A Fuzzy System for Detection of Interaction Demanding and Nodding Assent Based on Stereo Vision

Eugenio Aguirre, Miguel García-Silvente, Antonio González, Rui Paúl and Rafael Muñoz-Salinas

***Abstract*—Despite of the advances achieved in the past years in order to design more natural interfaces between intelligent systems and humans, there is still a great effort to be done. Considering a robot as an intelligent system, the determination of some typical interaction situations is an interesting ability to achieve. This paper shows a fuzzy system that allows the visual detection of possible interaction demands and the shaking or nodding of the head. To achieve this objective, the robot has first to carry out the detection and tracking of the people using the stereo vision system. Then, the level of interest of a person to interact with the robot is calculated by analyzing his/her position and the pose of his/her head. The head pose is estimated in real-time by a view based approach using Support Vector Machines (SVM). Whenever the level of interest achieves an high value, the person is analyzed in more detail to detect the position and the motions of his/her arms as well as whether the person is shaking or nodding his/her head. This information is managed by a fuzzy system to detect a possible interest demand or the intention of the person to say yes or no using his/her head. At the end of the paper, some experiments are shown to validate the proposal.Finally, future work is addressed.**

***Index Terms*—Human-Robot Interaction, Interest Detection, Attention Estimation, Head Shaking and Nodding, Support Vector Machines.**

I. INTRODUCTION

NOWADAYS, the interaction between Intelligent Systems and human beings is a topic that is focusing a great research effort. In particular, within the area of Robotics, the development of successful robotic systems applied to service tasks in home and office environments implies the generation of natural human-robot interfaces. In that sense, important issues that must be taken into account are how robots can detect the presence of people around them and how do they recognize when and how long a person is interested in establishing an interaction. In order to achieve this goal, it is necessary to solve several problems. First, a robot must be able to detect people in its vicinity and track their movements over time [1]. Second, once the surrounding people is spotted, it should be able to detect their interest in establishing an interaction. In this task, several types of signals from the human can be taken into account (both verbal and non-verbal). Some authors [2] use sound source localization or speech recognition combined with visual perception to detect which people are the most interested. In other cases, facial expressions [3] or hand gestures [4] are analyzed.

This paper shows the effort carry out by our research team in this topic and it represents a new step in the work shown in the last edition of the Physical Agents Workshop [5]. In this paper we are interested in the analysis of some typical interaction situations that can be integrated in a more complex system in a future. The proposed situations are: i) the interaction demanding through the position or motions of the arms; ii) the shaking and nodding of the head to express assent or negation. These analysis are carried out using visual information and dealing with the underlaying uncertainty and vagueness through fuzzy logic.

A. Related Work

From a long time ago researchers have been fascinated by the possibility of developing robots that could interact with other people and robots. In the 1940s Walter [6] built a robot capable of interacting with another in a seemingly "social" manner, although there was no explicit communication or mutual recognition between them.

Some years later, Deneubourg and his collaborators pioneered the first experiments on stigmergy (indirect communication between individuals via modifications made to the shared environment) in simulated and physical "ant-like robots" [7], [8]. This idea is based on the concept of colonies of insects which are able to work for a common goal, although each individual works alone.

Nowadays, social robots can be used in a wide range of applications like toys, educational tools, therapeutic aids, etc. As they play different roles (often operating as partners or assistants with different people) they have to show flexibility and adaptability. A survey and taxonomy of current applications is given in [9] where Fong et al. make a deep analysis of most of the aspects to take into consideration when designing a social robot.

This kind of robots have different shapes and functions ranging from robots aimed to engage people in social interactions (Kismet, Cog, etc.) to robots that are engineered to adhere to social norms in order to fulfill a range of tasks in human-inhabited environments (Pearl, Sage, etc.) [10], [11], [12], [13].

Another aspect is that it may also be desirable for a robot to improve its interaction skills over time. If we think about a

Eugenio Aguirre, Miguel García-Silvente, Antonio González and Rui Paúl are with Department of Computer Science and Artificial Intelligence at E.T.S. Ingeniería Informática at University of Granada. 18071 Granada, Spain.
E-mail: {eaguirre, mgs, gonzalez, ruipaul}@decsai.ugr.es
Rafael Muñoz-Salinas is with Department of Computing and Numerical Analysis at Escuela Politécnica Superior at University of Cordoba.
E-mail: rmsalinas@uco.es

pet robot that accompanies a child through his childhood we will get to the conclusion that this kind of robot may need to improve its skills in order to maintain the child's interest as he/she gets older. Learned development of social (and other) skills is a primary concern of epigenetic robotics [14], [15].

In order to robots to behave and perceive things similarly to humans they have to understand and react to the world and its situations as humans do. Furthermore they should have an intrinsic notion of sociality, develop social skills and bond with people, recognize social context and convention and show empathy and true understanding. At present, such robots remain a distant goal [16], [14]. Contributions from other research areas such as artificial life, developmental psychology and sociology [17] are of extreme importance for the achievement of these goals.

Their embodiment is also an important issue. To be easily accepted in the society, they should have the most adequate aspect related to the function they will execute [18]. For example, a robot that is designed for playing the role of a dog, should not just "bark" or behave like a dog but also have the morphology of a dog in order to be accepted by an human being as this species of animal.

Social robots also need human oriented perception. This means that they should be able to track people and human features such as bodies, faces, hands and others, to interpret human emotions including affective speech, discrete commands and natural language and to recognize facial expressions, gestures and other kind of human activity. Simulating human-like emotions is also a desirable task to achieve. For that reason it is also considered as an important and still under development topic in Human-Robot Interaction.

B. Our System

Our hardware system is comprised by a Nomad 200 mobile robot, a stereoscopic system with a binocular camera [19] and a laptop for processing all the data (see Fig. 1). The camera and the laptop are mounted on the top of the robot structure.

People detection and tracking problems are solved thanks to an own previous work based on stereo-vision [20] using this kind of vision device. The use of stereo vision brings several advantages when developing human-robot applications. First, the information regarding disparities becomes more invariable to illumination changes than the images provided by a single camera, being a very advantageous factor for the background estimation. Second, the possibility to know the distance to the person could be very useful for the tracking as well as for a better analysis of their gestures.

Once the people are located, the level of interest of each person to interact with the robot is calculated by analyzing his/her position and his/her degree of attention. The position of a person is analyzed using both his/her distance to the center of the robot and his/her angle in respect to the heading direction of the robot. With respect to the degree of attention, this is determined detecting the orientation of the head, i.e., a higher degree of attention can be assumed when a person is looking at the system than when it is backwards. This analysis is solved by a view based approach using Support Vector Machines

Fig. 1. Robot with stereo vision system.

(SVM)[21]. Thanks to SVM, head pose can be detected achieving a great percentage of success independently of the morphological features of the heads. Fusing this information with fuzzy logic, a level of interest in interacting with the robot can be computed for each detected person. When the level of interest is high, the person is analyzed in more detail to detect some of the interaction situations commented above.

The approach presented in this work is not only valid for robotic applications but also in ambient intelligence that use stereoscopic devices.

The remainder of this paper is structured as follows. Section II gives a short overview of the hardware and software system, giving a basic explanation of the method employed for the people detection and tracking. In Section III it is explained the SVM based approach to estimate the head pose and the fuzzy system employed for estimating the interest of people. In section IV it is described the method to detect the demand for attention and head shaking or nodding. In Section V it is shown the experimentation carried out, and finally, Section VI outlines some conclusions and future works.

II. People detection and tracking

The ability of detecting and tracking people is fundamental in robotic systems when it is desirable to achieve a natural human-robot interaction. They are achieved in our architecture by combining stereo vision and color using plan-view maps. Following, the process for people detection and tracking is explained in a summarized way. The readers more interested in this process are referred to [20].

Our robot has a stereo camera that is mounted on a pan-tilt unit (PTU). The stereo camera captures two images from slightly different positions (calibrated stereo pair) that are transferred to the computer to calculate a *disparity image* containing the points matched in both images (see Fig. 2). By knowing the intrinsic parameters of the stereo camera it is

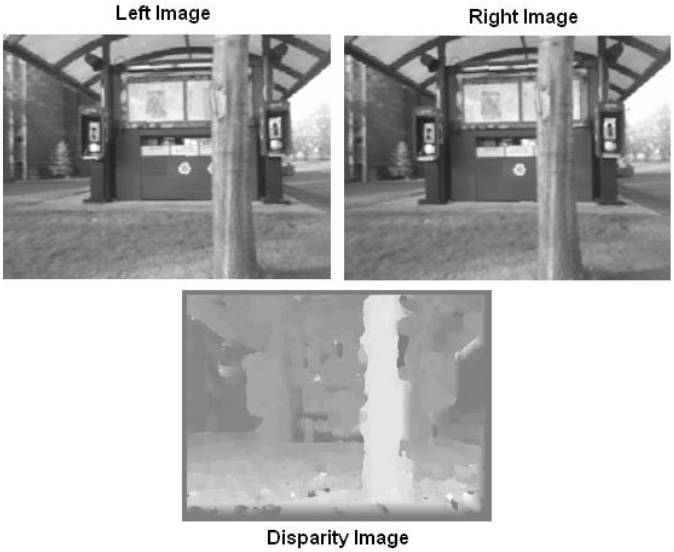

Fig. 2. Left and right images taken by the camera and the corresponding disparity image.

possible to reconstruct the three-dimensional position p_{cam} of a matched pixel (u, v). Then, the points captured are translated to a *"robot"* reference system, placed at the center of the robot at ground level in the direction of the heading of the robot. Generally, the number of points captured by our stereo system is very high. In order to perform a reduction of the amount of information, the points captured by the camera are orthogonally projected into a 2D plan-view map $\mathcal{O}$ named *occupancy map* [22], [23], [24].

Before the detection process begins, the environment must be registered. This step aims to register the structure and motionless objects of the environment by building an environment model. In a posterior, phase it will allow us to detect the objects that are not part of the environment and then we will be able to consider them as movable objects. Our approach is based on the creation of a geometrical height map of the environment $\mathcal{H}_{max}$, that divides the ground level into a group of cells a fixed size. Height maps have been used in mobile robotics in order to describe the environment and planning trajectories [25], [26]. The points identified by the stereo system are projected on $\mathcal{H}_{max}$, storing the maximum height of the projected points in each cell. To avoid adding the points of the ceiling on $\mathcal{H}_{max}$, the points that overcome the height threshold h_{max} are excluded from the process. Due to efficiency reasons, the points below the minimum height threshold h_{min}, are also excluded. The height range $[h_{min}, h_{max}]$ should be such that the majority the person's body to be detected should fit in it. On those cells $\mathcal{H}_{max}(x, y)$ on which there are no points located, we assume that there are no objects and therefore the height is h_{min}. Instead of building the height map from a single image, it is built from several observations that are fused using the median operator. With this method we can build the height map even in the presence of moving objects in the environment.

In Fig. 3 we can observe the creation of the height map of an environment. In this example the map has been created in the presence of two persons moving in the environment. Fig. 3(b) shows the environment height map. Dark areas represent the highest zones and the white areas represent the lowest ones (h_{min}). To create this map we have used the size of cells $\delta = 1$ cm and the range of height is $h_{min} = 0.5$ m and $h_{max} = 2.5$ m.

Once the height map $\mathcal{H}_{max}$ has been created, the people detection process can begin. The first step consists in creating an occupancy map $\mathcal{O}$, that indicates on each cell the surface occupied by the objects that do not belong to the environment. For this purpose, after capturing a stereo pair of the environment, stereo processing is performed. For each point it is evaluated if its height is within the limits $[h_{min}, h_{max}]$ and if it exceeds the value of the corresponding cell in $\mathcal{H}_{max}$. In that case, the equivalent cell in $\mathcal{O}$ is incremented by a value proportional to the surface that occupies in the real world. Points detected far from the camera, increment the corresponding cell with a higher value than nearer points. So, it is compensated the change in size observed for the same object when it is seen at different distances. Hence, the weight of its projection is independent of the distance.

When $\mathcal{O}$ is created, it is analyzed to detect the objects that appear on it. The first step consists in applying a closing process in order to link possible discontinuities on the objects. Afterwards, objects are detected by grouping cells that are connected and whose sums of areas overcome the threshold θ_{min}. In this way, we eliminate the potential noise that appears as a consequence of the stereoscopic process. On Fig. 3(c) we can observe the occupancy map $\mathcal{O}$ of the environment on Fig. 3(a) using a height map $\mathcal{H}_{max}$ from Fig. 3(b). The darker values represent the areas with higher occupancy density. On Fig. 3(c) we can see two detected objects bounded with frames after the closing process, grouping and thresholding. Finally, Fig. 3(d) and 3(e) show the 3D reconstruction of the scene of Fig. 3(a). While in Fig. 3(d) all the points detected by the stereo system are shown, in Fig. 3(e) there are only drawn those belonging to the foreground, i.e., those used for voting in $\mathcal{O}$.

The next step in our processing, is to identify the different objects present in $\mathcal{O}$ that could correspond to human beings (*human-like objects*). For that purpose, $\mathcal{O}$ is processed with a closing operator in order to link possible discontinuities in the objects caused by the errors in the stereo calculation. Then, objects are detected as groups of connected cells. Those objects whose area are similar to the area of a human being and whose sum of cells (occupancy level of the object) is above a threshold θ_{occ} are considered human-like objects. This test is performed in a flexible way so that it is possible to deal with the stereo errors and partial occlusions. However, the human-like objects detected might not belong to real people but to elements of the environment. The approach employed in this work to detect a person consists in detecting if any of the human-like objects found in $\mathcal{O}$ show a face in the camera image. In Fig. 4.c it is possible to see an example of the occupancy map with two objects detected.

Face detection is a process that can be time consuming if applied on the entire image, thus, it is only applied on regions of the camera image where the head of each object should

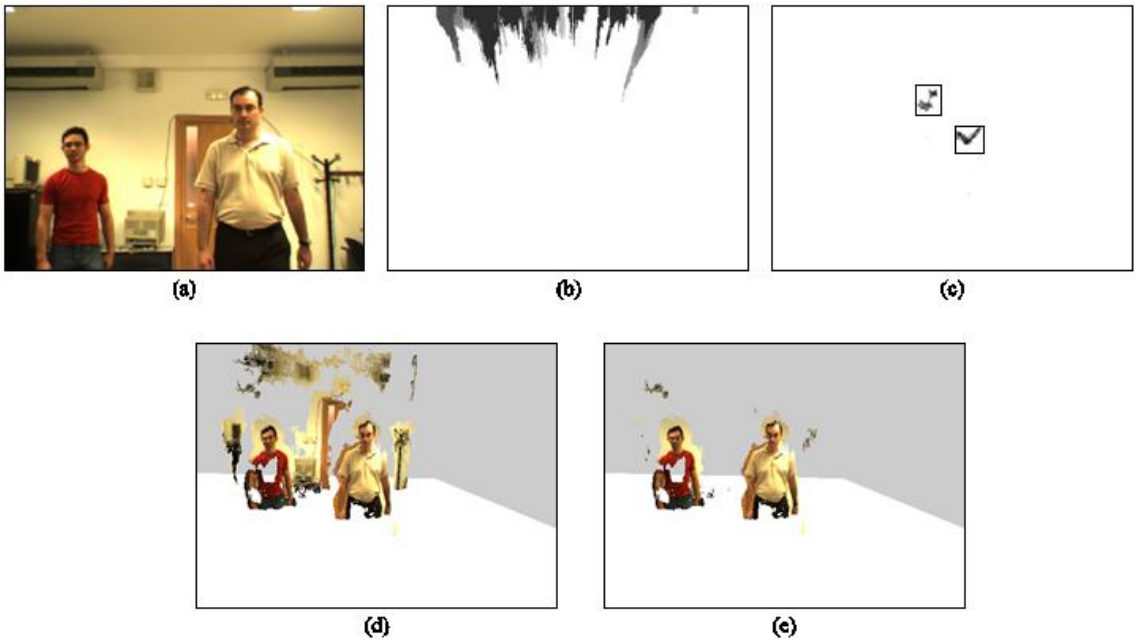

Fig. 3. (a) Image of our environment containing two people. (b) Background map $\mathcal{H}_{max}$ of our environment created in a previous phase. (c) Occupancy map $\mathcal{O}$ of the Image a. There can be seen the two objects detected. (d) 3D reconstruction of the scene including both background and foreground points. (e) 3D reconstruction of the background points, i.e., those projected in $\mathcal{O}$.

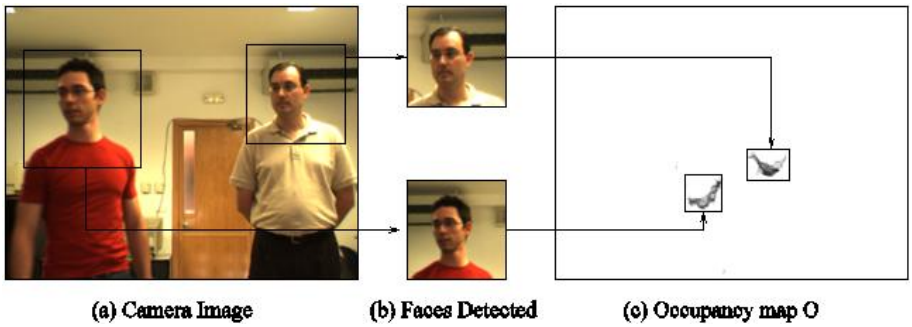

Fig. 4. (a) Image of the scene (b) Upper part of the objects in the scene that are analyzed to detect faces on them (c) Occupancy map $\mathcal{O}$ of the scene

be (head region). As the human head has a typical average width and height, the system analyzes first if the upper part of a human-like object has a similar size. If the object does not pass this test, the face detector is not applied to it. This test is performed in a flexible manner so that it can handle stereo errors and people with different morphological characteristics can pass it. If the human-like object passes this test, the corresponding region in the image is analyzed to detect if it contains a face. The face detector employed is based on the face detector of Viola and Jones [27] which was later improved by Lienhart [28]. We have employed the OpenCv's Library [29] implementation that is trained to detect frontal human faces and works on gray level images. In Fig. 4.a we have the camera image with the faces of the two previously detected objects bounded by a square.

Once a face has been detected on a human-like object, a color model of the person torso is created [30]. The idea is to assist the tracking process by capturing information about the color of the clothes of the user so that the robot can distinguish him/her from other people in the environment. Therefore, pixels around what it should be the chest of the person are used. The position of the chest in the camera image is estimated as 40 cm below the top of the head region. The size of the region used to create the color model depends on the distance of the person from the camera. When the object is far from the camera the region used is smaller to avoid including pixels from the background and it becomes bigger when the object is near to the camera.

Tracking consists in detecting in subsequent frames the human-like object that corresponds to the person being tracked. The Kuhn's well-known Hungarian Method for solving optimal assignment problems [31] is employed for that purpose. Two pieces of information are combined (position and color) to assign a value to each human-like object indicating its likelihood to be the person being tracked. On one hand, a prediction of the future position of the person being tracked is calculated using the Kalman filter. The nearer a human-like object is from the position estimated for the person being tracked, the higher likelihood it will have to be him/her. On the other hand, color information is employed to achieve a more

robust tracking. The more similar the color of a human-like object is to the clothes' color of the person being tracked, the higher likelihood it will have to be him/her. Both likelihood are combined so that when the person being tracked is near others, color information can help to distinguish him/her. The human-like object with highest likelihood is considered to be the person being tracked if its likelihood value exceeds a certain threshold. In that case, the Kalman filter is updated with the new observations and also the color model of the person is updated so that it can adapt to the illumination changes that take place.

When the position of the person being tracked is located, the system determines the location of his/her head in the camera image. In this work, the head is modeled as an ellipse whose size in the camera image is determined according to the distance of the person to the camera. Firstly, the system calculates an initial estimation of the head position in the camera image based on stereo information. Then, the initial position is refined by a local search process. For that purpose, the gradient around the ellipse perimeter is examined in order to determine the likelihood of a position using the Birchfield's method [32]. The position with higher likelihood is considered the person's head position.

III. Interest detection

This section explains our approach for estimating the interest of the detected people in interacting with the robot using fuzzy logic. The approach presented in this work is based on stereo vision but the system can be easily expanded to merge other sources of information. The advantages of using fuzzy logic are mainly three. Firstly, the robot has to deal with information from the stereo system that is affected by uncertainty and vagueness. Fuzzy logic is a good tool to manage uncertainty using linguistic variables. Secondly, the human knowledge can be usually expressed as rules. Fuzzy logic allows to establish relationships among the variables of a problem through fuzzy rules providing an inference mechanism. Finally, there are methods in fuzzy logic to fuse the results from several fuzzy rules in order to achieve a final overall result. Therefore, the system designed in this work, based exclusively on stereo information, could be easily integrated with other fuzzy systems using other types of information as source sound localization, gesture analysis or speech recognition systems. In this work, the determination of the degree of interest of a person is based on its position and its degree of attention. The position of a person is analyzed using both its distance to the center of the robot and its angle respect to the heading direction of the robot. The first feature is measured by the linguistic variable $Distance$ and the second one by the linguistic variable $Angle$. Each one of these linguistic variables has three possible values as shown in Fig. 6. These two variables are used to establish the following rule: if the person is detected near to the robot and more or less centered with respect to it, then we consider that the person is more interested in establishing interaction with the robot than when the person is far or at the left or right side of the robot. Nevertheless, the position of the person is not enough to determine his/her interest in interacting with the robot. Thus, the third feature shown in this paper is the person's attention detected by the analysis of the pose of the head. To detect the head pose we have employed a view based approach using SVM that is explained in the next section.

A. Estimating face attention using SVM

One of the most prominent cues to detect if a person is paying attention to the system is the orientation of the face, i.e., a higher degree of attention can be assumed when a person is looking at the system than when it is backwards. This section describes our approach for face attention estimation.

We have divided head poses in three main categories: "A" that comprehends all the frontal faces (faces looking directly at the camera), "B" that comprehends all the slightly sided faces (faces looking to some point slightly above, below or aside from the camera) and "C" that comprehends all the other faces (side faces, faces looking at some point in the ceiling or ground, backward heads). Figure 5 shows examples of each one of the categories employed.

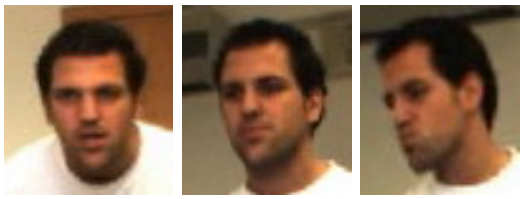

Fig. 5. Head Pose Estimation: Classes A, B and C.

We have created a head pose database comprised by a total of 4000 samples equally distributed among the three classes. The database contain images of 21 different people (men and women), of different races, with different hair cuts and some of them wearing glasses. The database samples were manually classified into categories "A", "B" or "C" according to where people were looking at. All the images are gray-scale and 48x40 sized.

In order to reduce the data dimensionality, we have applied Principal Component Analysis (PCA). PCA [33] is a technique widely employed for dimensionality reduction. When using PCA, an image is transformed into its principal components, i.e., those that contain the "most important" aspects of the data. PCA has the distinction of being the optimal linear transformation for keeping the subspace that has largest variance. A key aspect in PCA consists in deciding how many principal components are appropriate for a proper training. A low number of characteristics reduces the computing time required, but also reduces the accuracy since part of the information is discarded, and vice versa. Therefore, we have made tests with different number of characteristics and we have determined that 50 characteristics allow a good trade-off between classification accuracy and computing time.

The training process has been carried out using Support Vector Machines. Two are main advantages of Support Vector Machines over Artificial Neural Networks (ANN). First, most of the modalities of ANNs can suffer from multiple local minima while the solution supplied by SVM is global and unique. Second, the computational complexity of SVM does

not depend on the input data dimensionality, unlike ANNS. For that purpose, we have employed the libsvm library (free software available in Internet [34]). To certificate that results were satisfactory before applying the model we trained the SVM with 85% of the data set and kept the remainder 15% to test the model generated. The result on the test set was of 93.14% of accuracy.

For each detected person the SVM classifier estimates in real time the head pose in one of the three categories previously indicated. The output of the SVM classifier is translated into a numerical value by the definition of the variable $SVMOut$. The value of $SVMOut$ in the time t is

$$SVMOut_t = \begin{cases} 1 & \text{if output SVM = “A”;} \\ 0.5 & \text{if output SVM = “B”;} \\ 0 & \text{if output SVM = “C”.} \end{cases}$$

However, $SVMOut_t$ is an instantaneous value that does not take into account past observations. In order to consider past observations, we define the variable $HP_{(t)}$ as:

$$HP_{(t)} = \alpha HP_{(t-1)} + (1-\alpha) SVMOut_t \qquad (1)$$

where the initial value for HP is the first value of $SVMOut$ when the person is detected and α is a weighting factor that ponders the influence of past observations.

To deal with the uncertainty and vagueness in this process we use a linguistic variable called “Attention” and divide it into “High”, “Medium” and “Low” values (see Fig. 6). This variable will take as input values the measures of face attention estimation considered by HP (Eq. 1). Figure 6 it is possible to see the labels for the variable “Attention”.

B. Fuzzy system for interest estimation

Once the three linguistic variables have been defined, the rule base that integrates them are explained in this section. The idea that governs the definition of the rule base is dominated by the value of the variable $Attention$. If the attention has a high value then the possibility of interest is also high depending on the distance and the angle of the person to the robot. If the attention is medium then the possibility of interest has to be decreased but like in the former case depending on the distance and the angle. Finally, if the attention is low, it means that the person is not looking at the area where the robot is located and the possibility of interest is defined as low or very low depending on the other variables. The rules for the case in which $Attention$ is High are shown by Table I. The other cases are expressed in a similar way using the appropriate rules. The output linguistic variable is $Interest$ and has the five possible values shown by Figure 6(d).

Finally in order to compute the value of possible interest, a fuzzy inference process is carried out using the operator minimum as implication operator. Then the output fuzzy sets are aggregated and the overall output is obtained. The overall output fuzzy set can be understood as a possibility distribution of the interest of the person in the $[0, 1]$ interval. Therefore values near to 1 mean a high level of interest and vice versa.

TABLE I
RULES IN THE CASE OF HIGH ATTENTION.

IF			THEN
Attention	Distance	Angle	Interest
High	Low	Left	High
High	Low	Center	Very High
High	Low	Right	High
High	Medium	Left	Medium
High	Medium	Center	High
High	Medium	Right	Medium
High	High	Left	Low
High	High	Center	Medium
High	High	Right	Low

IV. RECOGNIZING TYPICAL INTERACTION SITUATIONS

After estimating the interest as described in the previous section, it is desirable that the robot centers its attention in the person that is more interested in interacting with it. This person could possible be willing to communicate with the robot in different ways.

The goal in this section is to compute whether a person, whose level of interest estimated before is high, is requesting attention from the robot. We are interested in the analysis of some typical interaction situations that can be integrated in a more complex system. The proposed situations are: i) the interaction demanding through the position or motions of the arms; ii) the shaking and nodding of the head to express assent or negation. These analysis are carried out using visual information and dealing with the underlaying uncertainty and vagueness through fuzzy logic.

After detecting the level of interest among the people in the surroundings of the robot, the system detects if the person is static (not moving or moving very slowly). If so, the system analyzes whether the person is standing (rising or extending) one or both arms as well as whether he/she is doing any movement with any of them.

As during an interaction people tend to ask and answer typical yes/no questions, we have also developed a method employed to detect whether the ”interested” person is shaking or nodding his/her head. This feature can be employed in a more complex system, when the robot is able to, for instance, ask questions to the user.

A. Gestures Detection

One of the most common ways to request somebody’s attention using gestures is to raise or to shake one or both arms. Therefore we focused our efforts in detecting whether a person who might be interested in communicating is doing this kind of gestures or not.

Using the information supplied by the stereo system it is possible to know the position and distance at which many of the image pixels are located, and therefore compute which objects are part of the foreground. To achieve that goal we use and algorithm based on the “Distance of Mahalanobis” to separate the pixels that are part of the background from those who belong to the foreground. The “Distance of Mahalanobis” is based not only in the euclidean distance but also in the correlation between two variables. Afterwards, we separate

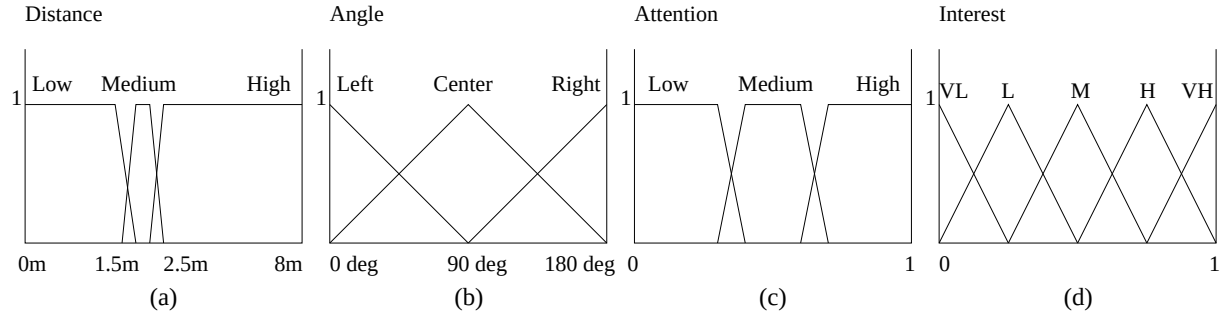

Fig. 6. Fuzzy sets of the linguistic variables: (a) Distance (b) Angle (c) Attention (d) Interest

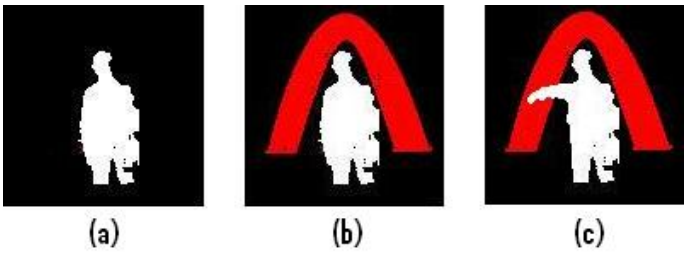

Fig. 7. (a) Silhouette of a person. (b) Silhouette image marked with area for detecting raised arms (red). (c) Silhouette with red marked area and with a raised arm inside it.

the objects and people belonging to the foreground, using the information of the position of each person given by the tracker. We apply a recursive algorithm called "Flood Fill" to build an image of the silhouette of that person. This algorithm computes whether the pixels surrounding the root pixel (the pixel assigned to be the center of mass of the person) are within a specific "distance" and, therefore, also belong to that person. If so it continues to search for pixels in the surroundings of the new pixels that were previously classified as belonging to the person. When no more pixels satisfy this condition we have the image of the mask of that person where each pixel that belongs to the person has the information about its distance to the camera and pixels not belonging to the person are set to value 0. A more simple version of this image is the silhouette of the person or the binary image of the person as seen in Fig. 7(a). In this picture, pixels in white belong to the person while pixels in black do not belong to the person.

By doing this it is possible to analyze if there are pixels around the person's body that could be part of a raised arm. In our system we search for pixels that are not set to 0 in an elliptic region around the person (see Fig. 7(b)). The inner border of the ellipse is just next to the person's exterior border while the outer border of the ellipse is located more externally in a way that any possible raised arm could fit inside the elliptic region. In Fig. 7(c) we can see an example of the silhouette of a person whose arm is inside the elliptic region, indicating that the person is raising it or moving it. We only examine the upper half of the ellipse because people's arms can never be below the hips whenever a person is standing up. We define the linguistic variable $RaisedArm$ that can take three values represented by the labels "Zero" "One" and "Two" (see Fig. 8) that represent the number of arms inside the region according to the number of pixels.

As it is also possible that the person is moving his/her arms forward in the region between the robot and the person, the system also analyzes the distance between each of the person's pixel to his/her mass center. In this way is possible to analyze the number of pixels that are not close to the mass center and that could potentially be part of an extended arm. We set the linguistic variable $ExtendedArm$ to represent this situation. This variable can also take three values represented by the labels "Zero", "One" and "Two" (see Fig.8) that represent the number of arms inside this region according to the number of pixels.

To analyze whether a person is moving an arm instead of only raising or extending it, the system analyzes the number of pixels in the last frames building an image made of the pixels existing in the elliptic area (for a raised arm) or in front of the person (for an extended arm) in the last frames. If a person is moving one arm in that area, the number of pixels in the last frames should be higher (double or triple) than for one arm that is only raised or extended. Analogously, we set two linguistic variables $MovingRaisedArm$ and $MovingExtendedArm$ that can take also values "Zero", "One" and "Two" (see Fig. 9) according to the number of arms given by the number of pixels in both regions explained before.

This information is given to two parallel fuzzy systems that compute the level of attention demand of each person. The first one will compute the level of attention demand using only the information about the number of raised and extended arms that are not moving (based only in position) while the second one will do the same using the information about moving arms in both regions. The output from the first fuzzy system is called RAP (Requiring Attention Position) while the output from the second is called RAM (Requiring Attention Moving). These linguistic variables are represented in Fig.8 and Fig.9. The rule base for each one of the two fuzzy system is represented in Fig. 8(d) and Fig. 9(d). The fuzzy variable $RAfuzzy$ (Requesting Attention fuzzy) takes the maximum value of variables RAP and RAM:

$$RAfuzzy = max(RAP, RAM)$$

The defuzzyfied value of $RAfuzzy$ will belong to $[0, 1]$ interval and means the instantaneous level of attention demand. This value is weighted with past observation in a similar way than the shown in Section III-A.

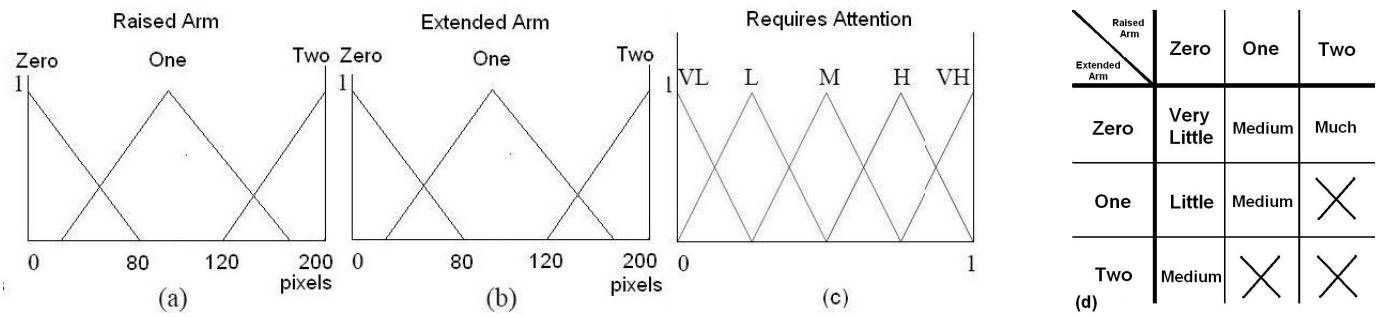

Fig. 8. Fuzzy sets of the linguistic variables: (a) Raised Arm (b) Extended Arm (c) Requires Attention Position (d) Rule base of the fuzzy system

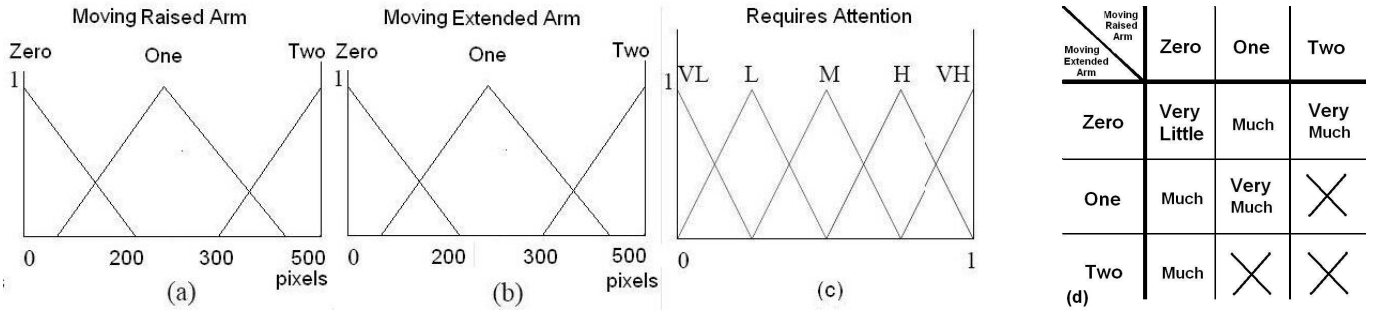

Fig. 9. Fuzzy sets of the linguistic variables: (a) Moving Raised Arm (b) Moving Extended Arm (c) Requires Attention Movement (d) Rule base of the fuzzy system

B. Shaking or nodding of the head

When communicating with people, head shaking or nodding is normally used to express agreement and disagreement with others. Similarly, in an interaction between a robot and a person, it is important to detect if the person is agreeing or disagreeing with respect to some statement or situation during the interaction process. Speech recognition could be used to detect this kind of situation, but as people tend to shake and nod the head while they speak, it could be a very precious feature to recognize and add the detection of this kind of gestures.

To detect this kind of situation the system uses the area of the face given by the face detector. After obtaining the face region, it is applied a Sobel filter to extract the gradient of the face. Then it is possible to analyze the direction in which the face has moved from the previous to the current frame. For that purpose, we compare whether the region of the current face image has moved to the surrounding pixels in the four main directions (up, down, left and right). To achieve this, the system computes the difference between the previous and current image gradients (see the following equation).

$$SD = \sum_{i=0}^{n} \sum_{j=0}^{m} |p(i,j)_t - p(i,j)_{(t-1)}| \qquad (2)$$

where n and m are the width and height of the face image and p(i,j) the gray level of pixel i,j.

Equation 2 is used to compare the current image with the previous image moved by a variable offset. After some experiments we realized that comparing images moved in the four main directions until a maximum offset value of 5 pixels is enough to detect head shaking and heading.

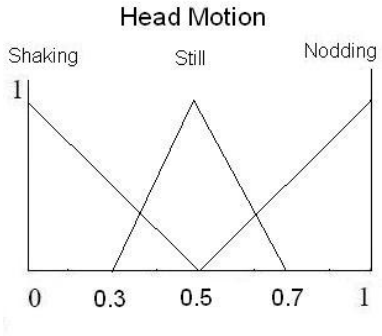

Fig. 10. Labels for variable Head Motion.

At the end, it is chosen the one that has the smallest value, indicating that it is the image most similar to the current one. Therefore it is possible to know what is the direction (if any) that the face has moved and how many pixels has it moved (the speed at which the person is moving his/her face).

For each frame the system estimates in real time the direction of the head in one of the five categories: up (U), if the direction that had the smallest error was the upper direction, in down (D) if it was the down direction, in left (L) if it was the left direction, in right (R) if it was the right direction and in still (S) if it was not none of the above. This output is translated into a numerical value by the definition of the

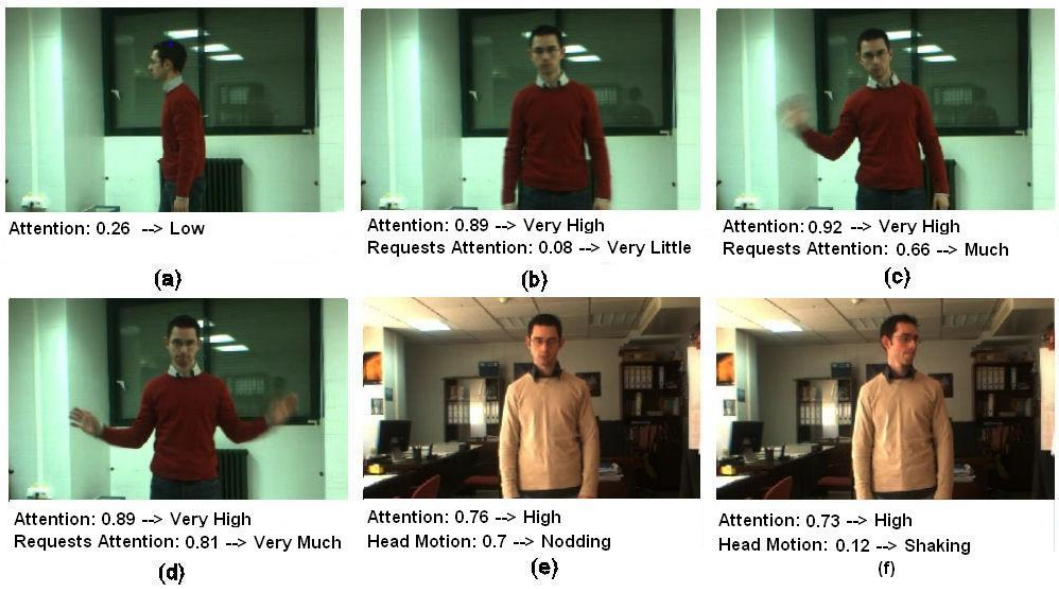

Fig. 11. Examples from the first video: Low interest (a), Highly interested and not moving the arm (b),Highly interested and moving one hand (c), Highly interested and moving both hands (d), Nodding the head (e), Shaking the head (f)

variable $HMcurr_t$. The value of $HMcurr_t$ in the time t is

$$HMcurr_t = \begin{cases} 1 & \text{if output = “U” or “D”;} \\ 0.5 & \text{if output = “S”;} \\ 0 & \text{if output = “L” or “R”.} \end{cases}$$

However, $HMcurr_t$ is an instantaneous value that does not take into account past observations. In order to consider past observations, we define the variable $HM_{(t)}$ as:

$$HM_{(t)} = \alpha HM_{(t-1)} + (1-\alpha)HMcurr_t \qquad (3)$$

where the initial value for HM is the first value of $HMcurr_t$ when the person is detected and α is a weighting factor that ponders the influence of past observations.

To deal with the uncertainty and vagueness in this process we use a linguistic variable called “Head Motion” and divide it into “Shaking”, “Still” and “Nodding” values (see Fig. 10). This variable will take as input values the measures of head motion estimation considered by HM_t (Eq. 3) so that values near to 0 mean a shaking and values near to 1 mean a nodding. In future works this fuzzy variable can be used to facilitate the communication with the people.

V. Experimentation

Several experiments have been done to validate our system. The results were very satisfactory in respect to interest estimation, attention demand and shaking and nodding estimation using our system. In this section we will describe two of the experiments. To perform the stereo process we have used images of size $320x240$ and sub-pixel interpolation to enhance the precision in the stereo calculation. The operation frequency of our system is about 30 Hz without considering the time required for stereo computation.

Regarding the interest estimation, we have checked that the interest degree assigned to each tracked person increases and decreases dynamically accordingly to the behavior of the person in relation to the robot, i.e., it depends on whether the person is looking at the robot, and on the distance from the robot, and on whether it is in front of it.

Fig.11 shows our first experiment with one person. In Fig.11(a) that the person is not looking at the camera and the level of interest is low while in the other frames the interest is higher. In frame Fig.11(b) the person has interest but is not requesting attention because is not doing any movement with the arms. In frame Fig.11(c) the person is demanding attention because it is moving one of the arms. In frame Fig.11(d) the person is moving not only one arm but two making the value of attention demand increase to higher values.

In respect to head shaking and nodding, it was possible to check that the system determined the movement most of times. It was possible to check that head shaking was detected more accurately than nodding as we can see by the results (around 86% accuracy to head shaking while head nodding was about 83%). Fig. 11 shows in frames (e) and (f) examples of head nodding and head shaking and the output of the system for the first experiment.

With respect to the second experiment two people were used to test the system as seen in Fig. 12. In this experiment one of the people did not show interest to the robot (the person on the left) while the other one was changing his behavior over the time. The system never examines whether the person is demanding attention or making any movement with his head when his interest towards the robot is less than “High” as we can see in frames (a), (c) and (e). In frame (a) we can see that no one of them was showing interest to the robot. The same situation happens in frames (c) and (e) with different levels of attention for each of them. Between these frames we can see that the person on the right changed his behavior towards the robot. In frame (b) that person was showing interest to

Fig. 12. Examples from the second video: Both people with no interest (a), High interest from person on the right (b), Both people with little interest (c), High interest and demand of attention by the person on the right (d), Both people with no interest (e), High interest from person on the right (f), High demand of attention from person on the right (g), Person on the right nodding his head (h), Person on the right shaking his head (i)

the robot but was not demanding attention neither making any movement with his head, while in frame (d) the same person was requesting the robot's attention by raising his left arm. In frame (f) and (g) we have similar situations to frames (b) and (d). Finally, in frames (h) and (i) we have examples of the person on the right nodding and shaking his head.

In order to better understand the performance of the system, several videos are available in the following web site http://decsai.ugr.es/∼ruipaul/interest.htm.

VI. Conclusions and future work

In this paper we have shown a system for estimating the interest of the people in the surroundings of a mobile robot and detecting motions related to attention demand and head nodding and shaking, using stereo vision, head pose estimation by SVM and fuzzy logic. While a person is being tracked, the fuzzy system computes a level of possibility about the interest that this person has in interacting with the robot. This possibility value is based on the position of the person with respect to the robot, as well as on an estimation of the face attention towards the robot. To examine the face attention we analyze the head pose of the person in real time. This analysis is performed by a view based approach using Support Vector Machines. Thanks to SVM, the head pose can be detected achieving a great percentage of success that is no dependent on the morphological features of the heads. For those people whose interest level was "High" or "Very High" our fuzzy system was also able to detect whenever those people were demanding attention by analyzing the movement of their arms. The system also achieved a good result concerning the detection of head movements such as head nodding and shaking. In the future our efforts will be centered in improving the performance of the system with learning methods that can compute the best values for the variables that are part of the system like fuzzy variables.

Acknowledgment

This work has been partially supported by the Spanish MEC project TIN2006-05565 and Andalusian Regional Government project TIC1670.

References

[1] L. Snidaro, C. Micheloni, and C. Chiavedale, "Video security for ambient intelligence," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 35, pp. 133 – 144, 2005.

[2] M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke, "Integrating vision and speech for conversations with multiple persons," in *IROS'05: Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2005, pp. 2523 – 2528.

[3] W. Song, D. Kim, J. Kim, and Z. Bien, "Visual servoing for a user's mouth with effective intention reading in a wheelchair-based robotic arm," in *ICRA*, 2001, pp. 3662–3667.

[4] S. S. Ghidary, Y. Nakata, H. Saito, M. Hattori, and T. Takamori, "Multimodal interaction of human and home robot in the context of room map generation," *Autonomous Robots*, vol. 13, no. 2, pp. 169–184, 2002.

[5] R. Muñoz-Salinas, E. Aguirre, M. García-Silvente, and A. González, "Un sistema visual difuso para la detección de interés en la interacción robot-persona," in *Actas del VII Workshop de Agentes Físicos*, 2006, pp. 49–56.

[6] O. Holland and G. Walter, "The pioneer of real artificial life," in *Proceedings of the International Workshop on Artificial Life, MIT Press*, Cambridge, MA, pp. 34–44.

[7] R. Beckers and et al., "From local actions to global tasks: Stigmergy and collective robotics," in *Proceedings of Artificial Life IV*, 1996, pp. 181–189.

[8] J. L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, and L. Chrétien, "The dynamics of collective sorting robot-like ants and ant-like robots," in *Proceedings of the first international conference on simulation of adaptive behavior on From animals to animats*. Cambridge, MA, USA: MIT Press, 1990, pp. 356–363.

[9] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonumous Systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[10] C. Breazeal, *Designing Sociable Robots*. MIT Press, 2002.

[11] I. Nourbakhsh, J. Bobenage, S. Grange, R. Lutz, R. Meyer, and A. Soto, "An affective mobile robot educator with a full-time job," *Artificial Intelligence*, vol. 114, no. 1-2, pp. 95–124, 1999.

[12] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, "Towards robotic assistants in nursing homes: Challenges and results," *Special issue on Socially Interactive Robots, Robotics and Autonomous Systems*, vol. 42, no. 3 - 4, pp. 271 – 281, 2003.

[13] B. Scassellati, "Foundations for a theory of mind for a humanoid robot," Ph.D. dissertation, 2001, supervisor-Rodney Brooks.

[14] K. Dautenhahn and A. Billard, "Bringing up robots or - the psychology of socially intelligent robots: from theory to implementation," in *Proceedings of the Third International Conference on Autonomous Agents (Agents'99)*, O. Etzioni, J. P. Müller, and J. M. Bradshaw, Eds. Seattle, WA, USA: ACM Press, 1999, pp. 366–367. [Online]. Available: citeseer.ist.psu.edu/dautenhahn99bringing.html

[15] J. Zlatev, "The epigenesis of meaning in human beings, and possibly in robots," *Minds Mach.*, vol. 11, no. 2, pp. 155–195, 2001.

[16] K. Dautenhahn, "I could be you — the phenomenological dimension of social understanding," *Cybernetics and Systems*, vol. 25, no. 8, pp. 417–453, 1997.

[17] S. Restivo, "Bringing up and booting up: Social theory and the emergence of socially intelligent robot," in *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 4, 2001, pp. 2110–2117.

[18] K. Dautenhahn and B. Ogden, "From embodied to socially embedded agents implications for interaction-aware robots," *Cognitive Systems Research*, vol. 3, pp. 397–428, September 2002.

[19] PtGrey, "Bumblebee. Binocular stereo vision camera system," *http://www.ptgrey.com*, 2005.

[20] R. Muñoz-Salinas, E. Aguirre, and M. García-Silvente, "People detection and tracking using stereo vision and color," *Image and Vision Computing*, vol. 25, no. 6, pp. 995–1007, 2007.

[21] N. Cristianini and J. Shawe-Taylor, *An Introduction To Support Vector Machines (and other Kernel Based Methods)*, C. U. Press, Ed. Cambridge University Press, 2000.

[22] M. Harville, "Stereo person tracking with adaptive plan-view templates of height and occupancy statistics," *Image and Vision Computing*, vol. 2, pp. 127–142, 2004.

[23] I. Haritaoglu, D. Beymer, and M. Flickner, "Ghost 3d: detecting body posture and parts using stereo," in *Workshop on Motion and Video Computing*, 2002, pp. 175 – 180.

[24] K. Hayashi, M. Hashimoto, K. Sumi, and K. Sasakawa, "Multiple-person tracker with a fixed slanting stereo camera," in *6th IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 681–686.

[25] C. Eldershaw and M. Yim, "Motion planning of legged vehicles in an unstructured environment," in *IEEE International Conference on Robotics and Automation (ICRA'2001)*, vol. 4, 2001, pp. 3383 – 3389.

[26] S. Thompson and S. Kagami, "Stereo vision terrain modeling for non-planar mobile robot mapping and navigation," in *IEEE International Conference on Systems, Man and Cybernetics*, vol. 6, 2004, pp. 5392 – 5397.

[27] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[28] R. Lienhart and J. Maydt, "An Extended Set of Haar-Like Features for rapid Object detection," in *IEEE Conf. on Image Processing*, 2002, pp. 900–903.

[29] Intel, *OpenCV: Open source Computer Vision library*, 2005.

[30] D. Comaniciu, V. Ramesh, and P. Meer, "Real-Time Tracking of Non-Rigid Objects using Mean Shift," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2000, pp. 142–149.

[31] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[32] S. Birchfield, "Elliptical Head Tracking Using Intensity Gradients and Color Histograms," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1998, pp. 232–237.

[33] G. Henry and Dunteman, *Principal Components Analysis*, S. Publications, Ed. SAGE Publications, 1989.

[34] C. Chang and C. Lin, "Libsvm. a library for support vector machines," *http://www.csie.ntu.edu.tw/ cjlin/libsvm/*, 2006.