Research Steps towards Human Sequence Evaluation

Jordi Gonzàlez [†], F. Xavier Roca [*], Juan J. Villanueva [*]

[†] Institut de Robòtica i Informàtica Industrial (UPC-CSIC), Edifici U, Parc Tecnològic de Barcelona, Barcelona 08028, Catalonia, Spain.

[*] Computer Vision Center, Edifici O, Campus UAB, Bellaterra 08193, Catalonia, Spain

Abstract *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.

1 Introduction

Hermeneutics, according to Wilhelm Dilthey, is the art of interpretation of hidden meanings. The name comes from HERMES, the God known as the messenger of the intentions of the Gods to the human beings. In particular, interpretation in cultural sciences requires to *know* its object, a human being, from the inside. That means, we can infer the intentions of a

person because we also are persons. Towards this end, the HERMES project will address basic methods for the extraction, description and animation of human motion in the same scenario (indoor or outdoor), and new methods for the interpretation of dynamic scenes.

The design and implementation of such a cognitive system still constitutes a challenge, even if the discourse domain will be drastically constrained within which it is expected to operate. An algorithmic system with analogous capabilities can be considered an instantiation of a 'cognitive system'. In particular, the term Human Sequence Evaluation (HSE) denotes the design, implementation and evaluation of such a cognitive system (Gonzàlez 2004). In general terms, we proposed to develop towards weakly embodied cognition within a system for understanding an environment containing autonomous agents. By understanding, we mean that the system must move beyond merely describing the scene: in addition it must be able to reason about the scene and give suitable explanations for various events and behaviours.

Thus, the generation of semantic descriptions conveys the meaning of motion, i.e. *where, when, what*, *how* and also *why* the motion is being detected. As a result, this high-level understanding provide a richer, broader and even more challenging domains of research, which will encompass not only research in Computer Vision, but also in Pattern Recognition, Artificial Intelligence and Computer Animation, to cite few.

At present, few video surveillance systems exploits all these aspects of cognition: in the HERMES project, we restrict cognition to assure HSE, that means, on the one hand, to develop transformation processes to perform human motion understanding and, on the other hand, to convey inferred interpretations to human operators by means of natural language texts or synthesized agents in virtual environments.

This paper presents how HSE considers the interpretation of human motion as a transformation process between raw video signals and high-level, qualitative descriptions. At least, this process will involve (i) the extraction of relevant visual information from a video sequence, (ii) the representation of that information in a suitable form, and (iii) the interpretation of visual information for the purpose of recognition and learning about human behaviour.

2 State of the Art

During the past three decades, important efforts in Computer Vision research have been focused on developing theories, methods and systems applied to video sequences (Moeslund and Granum 2001). Broadly speaking, research is focused on describing *where and when* motion is being detected by camera sensors. For this purpose, the goal is set to describe motion using quantitative values, such as the spatial position of a given agent over time, for example.

Suitable discourse domains are, e.g., well-frequented streets, pedestrian-crossings, bus-stops, reception desks of public buildings, railway platforms. This demand in surveillance systems is due to the huge amount of video which should be selected, watched, and analyzed by a small number of operators in real time. Current textual descriptions generated automatically from surveillance sequences helps to detect abnormal and dangerous situations on-line. As a long-term result of HSE, surveillance systems will not only recognize and describe, but also *predict* abnormal or dangerous behaviours on-line, instead of merely record video sequences

The basis of current research in any of the aforementioned domains is the detection of agents within the scene. Two different approaches are found in the literature, namely, *background modeling/substraction* and *motion detection*. The former necessitates implementing a suitable background model of the scene to determine foreground regions. Most referred publications use a background modeling-based approach (Haritaoglu et al. 2000; Stauffer et al. 2000; Li et al. 2004). On the other hand, motion detection computes motion information from consecutive frames. Consequently, an action can be described in terms of a proper motion characterization (Lipton et al. 1998; Ricquebourg and Bouthemy 2000; Masoud and Papanikolopoulos 2003).

Additionally, *tracking* procedures are usually incorporated in order to reduce segmentation errors (Sanfeliu and Villanueva 2005) . In recent years, new tracking techniques are defined based on a hypothesis/validation principle (Comaniciu et al. 2003). Thus, the tracking process is modeled using a probabilistic scheme, which is based on the Bayes' rule (Isard and Blake 1998; Sidenbladh et al. 2002; Nummiaro et al. 2003, Bullock and Zelek 2004).

Tracking techniques should embed knowledge about the human agent, such as its observed motion, appearance, or shape. This knowledge can be based on *image features* or *predefined body models*. On the one hand, the spatial information of the agent state in video surveillance systems is often represented using simple image features, such as points, lines, or regions.

Most popular representations are blobs (Li et al. 1998,) or blob attributes, such as the centroid, median or bounding box. On the other hand, model-driven approaches incorporate known physical constraints of limbs and extremities of the body to help both localisation and tracking. By providing a synthetic body model, anatomical information and kinematic constraints are incorporated into the action model, thus allowing tracking of limbs, synthesis of motion, and performance analysis. Most referred models are those based on stick figures (Dockstader et al. 2003; Karaulova et al 2002; Deutscher and Reid 2005), 2-D contours (Yamada et al. 1998; Wagg and Nixon 2004) and volumetric models (Gavrila and Davis 1996; Ben Aire et al. 2002; Ning et al. 2004).

Once the body model is properly tracked over time, it is possible to recognize predefined motion patterns and to produce high-level descriptions. In fact, the basis of motion understanding is *action recognition*. In order to deal with the inherent temporal and spatial variability of human performances, suitable analytical methods have been used in the literature for matching time-varying data. Most referred algorithms are Dynamic Time Warping (DTW), Hidden Markov Models (HMM) and Neural Networks (NN) (Galata et al. 2001, Wang et al. 2003).

Subsequently, human motion information is then combined with the known information about the environment in order to derive complex semantic descriptions (Gonzàlez 2004). From a semantic perspective, conceptual predicates extracted from video sequences are classified according to different criteria, such as *specialization relationship* (Karaulova et al. 2002), *semantic nature* (Remagnino et al. 1998) or *temporal ordering* (Intille and Bobick 2001). Likewise, suitable behaviour models explicitly represent and combine the specialization, semantic and temporal relationships of their constituent semantic predicates (Nagel 1988). For this purpose, semantic primitives involved in a particular behaviour are organized into hierarchical structures, such as networks (Sagerer and Niemann 1997) or trees (Wachter and Nagel 1999; Kojima et al. 2002) which allow motion understanding.

On the one hand, semantic interpretation is still mostly restricted to express the *relationships* of an agent with respect to its environment. However, the *internal state* of the agent has traditionally received little (or none) attention in human motion understanding. But human agents have inner states (based on emotions, personality, feelings, goals and beliefs) which may determine and modify the execution of their movement. These inner states are hard to be derived from a single picture. Instead, we need image sequences to evaluate emotions, like *sad*, *happy* or *angry*, in a temporal context.

Emotion descriptions will require high-detailed images which will be obtained by means of active cameras. In fact, camera's zoom are controlled to supply imagery at the appropriate resolution for motion analysis of the human face, thus facilitating emotion analysis (Cohen et al. 2003, Zhang et al. 2001). Current state-of-the-art is mainly concerned with posed facial expression recognition. In the proposed scenario, we would encounter spontaneous expressions that are considerably more difficult to handle. Only few publications can be found on spontaneous facial expression recognition and are mostly limited to very specific facial motions such as eye blinking.

On the other hand, semantic interpretation also leads to *uncertainty*, due to the vagueness of the semantic concepts utilized, and the incompleteness, errors and noise in the agent state's parameters (Ma and McKevitt 2004). In order to cope with the uncertainty aspects, integration and fusion methods can be learnt using a probabilistic framework: PCA, Mixtures of Gaussians (MoG) (Morris and Hogg 2000, Fod et al. 2002) and Belief Networks (BN) (Intille and Bobick 2001, Remagnino et al. 1998) provide examples. Alternatively, Fuzzy Metric Temporal Logic (FMTL) can also cope with the temporal and uncertainty aspects of integration in a goal-oriented manner (Schäfer 1997).

3 Approach to Research

The main objective of HSE is to develop a cognitive artificial system based in a framework model which allows both recognition and description of a particular set of human behaviors arising from real-world events. Specifically, we propose to model the knowledge about the environment in order to make or suggest interpretations from motion events, and to communicate with people using natural language texts, audio or synthetic

films. These events will be detected in image data-streams obtained from arrays of multiple active cameras (including zoom, pan and tilt).

HSE thus aims to design a Cognitive Vision System for human motion and behaviour understanding, followed by communication of the system results to end-users, based on two main goals. We assume that three different types of descriptions can be obtained, which depend on the resolution of the acquisition process: facial expressions, body postures and agent trajectories, where each topic demands its own specific requirements and computational models for a proper representation.

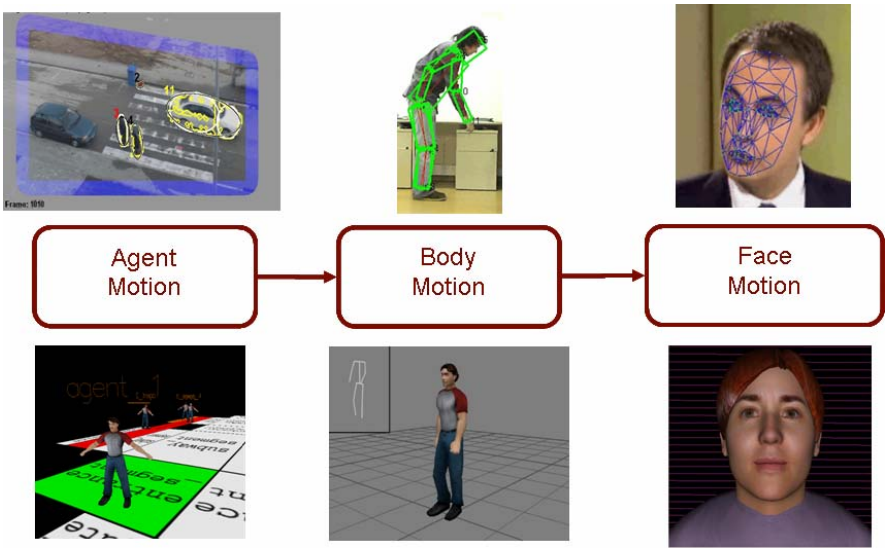

Fig. 1. . Human-Expressive Representations of Motion.

So the first goal is to determine which interpretations are feasible to be derived in each category of human motion, see Fig. 1. Consequently, for each category, suitable human-expressive representations of motion will be developed and tested. In particular, HSE will interpret and combine the knowledge inferred from three different categories of human motion, namely the motion of agent, body and face, in the same discourse domain.

The distinction between these motion categories is due to the fact that knowledge of different nature is required to interpret agent trajectories, body poses and facial expressions, since these types of interpretations strongly depend on the details of motion which can be inferred from active

video cameras. The strategy is to obtain the available information at a particular level (i.e., agent), thereby providing this incomplete knowledge to higher levels (i.e., body and face) which can update their representations as more information becomes available, and which can feedback the new information to the lower level.

The second objective of HSE is set to establish how these three types of interpretations can be linked together in order to coherently evaluate the human motion as a whole in image sequences. Such evaluation will require, at the very least, to *acquire* human motion from video cameras, to *represent* the recorded human motion using computational models, to *understand* the developments observed within a scene using high-level descriptions, and to *communicate* the inferred interpretations to a human operator by means of natural language texts or synthesized virtual agents as a visual language.

Thus, the main procedure of HSE will be the combination of:

- detection and tracking of agents while they are still some distance away from a particular location (for example a bus station, a pedestrian crossing, or a passenger in an airport, or a guest in a lobby);
- when these agents come closer to the camera, or when the active camera zooms in on these agents, their body posture will be evaluated to check for compatibility with behaviour hypotheses generated so far;
- if they are even closer and their face can be resolved sufficiently well, facial emotions will be checked in order to see whether these again are compatible with what one expects from their movements and posture in the observational and locational context which has been accumulated so far by the system.

Naturally, the interest is greater to integrate the three different components of human motion for someone approaching than someone leaving the camera. In addition, the most complex task (emotion evaluation) will come last, when the most is known already about the person in question from the preceding observations. Moreover, emotion recognition will become more specific because it can be embedded into the context of the preceding observations and it can exploit the rigid and non-rigid motion of the face.

A suitable discourse domain comprise two types of scenarios: (i) open worlds (such as well-frequented streets, pedestrian-crossings and bus-stops), and (ii) indoor worlds (airports, train stations or lobbies, for example). Multiple active cameras will record people to infer what the humans

intend to do. The main objective is the characterisation of humans to study the behaviour of people in these domains. It will be interesting whether the abilities to detect, track and characterize pedestrians would already be sufficiently advanced to reliably detect regional differences within the EU.

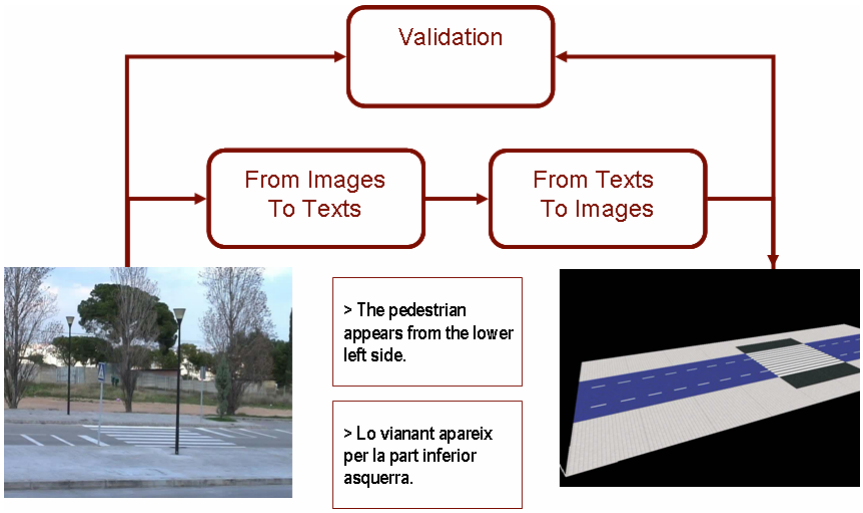

Fig. 2. Evaluation of Human Behaviours in Image Sequences.

By implementing the aforementioned tasks, HSE will fulfill two main objectives, see Fig. 2: on the one hand, the goal will be *description*, or the generation of conceptual descriptions based on acquired and analysed motion patterns. On the other hand, the aim will be communication using *visualization*, or the generation of synthetic motion patterns based on textual descriptions.

Firstly, natural language text generation will be accommodated within HSE based on the following considerations:

- Semantic descriptions will enable researchers to check details of the conceptual knowledge base.
- Semantic descriptions will allow communication with end-users of HSE in a most natural manner.
- Semantic descriptions will support conceptual abstraction, thereby facilitating the communication of short messages or essential details, possibly in response to inquiries communicated by a microphone near

the recording camera or by an UMTS mobile phone, for example for blind people.

Descriptive texts will be applied to outdoor or indoor scenes from different parts of the EU. The inclusion of videos from different parts of Europe will also constitute a means to prevent over adaptation of HSE to a small set of learning videos. In addition, once a system-internal conceptual representation has been built, it will be possible to enlarge this for natural language text generation in the languages of all groups cooperating within HSE. Also, we will test whether the same video recordings are interpreted in different manners in different parts of Europe (or similar situations just happen in a different manner, for example people nicely queuing up at a bus station in one country and habitually cluster around the bus doors in another). Thus, on the one hand, HSE will achieve automatic translation of visual information and, on the other hand, it will be able to investigate how and why human motion may produce different descriptions, due to the cultural characteristics of the areas where a given language is spoken.

Secondly, animation will be accommodated within HSE based on the following considerations:

- Analysis-by-synthesis at the three stages of human behaviour, i.e. motion of people, their posture analysis, and their face characterisation.
- Animated computer graphics as a visual language to quickly communicate essential aspects to involved people like bus-drivers, policemen for helping people at pedestrian crossings, waiters in a lobby, etc.
- Animated computer graphics, again at three motion categories, for checking the conceptual knowledge base underlying the entire approach. Since this knowledge base is expected to grow or need adaptation throughout the project, animated computer graphics will provide the means to quickly check larger parts.

Using both natural language text generation and animation, quantitative measures and qualitative descriptions will be developed to analyze the robustness and the efficiency of the proposed cognitive system. In fact, the performance of the system will be studied by considering the following strategy (Arens and Nagel 2003; Nagel 2004): let the system generate a synthetic image sequence using the textual descriptions obtained from a previously recorded image sequence. Both synthetic and original sequences can be compared to evaluate the suitability and correctness of the knowledge being considered so far.

Additionally, it will be possible to assess the results of the system by controlling the inference processes which are applied. The objectives will be:

- to trace the computational process which generates the result,
- to determine the internal information requested by the system, and
- to assess the selection of a particular interpretation.

As a result, it will become easier to debug the system. Therefore, the designer can decide to incorporate extra knowledge (by means of models, restrictions, and default options) for improving the performance of the system in terms of reliability. Also, there will be an increase of confidence of end-users in the results reported by the system: evaluation, in the sense of explanation, will ease the understanding of the results by non-expert users.

4 Innovation brought by HSE research

As an innovation, HSE proposes to develop an unified framework for human motion analysis which will be applied to confront both animation and description. Our basis is that procedures for synthetic video generation should rely on knowledge very similar to the knowledge required for textual description.

Image-sequence evaluation will be driven to incorporate assessment strategies to guide and validate the system results by:

- presenting the results of cognition using natural language texts or virtual animations, and
- arguing about inferred interpretations in order to assist and validate the system processes.

Using this know-how, we will be able to look for characterizing the behaviour of pedestrians approaching to a traffic-light-controlled pedestrian-crossing of a well-frequented inner-city street, for example. In this particular domain, a pedestrian-crossing could switch to green without grossly interfering with vehicle traffic by preparing the transition phase (green-yellow-red for vehicles) while vehicles are still some distance away. Also, switching back to green earlier, even saving gas, thus helping the environment, compared with stopping a cavalcade of vehicles in full drive after having had the pedestrians waiting for several minutes. A similar idea

could survey the environment around bus-stops with an associated gain in efficiency and comfort for all involved agents. Furthermore, provided one can extend this characterization of pedestrians reliably enough, it might become possible to design special help for handicapped people.

The basic procedure of HSE will be the combination of detection and tracking of agents while they are still some distance away from a particular location. Detecting and tracking people in crowded scenes is a challenging problem as people differ in their appearance caused by various types and styles of clothing and occluding accessories, undergo a large range of movements and moreover occlude each other. Previous approaches have either used appearance-based models or local features to detect people while a majority of trackers is still based on interactive initialization.

In HSE, cooperating pan-tilt-zoom sensors will also enhance this process of cognition via controlled responses to uncertain or ambiguous interpretations. Therefore, the challenge will be to provide sensor data for each of the modules by coupling the modules together in a sensor perception/action cycle (Nakazawa et al. 2002; Ukita and Matsuyama 2005). These cooperating pan-tilt-zoom sensors involved in acquisition will also serve the purpose of providing sensor data for each of the modules, but more importantly couple the other workpackages together in a sensor perception/action cycle. The use of zoom will provide an unification for interpretations of different resolution imagery, and will bestow the ability to switch the sensing process between different streams in a controlled fashion.

5 Conclusions

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 behaviour 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, identification, etc.).

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.

Acnowledgements

This work is supported by EC grants IST-027110 for the HERMES project and IST-045547 for the VIDI-video project, and by the Spanish MEC under projects TIN2006-14606, DPI-2004-5414 and CONSOLIDER-INGENIO 2010 (CSD2007-00018). Jordi Gonz`alez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

M. Arens and H.-H. Nagel. "*Behavioural Knowledge Representation for the Understanding and Creation of Video Sequences".* In Proceedings of the 26th German Conference on Artificial Intelligence (KI-2003), Hamburg, Germany; LNAI 2821, Springer-Verlag, 2003, pp. 149-163.

J. Ben-Aire, Z. Wang, P. Pandit, S. Rajaram, "*Human activity recognition using multidimensional indexing*", IEEE Trans. Pattern Analysis and Machine Intelligence 24 (8) (2002) 1091–1104.

D. Bullock, J. Zelek, "*Real-time tracking for visual interface applications in cluttered and occluding situations*", Image and Vision Computing 22 (12) (2004) 1083–1091.

I. Cohen, N. Sebe, A. Garg, L. Chen, and T.S. Huang. "*Facial expression recognition from video sequences: temporal and static modeling*". Computer Vision and Image Understanding, 91(1-2):160–187, 2003.

D. Comaniciu, V. Ramesh and P. Meer, "*Kernel-based object tracking*", IEEE Trans. Pattern Analysis and Machine Intelligence 25 (5) (2003) 564–577.

J. Deutscher, I. Reid, "*Articulated body motion capture by stochastic search*", International Journal of Computer Vision 61 (2) (2005) 185–205.

S. Dockstader, M. Berg, A. Tekalp, "*Stochastic kinematic modeling and feature extraction for gait analysis*", IEEE Trans. Pattern Analysis and Machine Intelligence 12 (8) (2003) 962–976.

A. Fod, M. Mataric, O. Jenkins, "*Automated derivation of primitives for movement classification*", Autonomous Robots 12 (1) (2002) 39–54.

A. Galata, N. Johnson, D. Hogg, "*Learning variable-length markov models of behaviour*", Computer Vision and Image Understanding 81 (3) (2001) 398–413.

D. Gavrila, L. Davis, "*3D model-based tracking of humans in action: A multiview approach*", in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'96), 1996, pp. 73–80.

J. Gonzàlez. "*Human Sequence Evaluation: the Key-frame Approach*". PhD Thesis. Universitat Autònoma de Barcelona. October 2004.

I. Haritaoglu, D. Harwood, L.S. Davis. "*W^4: Real-Time Surveillance of People and their Activities*". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No.8, pp. 809-830, 2000.

S. Intille, A. Bobick, "*Recognized planned, multiperson action*", International Journal of Computer Vision 81 (3) (2001) 414–445.

M. Isard, A. Blake, "*Condensation: Conditional density propagation for visual tracking*", International Journal of Computer Vision 29 (1) (1998) 5–28.

I. Karaulova, P. Hall, A. Marshall, "*Tracking people in three dimensions using a hierarchical model of dynamics*", Image and Vision Computing 20 (2002) 691–700.

A. Kojima, T. Tamura, K. Fukunaga, "*Natural language description of human activities from video images based on concept hierarchy of actions*", International Journal of Computer Vision 50 (2) (2002) 171–184.

Y. Li, S. Ma, H. Lu, "*A multiscale morphological method for human posture recognition*", in: Proceedings of Third Int. Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 56–61.

L. Li, W. Huang, I. Gu, Q. Tian, "*Statistical modeling of complex backgrounds for foreground object detection*", IEEE Transactions on Image Processing 11 (13) (2004) 1459–1472.

A. Lipton, H. Fujiyoshi, R. Patil, "*Moving target classification and tracking from real-video*", in: IEEE Workshop on Applications of Computer Vision (WACV'98), Princenton,NJ, 1998, pp. 8–14.

M. Ma, P. McKevitt, "*Interval relations in lexical semantics of verbs*", Artificial Intelligence Review 21 (3-4) (2004) 293–316.

O. Masoud, N. Papanikolopoulos, "*A method for human action recognition*", Image and Vision Computing 21 (8) (2003) 729–743.

T. Moeslund, E. Granum, "*A survey of computer vision based human motion capture*", Computer Vision and Image Understanding 81 (3) (2001) 231–268.

R. Morris, D. Hogg, "*Statistical models of object interaction*", International Journal of Computer Vision 37 (2) (2000) 209–215.

H.-H. Nagel, "*From image sequences towards conceptual descriptions*", Image and Vision Computing 6 (2) (1988) 59–74.

H.-H. Nagel, "*Steps toward a Cognitive Vision System*". AI Magazine, Cognitive Vision 25(2):31-50, 2004.

A. Nakazawa, H. Kato, S. Hiura, S. Inokuchi, "*Tracking multiple people using distributed vision systems*", IEEE Int. Conf. on Robotics and Automation 2002, pp. 2974-2981.

H. Ning, T. Tan, L. Wang, W. Hu, "*People tracking based on motion model and motion constraints with automatic initialization*", Pattern Recognition 37 (2004) 1423–1440.

K. Nummiaro and E. Koller-Meier, L.J. Van Gool. "*An adaptive color-based particle filter*". Image Vision Computing 21(1), pp. 99-110, 2003.

P. Remagnino, T. Tan, K. Baker, "*Agent oriented annotation in model based visual surveillance*", in: Proceedings of International Conference on Computer Vision (ICCV'98), Mumbai, India, 1998, pp. 857–862.

Y. Ricquebourg, P. Bouthemy, "*Real-time tracking of moving persons by exploiting spatio-temporal image slices*", IEEE Trans. Pattern Analysis and Machine Intelligence 22 (8) (2000) 797–808.

G. Sagerer, H. Niemann, "*Semantic networks for understanding scenes*", in: M. Levine (Ed.), Advances in Computer Vision and Machine Intelligence, Plenum Press, New York, 1997.

A. Sanfeliu and J.J. Villanueva, "*An approach of visual motion analysis*", Pattern Recognition Letters 26(3), pp. 355-368, 2005.

K. Schäfer, "*Fuzzy spatio-temporal logic programming*", in: C. Brzoska (Ed.), Proceedings of 7th Workshop in Temporal and Non-Classical Logics – IJCAI'97, Nagoya, Japan, 1997, pp. 23–28.

H. Sidenbladh, M. Black, L. Sigal, "*Implicit probabilistic models of human motion for synthesis and tracking*", in: A. Heyden, G. Sparr, M. Nielsen, P. Johansen (Eds.), Proceedings European Conference on Computer Vision (ECCV), Vol. 1, LNCS 2353, Springer-Verlag, Denmark, 2002, pp. 784–800.

C. Stauffer, W. Eric L. Grimson, "*Learning patterns of activity using real-time tracking*", IEEE Trans. Pattern Analysis and Machine Intelligence 22 (8) (2000) 747–757.

N. Ukita, T. Matsuyama, "*Real-time cooperative multiple-target tracking by communicating active vision agents*", Computer Vision and Iage Understanding 97(2), 2005, pp. 137-179.

S. Wachter, H.-H. Nagel, "*Tracking persons in monocular image sequences*", Computer Vision and Image Understanding 74 (3) (1999) 174–192.

D. Wagg, M. Nixon, "*Automated markerless extraction of walking people using deformable contour models*", Computer Animation and Virtual Worlds 15 (3-4) (2004) 399–406.

L. Wang, W. Hu, T. Tan, *"Recent developments in human motion analysis",* Pattern Recognition 36 (3) (2003) 585–601.

M. Yamada, K. Ebihara, J. Ohya, "*A new robust real-time method for extracting human silhouettes from color images*", in: Proceedings of Third International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 528–533.

Y. Zhang, E. Sung, E. C. Prakash "*3D modeling of dynamic facial expressions for face image analysis and synthesis*". International Conference on Vision Interface, Canada, 2001.