated into the distributed camera system. All these steps can be naturally included into the proposed taxonomy, since it focus the interaction among different sources of knowledge. Only data conceptualizations and model extensions need to be considered.

Some tasks need to be improved regarding a better and more general performance: first, prior knowledge for the concrete scenario needs to be reduced as much as possible, to facilitate the characterization of content without excessive constraints. Refining the described stage for automatic modeling of common

[3] The complete video sequence can be seen at `http://www.cvc.uab.es/ise`



470 - pedestrian (Agent1)
470 - appear (Agent1, upper_left)
492 - walk (Agent1, upper_sidewalk)

583 - turn (Agent1, right, upper_crosswalk)
591 - stop (Agent1, upper_crosswalk)
615 - leave_object (Agent3, Object2)
630 - pedestrian (Agent4)
630 - appear (Agent4, upper_right)

642 - walk (Agent4, upper_sidewalk)
656 - walk (Agent3, upper_sidewalk)
687 - abandoned_object (Object2, upper_crosswalk)
692 - meet (Agent3, Agent4, upper_crosswalk)

799 - enter (crosswalk, Agent4)
806 - vehicle (Agent8)
806 - appear (Agent8, left)
810 - enter (crosswalk, Agent3)

822 - danger_of_runover (Agent8, Agent3)
825 - stop (Agent3)
828 - brake_up (Agent8)
828 - danger_of_runover (Agent8, Agent4)
838 - back_up (Agent4)
842 - stop (Agent4)

852 - accelerate (Agent8)
862 - vehicle (Agent11)
862 - appear (Agent11, left)
872 - exit (Agent8, right)

891 - give_way (Agent11, crosswalk)
896 - walk (Agent4, crosswalk)
906 - walk (Agent3, crosswalk)

939 - accelerate (Agent11)
1000 - stop (Agent4, end_of_crosswalk)
1006 - stop (Agent3, end_of_crosswalk)
1018 - exit (Agent11, right)

1033 - pedestrian (Agent26)
1033 - appear (Agent26, upper right)
1049 - walk (Agent26, upper sidewalk)

1054 - object_left (Object27, lower_crosswalk)
1078 - turn (Agent26, left, upper_crosswalk)
1093 - enter (crosswalk, Agent26)
1168 - turn (Agent26, right, lower_crosswalk)
1186 - pick_up (Agent26, Object27)

1211 - run (Agent26, road)
1220 - theft (Agent26, object27, Agent4)

1241 - chase (Agent4, Agent26)
1276 - exit (Agent26, upper_left)
1278 - run (Agent4, road)
1294 - exit (Agent4, upper_left)

Fig. 5. Set of semantic annotations produced for the theft scene, which have been automatically generated for the fragment of recording comprised between frames 450 and 1301. Some captures showing the results after tracking processes have been provided, too, for illustration purposes. The number of frame appears in front of each produced annotation, and also in the upper-right corner of each capture. Detections of new agents within the scene have been marked in blue, annotations for activating predefined alerts have been emphasized in red.

paths may be useful for this matter. In addition, the contextualization of events needs to be strengthened to facilitate knowledge acquisition by means of object relationships within the scene. Considering more complex behavioral patterns will include *group* and *crowd* detection. The impending addition of face and posture trackers will contribute a big leap towards this end, too.

Acknowledgements

This work has been supported by EC grant IST-027110 for the HERMES project and by the Spanish MEC under projects TIC-2003-08865 and DPI-2004-5414. Jordi Gonzàlez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

1. M. Arens and H.-H. Nagel. Behavioral Knowledge Representation for the Understanding and Creation of Video Sequences. In *KI 2003: Advances in AI*, volume 2821 of *LNAI*, pages 149–163. Springer–Verlag, Berlin, Heidelberg, New York, 2003.
2. J. Assfalg, M. Bertini, C. Colombo, and A.D. Bimbo. Semantic Annotation of Sports Videos. *Multimedia, IEEE*, 9(2):52–60, 2002.
3. P. Baiget, C. Fernández, X. Roca, and J. Gonzàlez. Automatic Learning of Conceptual Knowledge for the Interpretation of Human Behavior in Video Sequences. In *3rd Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2007)*. Springer LNCS, 2007.
4. A.F. Bobick. Movement, Activity and Action: the Role of Knowledge in the Perception of Motion. *Philosophical Transactions: Biological Sciences*, 352(1358):1257–1265, 1997.
5. R. Cucchiara. Multimedia Surveillance Systems. *Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, pages 3–10, 2005.
6. J. Gonzàlez. *Human Sequence Evaluation: The Key-Frame Approach.* PhD thesis, Universitat Autònoma de Barcelona, Barcelona, Spain, 2004.
7. M. Haag, W. Theilmann, K. Schäfer, and H.-H. Nagel. Integration of Image Sequence Evaluation and Fuzzy Metric Temporal Logic Programming. pages 301–312. Springer-Verlag London, UK, 1997.
8. T.B. Moeslund, A. Hilton, and V. Kruger. A Survey of Advances in Vision-based Human Motion Capture and Analysis. *Computer Vision and Image Understanding*, 104:90–126, November-December 2006.
9. H.-H. Nagel. From Image Sequences Towards Conceptual Descriptions. *Image and Vision Computing*, 6(2):59–74, 1988.
10. K. Schäfer and C. Brzoska. F-Limette Fuzzy Logic Programming Integrating Metric Temporal Extensions. *Journal of Symbolic Computation*, 22(5-6):725–727, 1996.
11. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based Image Retrieval at the End of the Early Years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, 2000.
12. C.G.M. Snoek and M. Worring. A Review on Multimodal Video Indexing. *Multimedia and Expo, 2002. Proceedings of the ICME'02.*, 2, 2002.