Implementation of an Intentional Vision System to Support Cognitive Architectures

Ignazio Infantino, Carmelo Lodato, Salvatore Lopes and Filippo Vella

Istituto di Calcolo e Reti ad Alte Prestazioni
Consiglio Nazionale delle Ricerche
ICAR-CNR sede di Palermo
edif. 11, Viale delle Scienze, 90128, Palermo, Italy

Abstract. An effective cognitive architecture has to be able to model, recognize and interpret user wills. The aim of the proposed framework is the development of an intentional vision system oriented to man-machine interaction. Such system will be able to recognize user faces, to recognize and tracking human postures by video cameras. It can be integrated in cognitive software architecture, and could be tested in several demonstrative scenarios such as domotics, or entrainment robotics, and so on. The described framework is organized on two modules mapped on the corresponding outputs to obtain: intentional perception of faces; intentional perception of human body movements. Moreover a possible integration of intentional vision module in a complete cognitive architecture is proposed.

1 Introduction

This paper describes a cognitive architecture developed with the aim of detecting human movements and perceiving actions and intents. In particular, we implemented an “intentional” vision system, that is, a system that “looks at people” and automatically perceives information relevant to interpret the human behavior [8]. The use of word "intentional" in this context concerns the purpose of generating a stream of pre-processed data useful for reasoning, recognition, reacting, and interacting when a human and his activity are objects of observation from the artificial system. The raw data coming from multiple sources of images and videos are filtered and processed in order to retain information useful to understand the human will, state and condition. Video sensors are non-intrusive as they do not require any contact with the user; they can also be used for security, monitoring, entertainment, domotic and other imaging purposes.

In order to model, recognize, and interpret human behavior, several tasks must be addressed [14]. Many of them are currently object of research investigations: face detection, location, tracking and recognition; facial expression analysis, human emotion recognition; audiovisual speech recognition; eye-gaze tracking; body tracking; hand tracking; gait recognition; recognition of postures, gestures, and activity.

Infantino I., Lodato C., Lopes S. and Vella F. (2008).
Implementation of an Intentional Vision System to Support Cognitive Architectures.
In *VISAPP-Robotic Perception*, pages 53-62
DOI: 10.5220/0002341100530062

Face detection and face recognition have reached a high level of precision and efficiency [10], by employing a number of computational models, based on features, shape, texture, and combinations thereof [18]. Models widely used are Active Appearance Models (AAMs), Principle Component Analysis (PCA, Eigenfaces), Linear Discriminant Analysis, Gabor wavelets, and statistical learning approaches that run in real time [16]. In real environment, head and face tracking algorithms suffer of various problems, and that is also true for facial expression analysis [13], which typically depends on the accuracy of facial feature tracking [7].

In order to allow interaction, tracking the human body and its constituent parts [9] is only the first step, that must be integrated by recognizing the behavior at different time scales. In such context, gesture recognition may be implemented as a problem to match a temporal sequence of body parameters, or to reasons about statistically defined gesture models. It is important to distinguish between unintentional human movements [8], movement for manipulating objects, and gestures used for communicating. Hidden Markov Models [12] and Bayesian networks have also shown promise for gesture recognition [1] have been extensively used to model and recognize gestures.

2 Intentional Framework Description

The proposed framework is named SeARCH In (Sensing-Acting-Reasoning: Computer understands Human Intentions). The relevant modules of the proposed architecture and their functional interconnections are depicted in figure 1. The core of intentional vision system is composed by two specialized modules: Intentional Perception of Body (IPB) and Intentional Perception of Face (IPF). The first module (IPB) deals with the detection and tracking of human bodies. In particular, it tries to locate silhouette, head, and hands of the people detected in the scene and performs their posture recognition. Furthermore, IPB detects and tracks also relevant objects moved by the hands. The output of this module consists in sequences of positions, and shape descriptors corresponding to all the detected entities. The second module (IPF) performs the recognition of the human detected in the scene and his face expression analysis. This recognition could be made more robust also including the silhouette detection arising from posture recognition task. The main output of IPF module is a temporal sequence of recognized facial expressions characterizing the human mood.

The sequences coming from both modules are linked to the relevant human states (hungry, sleeping, and so on) by the Reasoning Module (RM). RM outputs the interpreted human wills (to eat, to sleep, etc.) on the basis of IPF, and IPM data stream. Its effectiveness is improved on if knowledge of the individual is stored in the Collection of Habits (CH) that represents the memory of RM. Finally, the Action Planner module (AP) decides if and how the system has to interact, collaborate, or assist the human. A more detailed explanation of the single modules and their functions is reported in the following sections. Moreover, a simple example is reported in order to show a realistic scenario for the proposed framework. A robot perceives that its human companion is hungry and it will help him/her to lay the table. Actually, RM has been implemented as a simple rule based algorithm. The knowledge stored in the Collection Habit has been built by means of a supervised learning phase.

Finally, an imitation based approach has been used to record in the Action Planner all the operations necessary to accomplish the task just as the human is used to do it.

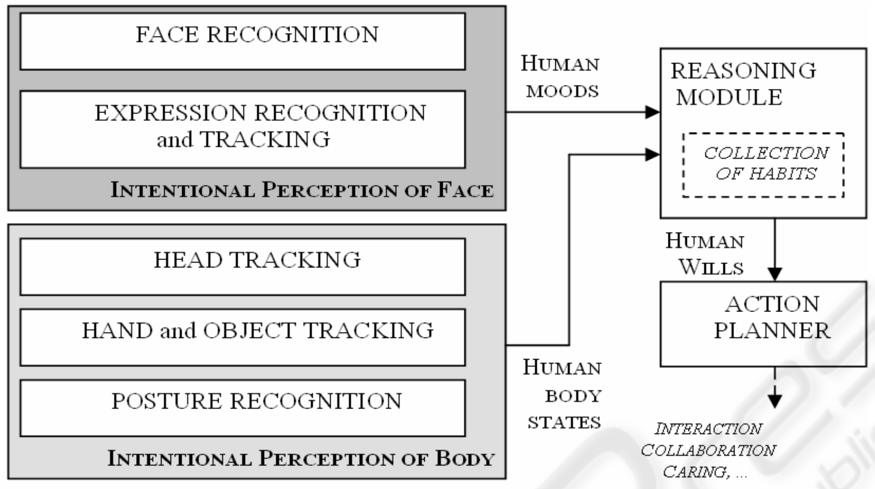

Fig. 1. *SeARCH In*: *S*ensing-*A*cting-*R*easoning: *C*omputer understands *H*uman *In*tentions. Intentional vision framework scheme.

3 Intentional Perception of Faces

The IPF module is devoted to recognize the people detected in the scene and to analyze their facial expression. The face recognition can be made more robust and effective using the results produced by the IPB posture recognition task (as explained in section 4.1). The main output of IPF module is a temporal sequence of recognized facial expressions that is sent to the reasoning module.

3.1 Face Recognition

The face recognition task is accomplished in few steps. First, the technique described in [16] is used for detecting the presence of faces in the scene. The Ada Boost algorithm is applied to simple detectors based on the Haar filters and the images areas related to the faces are then extracted and normalized. In fact, any robust identification process can compare faces when these are expressed in a canonical form. The relevant facial features, i.e. eye and mouth corners, central point of the nose, are localized by means of a template base detector. These points (at least three) are used to apply an affine transformation thus obtaining face images that are comparable with each other [2]. The rectified images are projected on the eigenspace derived from the most relevant eigenvectors (eigenfaces) generated from a training set [15]. A simple metric evaluation using the Mahalanobis distance (or Euclidean norm) allows estimating the degree of similarity in respect to the faces already stored in the database. If a positive match can be found, the corresponding label is attached to the current output data stream; otherwise the unknown face is recorded in the database.

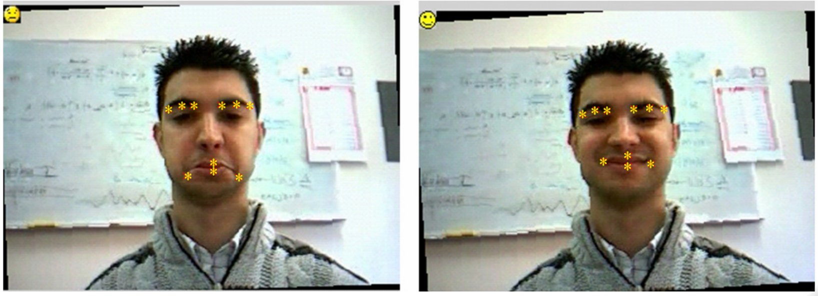

Fig. 2. Facial expression tracking and recognition: a rule-based recognition algorithm classifies simple expressions (anger, disgust, etc.) that are indicated by the yellow small icon depicted on upper-left corner of image; sequences of elementary emotions are considered to recognize human mood.

3.2 Face expression recognition

The facial expression recognition is performed by the tracking of eyebrows and lips movements. The Particle Filters are a well-suited solution for tracking multiple features and have been successfully used to follow the face movements [19]. The Tracking is initialized by the face detector [16] whereas the eyes and lips positions are detected by a correlation based template searching. Each frame of the video sequence is rectified in respect to the facial plane individuated by the eyes and the mouth (see figure 2). The distance between the eyes is employed for normalizing the feature positions. The relative position between eyebrows and eyes and between lips corners and the mouth centre are considered for recognizing the facial expression.

We have implemented a Facial Action Coding System (FACS) [7][6] classifying simple expressions: anger, disgust, fear, joy, sadness, surprise The sequences of these elementary emotions are recorded to the aim of building a sort of signature representative of a particular human condition in the scene: he/she is hungry, he/she is bored, and so on. In a learning phase, relevant sequences are recorded, and they will be included in the Collection of Habits of a particular person.

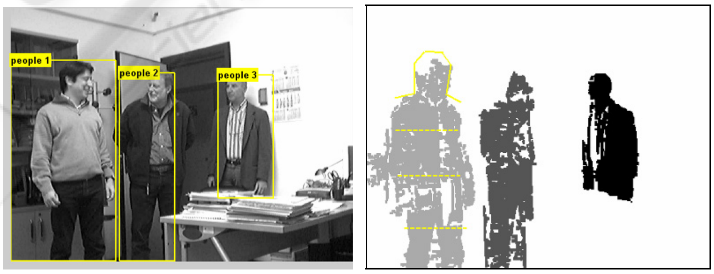

Fig. 3. Perception of human movements. Example of multiple people tracking and silhouette extraction: only sub-image region on the left shows a well defined shape, allowing to locate head.

4 Intentional Perception of Human Body

The Intentional Perception of Body module has been designed for the detection of human presence and activity. It accomplishes three main tasks that are the human silhouette localization, posture recognition and hands tracking. The input to IPB module comes from at least one fixed camera observing an indoor environment whose map is known. Two more cameras placed on a mobile robotic platform are then used to take close views of human standing in an area "of interest" monitored by fixed cameras.

4.1 Posture and Head Detection

We use the following approach to track the people [17] and recognize postures [4]. A Condensation algorithm, which belongs to the family of particle filters, has been implemented for tracking the people in a cluttered environment (see figure 3 on the left). The posture recognition is performed adopting a modified eigenspace technique and PCA in order to classify among several different postures. It is possible to perform the process of posture recognition which works together with the tracking algorithm because a posture feature space has been coded. Given the silhouette of a person extracted from the image by the tracking algorithm, it is first projected onto the feature space and then classified as belonging to one of known classes by using a PCA based clustering algorithm. As soon as the tracking algorithm individuates a person to track, the vision system also starts to estimate his/her posture. The posture recognition phase is interrupted only when some occlusions occur and the whole silhouette is no more visible. In this case, the vision system will output the last estimated posture before the occlusion has occurred, and it will try to re-capture it as soon as possible. If the silhouette is constituted by a single region and covers the most part of the tracking area, the head is located by using an omega shape template on region contours (see figure 3 on the right). The color histogram (in RGB space) of the silhouette sub-image (excluding head) is used to "attract" a robotic platform near a human.

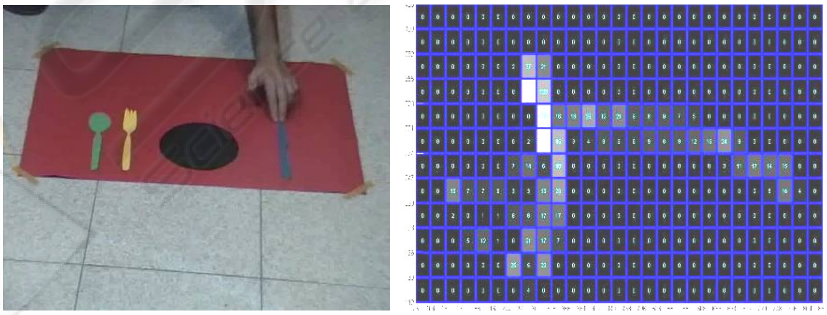

Fig. 4. Robot observation of a task showed by the human user: the task of "to lay the table" is simulated by the placing of simple planar objects (cutlery and plate) on red area (the table). Trajectories of object movements are recorded in order to build an occurrence matrix (see left side of the figure related to "fork"). This matrix is used for finding positional relation of objects in the table, and for calculating paths to place an object.

4.2 Hands and Objects Detection

When robot is observing an area “of interest”, the system attention will be focalized on hands and objects movements. The aim is to learn a human task by means of observation, or to collaborate/interact with the person if the task is known. The approach used for implementing both capabilities is inspired to the work of [11], where a simple statistical analysis is employed. The robot camera view is rectified in respect to the plane of the working area, where human hands manipulate several objects. The detected features are: position of centre of mass, color, and shape (by Fourier descriptors) of objects, and hands position. An occurrences matrix for each entity records the number of times that the object is detected in a particular location of the working plane. For example, a frame of the task “lay the table” is reported on the left of figure 4 whilst the occurrences matrix corresponding to the fork (yellow object) is shown on the right. Also in this case a particle filter is used to track working area entities. When a new working area is observed, or an object never seen before is noticed, occurrences matrices and features values are built or updated. If the knowledge of the working area is already acquired and known objects are detected, the intentional system could execute the task replacing the human actions or collaborate with him/her to place some objects until the final configuration is obtained.

5 Effective Implementation of an Intentional Vision System

Previous described modules could be composed in a flexible and dynamical way. The aim is to have an adaptable “intentional vision software system” which is capable to act in different situations or scenarios. For example, the recognition of behavior patterns may be considered more relevant than other aspects in the case of domotic scenario, whilst enabling natural and emotion-based human-machine interactions may be considered relevant in a robotic entertainment scenario. Generally, we could suppose to have a series of color cameras and video streams accessible by a large band network. Different computers will be used in order to process video streams coming from acquisition sources, and they will execute the same procedure related to the vision system. A coordinator software agent named CA will collect all visual information resulting from various sources, and will provide an aggregate view of the whole scenario, making available this data to the cognitive architecture for a further analysis. In simple experiments, CA agent could include reasoning module, action planner module, and collection of habits, allowing to have a complete cognitive architecture. The knowledge managed by intentional vision subsystem, updated at regular time interval, will record the following data: label of identified person, his/her localization, state of motion, posture, and facial expression, positions of his/her visible body parts, behavior pattern. Other information could be considered in order to completely satisfy the requirements of the cognitive architecture.

5.1 Example of Application

An experimental scenario has been arranged by using a mobile robotic platform (see figure 5 on the left) equipped with two pan-tilt web cameras. Moreover, a fixed camera is placed in a room where the robot can freely move. Two personal computers connected by a wireless network host the components of the intentional system. The on-board laptop manages robot movements, processes video and infrared signals to avoid obstacles allowing it to reach interesting places communicated by CA agent. This agent implements a rule-based algorithm that provides both reasoning and action planning capabilities. Details about each module of proposed architecture are reported in the following:

Face Recognition. We use a face dataset of 15 persons viewed in different illumination conditions. The recognition rate using the previously described approach is 95% of 750 detected faces. A binary value f_i is associated to each person (code 0000_2 means that face is not recognized).

Expression Recognition. Detected elementary emotions e_i are: neutral (001_2), anger (010_2), disgust (011_2), fear (100_2), joy (101_2), sadness (110_2), surprise (111_2), and not classified (000_2). Considering a video sequence of 1500 frames performances of expression detector are: correct tracking 85%, correct location of features (eyebrows, lip) 66%, correct detected emotion 63%. Human mood is defined by a sequence of elementary emotions. It is considered a temporal interval of 5 seconds, and 10 frames per seconds are processed to extract emotion (temporal sliding is 0.1 seconds). Then the vector $m_k = [e_1,e_2,\ldots,e_{50}]$ is the signal coding the human emotive state. At the present, we use a *k-mean* clustering algorithm for classifying 15 different states that resulted an optimal choose in our experimentation. Two clusters have been manually labeled as “hungry” and “confused” and could be respectively linked “to eat” and “to tidy”. We are investigating for setting up a complete meaningful taxonomy, and for finding a suitable number of clusters. We observed a rate of correct mood classification of 25% by interviewing involved users, but it is intrinsically difficult to establish an evaluation metric.

People and Posture Tracking. We have considered 7 body postures, with a success recognition rate of 95%. Positions in respect to the floor and sequences of 20 postures (10 frames per seconds) allow to classify people movements in the scene as “slowly walking”, “quickly walking”, “standing without acting”, “standing and acting”, “other”. We code these human body states by binary words p_i (3 bits). When the silhouette is composed by a single region covering at least 75% of tracking area, and if head is successfully located by an omega shape template the rest of body is considered as clothing. Four parallel sections subdivide the body (see figure 3), and mean RGB color and variance are recorded for each part. Signatures $d_i=[r_1, g_1, b_1, v_{r1}, v_{g1}, v_{b1}, \ldots, r_4, g_4, b_4, v_{r4}, v_{g4}, v_{b4}]$ are linked to recognized faces, and are also used for controlling the “curiosity” of robot.

Hands and Objects Tracking. The experimental scenario has two special places indicated as working areas and their position is supposed known. The red working area has been used to show how “to lay the table” using spoon, fork, knife, plate, and glass; the blue working area has been used to show the task “to tidy”, using book, pencil, stapler, and eraser. In these areas, the robot observes actions in order to learn tasks by examples given to it by humans. Objects and hands movements are detected on frames by segmentation, and occurrences matrices are built for determining relevant positions and trajectories. The mean error of objects and hands location is of 2.5 pixels. The wished final configuration to obtain is given by most probable objects positions. Each object o_i is described by a vector reporting 15 components of Fourier shape descriptor, and its mean color.

Reasoning and Planning. The knowledge managed by intentional system results from aggregating previously described data that constitute the "collection of habits" is coded as follows: $h_{ik} = [f_i, d_i, m_k, t_k]$, where *i* indicates a person, *k* one of his/her possible moods, t_k the task to perform (000= "nothing to do", 001= "make task 1", 010= "make task 2", 101= "learn task 1", 110= "learn task 2"). A list of objects and occurrences matrices $\{o_j, M_j\}_k$ corresponds to each task t_k. During the learning phase, when a human is near to a working area, the robot goes there to recognize him/her and observe actions. In normal activity, after the learning phase, the human-robot interaction is regulated by following set of simple rules:

- if the people tracking module detects a person close to a working area, and d_i is similar to a known one, the CA agent sends a command to make the robot approach such a place;
- if the face is recognized, then the robot observes the face expressions in order to determine his/her mood; else a new person is introduced in face database;
- afterwards the robot searches and selects a task among the available "collection of his/her habits" given the recognized mood. This task represents the human will to satisfy. We have performed 10 experiments for each task ("to lay the table", and "to tidy"): 5 are related to the learning phase and 5 to the collaboration one. Even if this is a preliminary experimentation, we report only 3 failures: 2 are due to erroneous recognition of human moods, and the other to erroneous recognition of the human face.

Interaction or Collaboration. Because the robot has no gripper or arm to manipulate objects, its collaboration during a task consists in indicating to the human which action to perform. For example "Please place plate on the centre of table", "Please place fork near spoon on the right", and so on. Considering the 10 "collaboration" experiments the rate of correct indications to human is near to 60%.

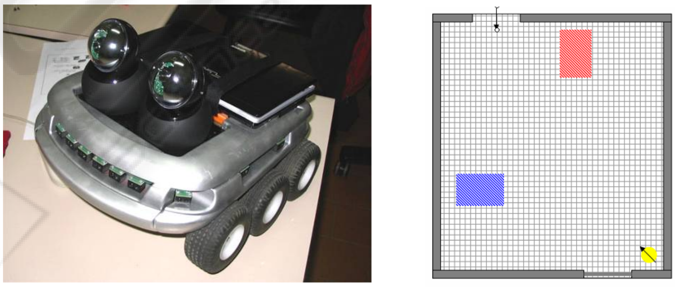

Fig. 5. On the left: Koala robot used to show an example of application of the proposed framework. On the right: map of environment that Coordinator Agent (CA) knows. There are two working areas indicated by red and blue rectangles, and one fixed camera observes the scene and is indicated by yellow circle. Obstacles and people positions are unknown, the access to the room and is indicated by the arrow in upper part of the map.

6 Conclusions and Future Works

The described framework aims to obtain a vision systems focused on the extraction of information useful to understand human wills. We have described a possible composition of several standard artificial vision algorithms for implementing an intentional vision system to insert in a cognitive architecture. More extensive experimentation is in progress to have better structure of collection of habits in order to gain efficiency and precision. Different applicative scenarios will be considered to have an exhaustive testing phase of the proposed architecture. Our intent is to include more sophisticated reasoning and planning modules. For example, it would be really interesting if the system could recognize when the human find difficult to accomplish a task. Moreover, we are investigating on suited qualitative metrics to evaluate the effectiveness of the robot behavior during collaboration phases.

Acknowledgements

We thank Giuseppe Arduino, and Dario Zangara for contributing hardware setup and implementation of software code; we are also grateful to Professor Antonio Chella and Dindo Haris for fruitful discussions about paper topics.

References

1. Bauckhage, C., Hanheide, M., et al., (2004), “A cognitive vision system for action recognition in office environments”, proc of. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004), vol. 2, pp. 827-833.
2. Berg, T.L., Berg, A.C., et al., (2004), “Name and faces in the news”, proc of. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004), vol. 2, pp. 848-854.
3. Chella, A., Dindo, H., and Infantino, I., (2006), “People Tracking and Posture Recognition for Human-Robot Interaction”, in proc. of International Workshop on Vision Based Human-Robot Interaction, EUROS-2006.
4. Chella, A., and Infantino, I., (2004), “Emotions in a Cognitive Architecture for Human Robot Interactions”, the 2004 AAAI Spring Symposium Series, March 2004, Stanford, California.
5. Ekman, P., (1992), “An argument for basic emotions”, in Cognition and Emotion, vol. 6, no. 3-4, pp.169–200.
6. Ekman, P., and Friesen, W.V., (1978), Manual for the Facial Action Coding System, Consulting Psychologists Press, Inc.
7. Fasel, B. and Luettin, J., (2003), “Automatic Facial Expression Analysis: A Survey”, Pattern Recognition, vol. 36, no 1, pp.259-275.
8. Kuno, Y., Ishiyama, T., et al., (1999), “Combining observations of intentional and unintentional behaviors for human-computer interaction”, in proc. of the SIGCHI conference on Human factors in computing systems, Pittsburgh, Pennsylvania, USA, pp. 238-245.
9. Moeslund, T.B., and Granum, E., (2001), “A survey of computer vision-based human motion capture”, Computer Vision and Image Understanding, vol. 18, pp. 231-268.

10.Phillips, P.J., Flynn, et al., (2005), "Overview of the face recognition grand challenge", in proc. of Computer Vision and Pattern Recognition, 2005 (CVPR 2005), pp. 947-954.

11.Rao, R.P.N, Shon, A.P., and Meltzoff, A.N., (2007) "Imitation and Social Learning in Robots, Humans and Animals", in "Imitation and Social Learning in Robots, Humans and Animals", Cambridge Press, pp. 217-248.

12.Starner, T., and Pentland, (1995), "A. Visual recognition of American Sign Language using Hidden Markov Models", in proc. of International Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, pp. 189–194.

13.Tian, Y.L., Kanade, T., and Cohn, J.F, (2001), "Recognizing action units for facial expression analysis", IEEE Trans. on Pattern Analysis and Mach. Intell., vol. 23, no. 2, pp. 97-115.

14.Turk, M., (2004), "Computer Vision in the Interface", Comm. of the ACM, vol. 47, no 1.

15.Turk, M., and Pentland, A., (1991), "Face recognition using Eigenfaces", in proc. of Computer Vision and Pattern Recognition 1991, pp.586-591.

16.Viola, P., and Jones, M. J., (2004), "Robust Real-Time Face Detection", International Journal of Computer Vision, vol. 57, no 2, pp. 137-154.

17.Wren, C., Azarbayejani, A., et al., (1997) , "Pfinder: Real-time tracking of the human body", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 780-785.

18.Zhao, W., Chellappa, R., et al., (2003), "Face recognition: A literature survey", ACM Computing Surveys (CSUR), vol. 35, no. 4, pp 399-458.

19.Zhou, S.K., Chellappa, R., Moghaddam, B., (2004), "Visual Tracking and Recognition Using Appearance-Adaptive Models in Particle Filters", IEEE Trans. On Image Processing, vol. 13, no. 11, pp. 1491-1506.