A fuzzy system for visual detection of interest in human-robot interaction

Rafael Muñoz-Salinas, Eugenio Aguirre, Miguel García-Silvente, Antonio González
Department of Computer Science and Artificial Intelligence
E.T.S. Ingeniería Informática. University of Granada.
18071 Granada, SPAIN
email: {salinas, eaguirre, M.Garcia-Silvente, A.Gonzalez}@decsai.ugr.es

***Abstract*— The development of natural human-robot interfaces is necessary to achieve intelligent robots applied to service tasks in environments where humans operate. In that sense, a very useful capability for a robot is to be able to detect the interest that the people in the environment have to interact with it. This knowledge can be used to establish a more natural communication with humans as well as to create an appropriate policy for the assignment of the available resources. In this paper, it is proposed a fuzzy system that establishes a level of possibility about the degree of interest that the people around the robot have in interacting with it. First, a method to detect and track persons using stereo vision is proposed. The method uses a height map of the environment and a face detector besides the Kalman filter to detect and track the persons in the surroundings of the robot. Then, the interest of each person is computed using fuzzy logic by analyzing its position and its level of attention to the robot. The level of attention is estimated by analyzing if the person is looking or not at the robot. Although the proposed system is based only on visual information, its modularity and the use of the fuzzy logic make easy the incorporation in the future of other sources information to estimate with higher precision the interest of the people. At the end of the paper, some experiments are shown that validate the proposal and future work is addressed.**

I. Introduction

The development of successful robots systems applied to service tasks in home and office environments implies the generation of natural human-robot interfaces. In the case of mobile robots, they usually operate in environments where several people are moving around and it is not always clear for the robot which of the persons have really some interest in interacting with it. Thus, it is needed the development of the appropriate techniques that allow a mobile robot to autonomously 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, the robot should be able to detect persons in its vicinity and track their movements over time. This tracking is not an easy task since several persons could be moving at the same time, crossing their trajectories and occluding each others. In the literature we can found many works on this topic [1], [2], [3]. The techniques to perform the detection and tracking are frequently based on the integration of different information sources such as: skin color, face detectors, visual analysis of the motion or laser range finder. Although the people detection and tracking with a single camera is a well explored topic, the use of the stereo technology for this purpose concentrates now an important interest. The availability of commercial hardware to solve the low-level problems of stereo processing, as well as the lower prices for these types of systems, turns them into an appealing sensor with which intelligent systems could be developed. The use of stereo vision provides a higher grade of information that bring several advantages when developing human-robot applications. On one hand, 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 [4]. Furthermore, the possibility to know the distance to the person could be of great assistance for the tracking as well as for a better analysis of their gestures.

Once that the robot is able to recognize and track the persons in its vicinity, then it should be able to detect their interest in establishing an interaction with it. In this task, several types of signals from the human can be taken into account (both verbal and non-verbal). Some authors [5], [6] use sound source localization or speech recognition besides visual perception to detect which persons are the most interested. In other cases, facial expressions [7] or hand gestures [8] are analyzed. Finally, other authors [9] propose the use of non-verbal signals present in physiological monitoring systems that include skin conductance, heart rate, pupil dilation and brain and muscle neural activity.

In this work we present a fuzzy system able to establish a possibility value on the interest of a person to interact with a mobile robot. First, we have developed an module to detect and track people through stereo vision. The detection method is based on the initial creation of a height map of the environment. This map contains information about the structure of the environment and could even be created while people are moving around. Using this structural map, it will be possible to detect the objects in motion. These are potential candidates to be persons that will be detected through a face detector. Once a person has been detected, the Kalman filter [10] is used to keep track of its position. While the person is being tracked, a fuzzy rule based system analyzes its position in the environment and its degree of attention by analyzing if the person is looking at the camera. The fuzzy system generates a possibility value on the interest of the person to interact with the robot.

A. Related works

Among the most important projects related to people detection and tracking using stereo vision, we find the work of Darrel et al [11]. This paper presents an interactive display system capable of identifying and tracking several people. The detection of people is based on the integration of the information provided by a skin detector, face detector and the map of disparity of the environment. On one hand, independent objects (*blobs*) are detected on the disparity image that will be candidates to people. On the other hand, the color of the image is analyzed to identify those areas that could be related to skin. And finally, a face detector is applied. These three items are merged in order to detect the visible people. To perform the tracking, information on hair color, clothes and past history of the located people are used. In this way, the people can be identified even though they disappear from the image for a while.

In [12] a method to locate and track people in stereo images is presented using occupancy maps. Before the people detection process takes place, an image of the environment is created through a sophisticated image analysis method. Once the background image is created, the objects that do not belong to it are easily isolated, and both an occupancy map and a height map are built. The information from both maps is merged to detect people through the use of simple heuristics. The people tracking is performed by using a Kalman filter combined with deformable templates. In this work, a stereoscopic system is used and it is located three meters above the ground, on a fixed position.

On the majority of the works, elevated positions of the cameras are used [12], [13]. However, on some other papers that seek the interaction with the user, the position of the camera is usually lower than the height of the person [11], [14]. Besides improving the visibility of the face and arms of the person, these methods are more adequate for their implementation in robotic systems that require interacting with people. Studies performed show that in order to improve the acceptance of the robots by the humans it is important that they are of less height than the later [15]. Otherwise the person could feel intimidated.

In [5] a multi-modal attention system is shown. This approach uses a pan-tilt camera for face recognition, two microphones for sound source localization and a laser range finder for leg detection. Shifting attention is carried out by turning the camera towards the person which is currently speaking. In [6] the authors present a system that makes use of visual perception, sound source localization and speech recognition to detect, track and involve people into interaction. In [6] the goal is that the robot interacts with multiple persons and does not focus its attention on only one single person.

In regard to the role of the Fuzzy Logic [16] in the robotics, in [17] an extensive catalogue of the uses of fuzzy logic in autonomous robots can be found. Fuzzy logic has been applied to design control behaviors for navigation, the coordination of these behaviors, the building of maps of the environment and to achieve a suitable integration of deliberative and reactive layers. Lately, fuzzy logic have been also applied to the area of the human-robot interaction. In [18] several soft computing techniques are applied to service robotic systems for comfortable interaction and safe operation. Fuzzy logic is used for recognizing facial emotional expression and for coordinating bio-signals with robotic motions. In [9] several sets of fuzzy rules are used for estimating intent based on physiological signals.

In this work we are interested in computing a value of possibility of the interest of a person to interact with the robot. This value is computed only taking into account visual information, but the modularity of the system makes easy the posterior incorporation of other types of input data as sound or laser range finder. The interest is computed according to the position of the person and its degree of attention. The people detection and tracking is performed by a stereoscopic system located at inferior levels from the people's height. In order to ease the processing, a model of the environment is built in a previous phase. A distinguished characteristic of the map building process is that it can be created even in the presence of movable objects. The use of this map will allow us to easily detect the objects that do not belong to the environment and narrow the people detection process to only those objects. The reduction of the information to be analyzed will enable us, besides to reduce the computer costs, to eliminate false positives produced by the face detector used.

The remainder of this paper is structured as follows. Section II gives an general overview of the hardware and software system. Section III describes the method employed for background modelling and foreground segmentation. Section IV shows how the detection and tracking of people in the surroundings of the robot is performed. In Section V it is explained the fuzzy system for estimating the interest of the people in interacting with the robot. In Section VI it is shown the experimentation carried out, and finally, Section VII outlines some conclusions and future works.

II. System Overview

A. Hardware system

The hardware system is comprised by a laptop to process the information, a stereoscopic system with 2 cameras [19] and a Nomad 200 [20] mobile robot.

The use of our stereoscopic system enables us to capture two images from slightly different positions (stereo pair) and to create a disparity image I_d. Knowing the internal parameters of the stereoscopic system it is feasible to estimate the three-dimensional position p_{cam} of a point in I_d. These points are translated to a reference static system that has as center the center of the robot at the ground level through the Equation 1. The T transformation matrix is created by using the intrinsic parameters of the system (provided by the manufacturer) and extrinsic ones that are previously estimated.

$$p_w = T p_{cam} \tag{1}$$

B. Detection, tracking and interest estimation process

The process of people detection and tracking besides the determination of the interest of a person in interacting, is divided in two phases.

In a first phase, a map of the environment $\mathcal{H}_{max}$ is created (background creation) that registers the position of the motionless objects. This map divides the environment into cells of a fixed size and indicates on each one of them the maximum height of the detected objects. This map will be employed to easily detect in the following images the objects that move in the environment (foreground extraction).

Once the environment has been registered, in the second phase, the system starts an iterative process consisting in: (i) detecting new people that could enter in the scene, (ii) tracking the people that have already been detected in previous instants and finally (iii) estimating their interest in interacting with the robot.

The second phase starts by creating an instantaneous occupancy map $\mathcal{O}$. On this map we will be able to identify those objects that are in the scene but were not registered as motionless objects in $\mathcal{H}_{max}$, in other words, those objects that are in motion. The objects present in $\mathcal{O}$ are identified and analyzed to determine which of them are people. For this purpose we have applied a face detector [21] on one of the camera's images. The false positives generated by the face detector will be rejected thanks to the integration of the information of the disparity image and $\mathcal{O}$. Once an object has been detected as a person, the system must keep track of this person. An important point to have into account is that several persons can be detected by the system in the same scene. The tracking problem consists in detecting, in the following scenes, which is the object with higher probability of being each person. In order to achieve a robust tracking, we employ the Kalman filter to predict the new position of each person using a linear movement model and considering the position and velocity of the person within the state vector in the Kalman filter.

The next step consists in the determination of the possibility of interest of each tracked person to interact with the robot. In this proposal we are going to define the interest depending on the position and the face attention of each person. We are conscientious that actually the interest of interact is a more complex topic and that other information sources would be necessary, for example speech recognition, facial expressions or gesture analysis. However, thanks to the flexibility of the fuzzy logic it is easy the posterior incorporation of more sources of information (like sound or a gesture analysis) to estimate with higher precision the degree of interest of a person in interacting with the robot.

In the following sections the more relevant processes previously mentioned will be elaborated in more detail.

III. BACKGROUND CREATION AND FOREGROUND EXTRACTION

Before 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 it and then we will be able to consider 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 [22], [23] in it. The points identified by the stereo system p_w are projected over $\mathcal{H}_{max}$, that stores the maximum height of the projected points in each cell. To avoid adding the points of the ceiling on $\mathcal{H}_{max}$, those that overcome the height threshold h_{max} are excluded from the process. Due to efficiency reasons for the calculation, 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 will be fused using the median operator. With this method we can built the height map even in the presence of moving objects in the environment.

On Figure 1 we can observe the creation of the height map of an environment. In this example the map has been created in presence of the two persons moving in the environment. Figure 1(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 can begin. The first step consists in creating an occupancy map $\mathcal{O}$, which will indicate on each cell the surface occupied by the objects that do not belong to the environment ($\mathcal{H}_{max}$). For this purpose, after capturing a stereo pair of the environment, the position of the points detected p_w is calculated. For each point p_w 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 the registered point. Points detected far from the camera increment the corresponding cell with a higher value than nearer cells. It compensates the change in size observed for the same object when it is seen at different distances and makes the weight of its projection independent of the distance.

When $\mathcal{O}$ is created, it is analyzed to detect the objects that appear on it. On a first step, a closing process takes place with the purpose to link possible discontinuities on the objects. After this, the objects are detected by grouping cells that are connected and that their sum of areas overcomes the threshold θ_{min}. On this way, we eliminate the potential noise that appears as a consequence of the stereoscopic process. On Figure 1(c) we can observe the occupancy map $\mathcal{O}$ of the environment on 1(a) using a height map $\mathcal{H}_{max}$ from 1(b). The darker values represent the areas with higher occupancy

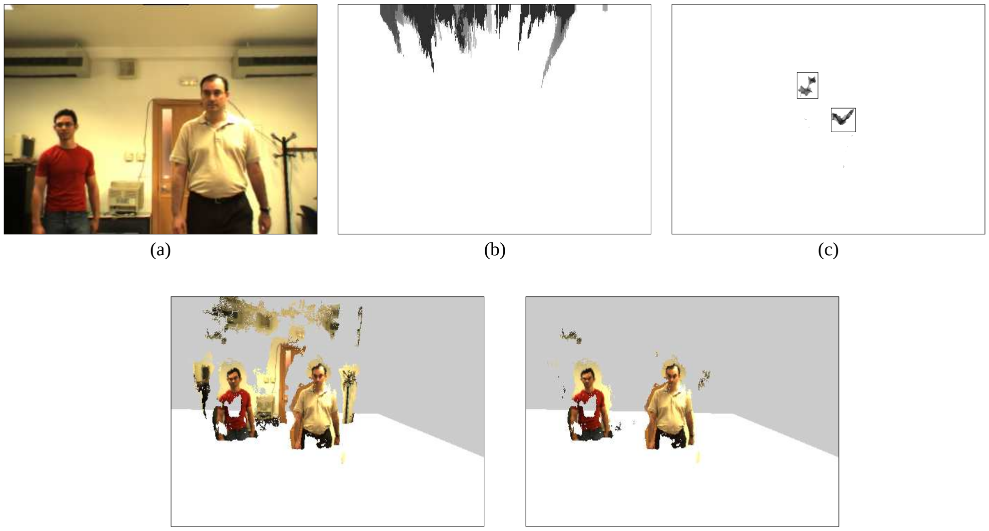


Fig. 1. (a) Image of our environment containing two people. (b) Background map $\mathcal{H}_{max}$ of our environment created in a previous phase. (c) Occupancy map $\mathcal{O}$ of the Image *a*. There can be seen the two objects detected. (d) 3D reconstruction of the scene including both background and foreground points. (e) 3D reconstruction of the background points, i.e., those projected in $\mathcal{O}$.

density. On Figure 1(c) we can see bounded with frames the two objects detected after the closing process, grouping and thresholding. Finally, Figures 1(d) and 1(e) shows the 3D reconstruction of the scene in Figure 1(a). While in Figure 1(d) it is shown all the points detected by the stereo system, in Figure 1(e) are only drawn those belonging to the foreground, i.e., those used for voting in $\mathcal{O}$.

IV. People Detection and Tracking

Our approach for people detection and tracking is based on the assumption that if a person is visible (as in Figure 1(a)), it is detected as an object by the previous process (Figure 1(c)). Nevertheless, it is important to notice that although every person is projected as an object, an object does not always represents a person but other things that move around in the environment. Therefore, it is necessary to analyze if an object detected is a person or not.

Our approach consists in analyzing if the object contains a human face, and if so, it is assumed that this object represents a person. Nevertheless, if it does not contain a human face, we can not say that it is not a person. It could be a person whose face is not visible yet, but it could be visible later.

Once an object has been classified as a person on time t, it is not necessary to apply the face detector on it on future times. We only need to know where this object has moved in the following scene. For that reason, we proceed to track this object. We only apply the face detector on these objects that have not been recognized as people yet.

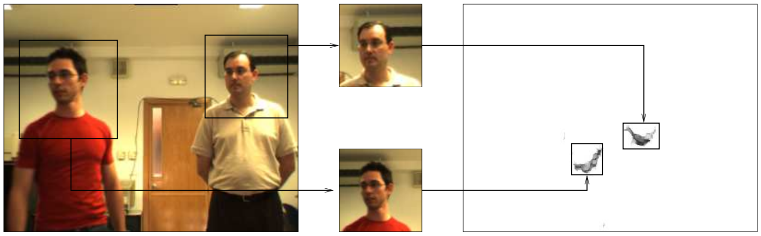
The process of analyzing if a object is a person is what we call people detection. The process of knowing where an object that is a person has moved is called tracking. Next we explain these processes in detail.

A. People Detection

People detection has been performed in different ways in the literature. The main two approaches are either to consider a person as an object in the occupancy map $\mathcal{O}$ with sufficient weight [13], [12] or to look for people's faces in the image [11], [14]. The first approach is most commonly used when the camera is placed at elevated positions because the sight of the face is usually more difficult to detect. We have opted for using a face detector on the image because it is a powerful indicator of the presence of a person and our camera position allows it.

Face detection is a process that can be time consuming if applied on the entire image. To avoid that, we estimate the region in the camera image where the head should be if the object is a person (the upper part). As the human head has an average width and height, we analyze if the upper part of the object have similar dimensions. Because of we have the depth information, we can perform this verification independently of the person's distance to the camera. If the object does not pass this test, it is rejected as possible person. We perform this test in a flexible manner so that people with different morphologic characteristics can pass it. If the upper part of the object has a dimension similar to the typical human's head, the corresponding region of the image (slightly enlarged) is extracted and analyzed to detect if it contains a face. For that purpose, we have used a face detector [24] that can detect both frontal and lateral views of human faces in grey level images.

Figure 2(a) shows a scene where there are two people that have entered in the environment. In that scene there are detected two objects (see Figure 2(c)). The portion of the images where their heads should be is calculated using the information about the height of the object and its distance to the camera. Figure 2(b) shows these images that are analyzed



(a) Camera Image (b) Faces Detected (c) Occupancy map O

Fig. 2. (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

using the face detector. In that case, it is detected that both are persons. If a face is detected in a object in the time instant t, it is assumed that this object represents a person in the future. Therefore, it will not be necessary to use the face detector on the object that represents this person in the following images, it is only necessary to know where this object has moved. This is the main idea behind the tracking process explained below.

Besides the reduction in computing time, we have observed that the method proposed in this section greatly reduces the false positives of the face detector. Nevertheless, the face detector remains having many false negatives, specially at far distances (above 3 meters).

B. People Tracking

Once an object has been detected as a person by the procedure explained before, we must keep track of this person. In order to achieve a robust tracking, we employ the Kalman filter to predict the next position of a person using information about its movement in the last scenes. The state vector used by the Kalman filter to track each person is $\vec{s} = (x_p, y_p, v_x, v_y)$, where the values (x_p, y_p) represents the position of the person. The pair (v_x, v_y) represent the velocity of the person in the x and y axis of the occupancy map respectively. We use as system model a first order movement model:

$$x(t+1) = x(t) + v_x;\ \ y(t+1) = y(t) + v_y$$

Kalman is used to predict the new position of each person $\vec{s}_{pred}$ and it is used to analyze which one of the objects in the current scene has a higher probability of being that person. For that purpose we employ the Equation 2, that computes a value in the range $[0, 1]$ indicating the probability that the object o_i^t corresponds to a tracked person. The Equation measures the distance between the positions and heights of the object and the prediction for the person. Values near to 1 indicates high probability and vice versa. The parameters σ_x and σ_y, indicates the uncertainty associated to each one of the values (x, y) and are given by the Kalman filter prior uncertainty matrix. For each person, Equation 2 is calculated to detect the object that best matches its position. The method assumes that the object with higher value of Equation 2 is that person.

$$S(o_i^t, \vec{s}_{pred}) = e^{-\left(\frac{(x_o - x_p)^2}{2\sigma_x^2} + \frac{(y_o - y_p)^2}{2\sigma_y^2}\right)}. \qquad (2)$$

In order to obtain a robust tracker we must deal with occlusions. Partial occlusion is dealt by our method; if the person is partially visible, it will be detected as an object if the weight in $\mathcal{O}$ of its visible points is higher than θ_{min}. When a person is almost totally occluded (the weight of its visible points is below θ_{min}) or it is not visible at all, it is not detected an object for this person in $\mathcal{O}$. In that case, we keep predicting its position using the Kalman filter and still looking for it for a maximum number of times. If the person remain unseen for too long, we assume that the person has definitively leaves the scene and the system stop tracking him.

V. Interest Detection

In the previous sections we have described as the robot is able to detect the persons in its vicinity by using the stereo system and track the detected persons. In this section, fuzzy logic is used to define a possibility of interest for the detected persons to interact with the robot. The advantages of the use of fuzzy logic are mainly three. First, the robot has to deal with information from the stereo system that is affected by the uncertainty and the vagueness and fuzzy logic is a good tool to manage them using linguistic variables. Second, the human knowledge can be usually expressed as rules. Fuzzy logic allows to establish relationships among the variables through fuzzy rules giving also the inference mechanism. Finally, there exist methods in fuzzy logic to fuse the results from each fuzzy rules set in order to achieve a final overall result. Therefore, in this paper we can build a fuzzy rules set to compute the possible interest based only on visual information and in future works, other fuzzy rules sets will be added using other types of information as source sound localization, gesture analysis or speech recognition systems.

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 in 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 $Angle$. These linguistic variables have three possible values each of them, that are shown by Figure 4(a-b). The sense of these two variables is the following: 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

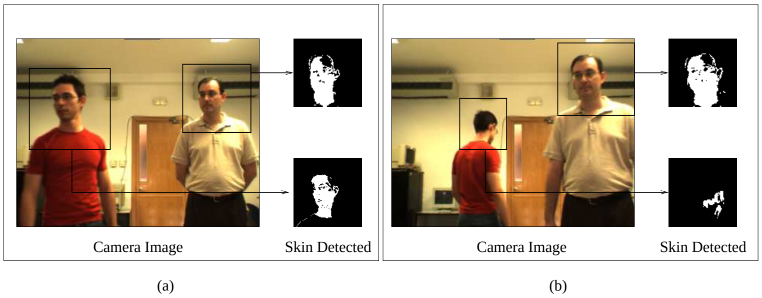


Fig. 3. Two examples of estimating face attention (a) Scene with two people looking a the camera (b) Scene in which the person on the left turns backwards while the person on the right is looking at the camera

establish the interest of a person in interacting with the robot. The third feature shown in this paper is the attention detected by the analysis of the face of the person. This analysis could be done in multiple ways but we have solve it by detecting the amount of skin in the head of the person. If the amount of skin detected is too low we can determine that the person is backwards and thus assign a low degree of attention than when a large amount of skin is found. We have employed this approach because it is not time-consuming and the results have shown to be good according to our experimentation. Following, this third feature is explained in detail.

A. Estimating face attention using color

One of the most prominent cues to detect if a person is paying attention to the system is the orientation of his face, i.e., a higher degree of attention can be assumed when a person is looking at the system than when it is backwards. For that purpose we estimate if the amount of skin visible in the head of a person corresponds to the amount of skin expected for a subject that is looking at the camera. We assume that if a person looks at the camera, the amount of skin in the head must be higher than when the person is backwards. Of course, it is only true if the person is not bald. We have opted for that heuristic instead of relying on the face detector because it is very prone to false negatives. In the future we plan to use more sophisticated methods to detect the orientation of the head that cover a broader range of the population.

The first step to detect the amount of skin visible for a person is to use a skin color detector. There have been used several techniques for skin color detection in the literature [25]. Most of them use illumination invariant color spaces and require an initial training phase in which the skin of several people is analyzed under different illumination conditions. Nevertheless, the automatic white balance of many cameras (like ours) makes very difficult and tedious this approach. For these reasons, we have opted for dynamically creating the skin color model of each person and to update the color model periodically to adapt it to the changing illumination conditions of the real environments. We model the skin color in the HS plane of the HSV color space in order to make more robust against brusque illumination changes.

The skin color model is created when the face of a person is detected at near distances (less than $2,5$ meters). When it is detected that there is a face in an object (as explained in Section IV-A), we analyze the color of the three-dimensional points detected in the region indicated by the face detector and calculate the median of its color distribution S_m. To avoid the shading and saturation effects, points too dark and too bright are discarded. The skin color segmentation is performed by analyzing if the distance of a color to S_m is below a fixed threshold.

In order to know if a person is looking at the camera, we measure in the images captured the degree in which the amount of skin found in the upper part of a person (belonging to its face and neck) corresponds to the amount of skin of a person who is looking at the camera. For that purpose, we have previously estimated the area of the visible skin of different people when they look at the camera (A_f). Using the knowledge about the intrinsic parameters of the camera and the distance of a person to it, we can calculate the area in the camera image A_{img} that the real area A_f must occupy, i.e., the value A_{img} is the number of pixels (in the camera image) that should be visible if the person were directly looking at the camera.

To perform the comparison, we first determine the region of the image that contains the head of a person (slightly enlarged) and threshold it using its skin color model. Then, we count the number of pixels in that region that belong to skin (N_{skin}). Finally, using Equation 3 we estimate the degree in which the amount of visible skin corresponds to the expected visible skin for a person that directly looks at the camera. The parameter $FaceAttention$ is in the range $[0, 1]$; Values near to 1 indicates that the amount of skin found is similar to the expected if the person were looking at the camera. Values near to 0 indicates that the person is backwards. Due to the error we commit in estimating the distance of a person to the camera and to the fact that we use the same value A_f for all the persons, the expected maximum skin area A_{img} might be slightly different from the real value. For that reasons, the fraction $\frac{A_{img}}{N_{skin}}$ could exceeds the value 1. We must therefore treat this value as a indicator of the amount of face visible but subject to uncertainty that is dealt with using fuzzy logic. The linguistic variable to manage this indicator is the variable $Attention$ that has three possible values shown by Figure 4(c). The linguistic variable $Attention$ takes as input the value of

$FaceAttention$ that is computed by Equation 3.

$$FaceAtention = \begin{cases} \frac{A_{img}}{N_{skin}} & if \ \frac{A_{img}}{N_{skin}} \leq 1 \\ 1 & otherwise \end{cases} \quad (3)$$

Figure 3 shows an example of the process explained. The Figure 3 shows two scenes with two people being tracked in it. At the right side of each camera image, it is shown the thresholded images using the skin color filter for each person. As we can observe, in Figure 3(a) both people are looking at the camera and thus we found a high amount of skin in their corresponding head images. In that case, their $FaceAttention$ values are near to 1. However in Figure 3(b) one person is nearly backwards and therefore there are less skin pixels in his corresponding image.

B. Fuzzy System for interest estimation

Once the three linguistic variables have been defined, the rules bases are explained in this section. The idea that governs the definition of the rules base is dominated by the value of the variable $Attention$. If the attention has an high value 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 decrease but like in the former case depending on the distance and angle. Finally if the attention is low, it means that the person has turned backwards 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 $Interest$ that has five possible values shown by Figure 4(d).

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

TABLE I

RULES IN THE CASE OF HIGH ATTENTION.

Finally 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 on the $[0, 1]$ interval. This information on the interest based on visual data could be fused with information supplied by other perceptual systems. In order to validate the performance of the system, in this work the overall output fuzzy set is defuzzifyed to obtain a value belonging to $[0, 1]$ interval. Therefore values near to 1 mean a high level of interest and vice versa.

VI. EXPERIMENTATION

During the explanation of the model we have shown examples of its performance. A broader experimentation has been done but we are unable to show it with images due to space reasons, although we will briefly explain. This experimentation refers to the detection, tracking and interest estimation of different people under different illumination conditions and different distances from the vision system. 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 4 Hz. The half of the time is dedicated to image capture and stereo computation (about 120 ms) and the rest of the processing time to detection and tracking (about 100 ms). The computing time could substantially be decreased if it were feasible to optimize the code of the stereo process or employing specific hardware for depth computation.

We have proven the accurate performance of the people detection method that satisfactorily eliminates the false positives produced by the face detector. The more appropriate distances to detect people vary within $0, 5$ and 2 meters. However, once a person has been located, it can be tracked up to distances of 5 meters.

In order to check the performance of the tracking procedure, there have been designed different tests. The system is able to successfully track a maximum of three people moving freely in the environment. Nevertheless, it is well known that the Kalman filter can fail when there is not enough information to correctly associate the observation with the estimation. It might happen when two people becomes too closed each others. We are currently working in using color information techniques to solve that problem. On the other hand, we have checked that a higher number of persons makes the system get confused because of the excessive total occlusion that takes place. It is due to the low position of the camera that makes impossible that more than three persons appear in the image without being occluded.

Regarding the interest estimation, we have checked the interest degree assigned to each tracked person increases and decreases dynamically accordingly to the behavior of the person in relation to the robot. When the person approaches to the robot looking at it we obtains the higher degree of interest. If the person is at near distances from the robot and turns back, the interest decreases drastically. As the person goes away, the level of interest decreases gradually. In order to better understand the performance of the system, several videos are available in the following web site http://decsai.ugr.es/∼salinas/humanrobot.htm.

VII. CONCLUSIONS AND FUTURE WORK

In this paper we have shown a system for detecting, tracking and estimating the interest of the people in the surroundings of

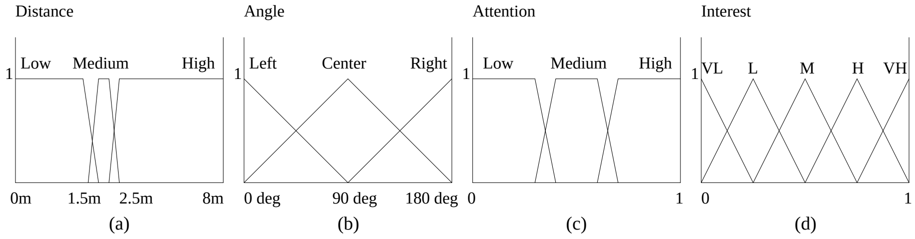


Fig. 4. Fuzzy sets of the linguistic variables: (a) Distance (b) Angle (c) Attention (d) Interest

a mobile robots, using stereo vision and fuzzy logic. As first step, it is necessary to solve the problem of people detection and its tracking. The proposed method initially creates a height map of the environment that registers its motionless characteristics. This map is later used to identify the movable objects in the environment and to search among them potential people by using the face detector. Once a person has been detected, the system can keep track of him/her as well as of the rest people detected using the Kalman filter. 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 in relation with the robot, as well as on an estimation of the attention that the person pays to the robot. To examine the attention that a person pays to the robot we analyze if he/she is looking at the camera of our robot. The experimentation shows that the system is able to detect the persons present in its vicinity, track their motions and give a value of possible interest on the interaction of the persons with the robot. As future work, some problems in the tracking through Kalman filter will be solved using color information. Additionally the proposed method can be easily updated in future works to analyze other types of input data as sounds or laser range finder. Also, the degree of interest will be useful to plan the actions of the robots towards the persons in order to allow a more natural human-robot interaction.

ACKNOWLEDGMENT

This work has been partially supported by the Spanish MEC project TIC2003-04900.

REFERENCES

[1] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, and G. Sagerer, "Multi-modal anchoring for human-robot interaction," *Robotics and Autonomous Systems*, vol. 43, no. 2-3, pp. 133–147, 2003.

[2] W. Liang, H. Weiming, and L. Tieniu, "Recent developments in human motion analysis," *Pattern Recognition*, vol. 36, pp. 585–601, 2003.

[3] 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.

[4] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb, "Plan-view trajectory estimation with dense stereo background models," in *Eighth IEEE International Conference on Computer Vision (ICCV 2001)*, vol. 2, 2001, pp. 628 – 635.

[5] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, "Providing the basis for human-robot-interaction: a multi-modal attention system for a mobile robot," in *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2003, pp. 28–35.

[6] M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke, "Multimodal conversation between a humanoid robot and multiple persons," in *AAAI'05: Proceedings of the Workshop On Modular Construction of Human-Like Intelligence*, 2005, p. To appear.

[7] 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.

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

[9] D. Kulic and E. Croft, "Estimating intent for human robot interaction," in *International Conference on Advanced Robotics*, 2003, pp. 810–815.

[10] M. Grewal and A. Andrews, *Kalman Filtering. Theory and Practice.* Prentice Hall, 1993.

[11] T. Darrell, G. Gordon, M. Harville, and J. Woodfill, "Integrated Person Tracking Using Stereo, Color, and Pattern Detection," *Int. Journ. Computer Vision*, vol. 37, pp. 175–185, 2000.

[12] 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.

[13] 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.

[14] D. Grest and R. Koch, "Realtime multi-camera person tracking for immersive environments," in *IEEE 6th Workshop on Multimedia Signal Processing*, 2004, pp. 387–390.

[15] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, pp. 143–166, 2003.

[16] L. Zadeh, "The concept of linguistic variable and its applications to approximate reasoning," *Part I Information Sciences vol. 8, pages 199-249, Part II Information Sciences vol. 8, pages 301-357, Part III Information Sciences vol. 9, pages 43-80*, 1975.

[17] A. Saffiotti, "The uses of fuzzy logic in autonomous robot navigation," *Soft Computing*, vol. 1, pp. 180–197, 1997.

[18] Z. Bien and W. Song, "Blend of soft computing techniques for effective human-machine interaction in service robotic systems," *Fuzzy Sets and Systems*, vol. 134, no. 1, pp. 5–25, 2003.

[19] PtGrey, "Bumblebee," *http://www.ptgrey.com/products/bumblebee/index.html*.

[20] N. Technologies, *User's Manual*, 1995.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] Y. Ming-Hsuan, D. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 34 – 58, 2002.