People detection and tracking through stereo vision for human-robot interaction.

Rafael Muñoz-Salinas, Eugenio Aguirre,
Miguel Garcia-Silvente, Antonio Gonzalez
{salinas, eaguirre, M.Garcia-Silvente, A.Gonzalez}@decsai.ugr.es

Depto. de Ciencias de la Computacion e Inteligencia Artificial.
E.T.S. Ingenieria Informatica University of Granada- 18071 Granada, Spain

Abstract. In this document we present an agent for people detection and tracking through stereo vision. The agent makes use of the active vision to perform the people tracking with a robotic head on which the vision system is installed. Initially, a map of the surrounding environment is created including its motionless characteristics. This map will be later on used to detect objects in motion and to search among them people by using a face detector. Once a person has been spotted, the agent is capable of tracking them through the robotic head that allows the stereo system to rotate. In order to achieve a robust tracking we have used the Kalman filter. The agent focuses on the person at all times by framing their head and arms on the image. This task could be used by other agents that might need to analyze gestures and expressions of potential application users in order to facilitate the robot-human interaction.

1 Introduction

One critical aspect of the creation of certain intelligent systems is to detect the human presence. The topic human-machine interaction has drawn a lot of attention in the last decade. The objective is to be able to create more intelligent interfaces capable of extracting information about the context or about the actions to be performed through the natural interaction with the user, for example through their gestures or voice.

One fundamental aspect in this sense is the people detection and tracking, with plenty existing literature about this topic [3, 9, 10]. The techniques to perform the detection are frequently based on the integration of different information sources such as: skin color, face detectors, motion analysis, etc.

Although the people detection and tracking with an 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 resolve low-level problems with stereoscopic cameras, as well as lower prices for these types of systems, turns them into and 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-machine 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 environment estimation (background). 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.

On this research we present an agent able to detect and track people through stereo vision. The agent uses active vision to perform the tracking using a robotic head on which the vision system is installed. This agent will serve as a base for the work of other agents in charge of tasks such as gesture analysis and expressions of potential users. 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 possible to detect the objects in motion. These are potential candidates to people that would be detected through a face detector. Unlike other works that only map the environment for a static camera, our map covers the entire visible region by the stereo system. Once a person has been detected, the robotic heads allows the stereo system to rotate in order to follow them through the environment. In order to have a robust tracking we have used the Kalman filter. The tracking method is designed to maintain visible, as long as feasible, the head and arms of the person and therefore facilitate the gesture analysis.

1.1 Related works

Among the most prestigious projects related to people detection and tracking using stereo vision we find the one by Darrel et al [1]. 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. 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 is used. In this way, the people can be identified even though they disappear from the image for a while.

In [4] it is shown a system for people detection and tracking for the interaction in virtual environments. The system allows the user to navigate in a virtual environment by just walking through the room using virtual reality glasses. In this work the face detection is a crucial aspect that has been resolved by using the face detector proposed by Viola and Jones [12]. Once the person is located, a histogram of the colors of the face and chest is used and a particle filter estimates the position of the person based on the information. The stereo information assists on knowing the position of the person in the room and therefore identifies their position on a virtual environment. On this work, the stereo process is performed by using the information gathered by different cameras located at different points of the room.

In [6] a method to locate and track people in stereo images is presented by 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, a map of occupancy is created, as well as a height map. 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 [5–7]. 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 [1, 4, 11]. 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 human-machine interaction. 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 [2]. Otherwise the person could feel intimidated.

In this work we propose a method for people detection and tracking by using a movable stereoscopic system located at inferior levels from the people's height. Unlike most of the documents reviewed, that only model the environment for unmovable cameras [1, 4–7], we propose a method to create a map that models all visible environment by the stereoscopic system when rotating the robotic head. A distinguished characteristic of this method is that even with movable objects present, the map can still be created. The use of this map will allows 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 agent that has been created uses active vision to track the person movements through all the room. This situation allows us to track the person on a wider environment, than if we had used immovable cameras, and thus makes feasible a more natural and comfortable interaction.

2 Hardware system

The hardware system is formed by a laptop to process the information, a stereoscopic system with 2 cameras and a robotic head. The robotic head (Pan-Tilt Unit or PTU) has two degrees of freedom, one on the X axis (pan) of $\phi = [-139, 139]$ degrees and the other one on the Y axis (tilt) of $\psi = [-47, 31]$ degrees.

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. Due to the fact that the camera is subject to movements, these points are translated to a reference static

system that has as center the robotic head located 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.

$$p_w = Tp_{cam} \tag{1}$$

3 People Detection and Tracking Process

The method for people detection and tracking proposed on this document, is an iterative process that has been outlined in Figure 1.

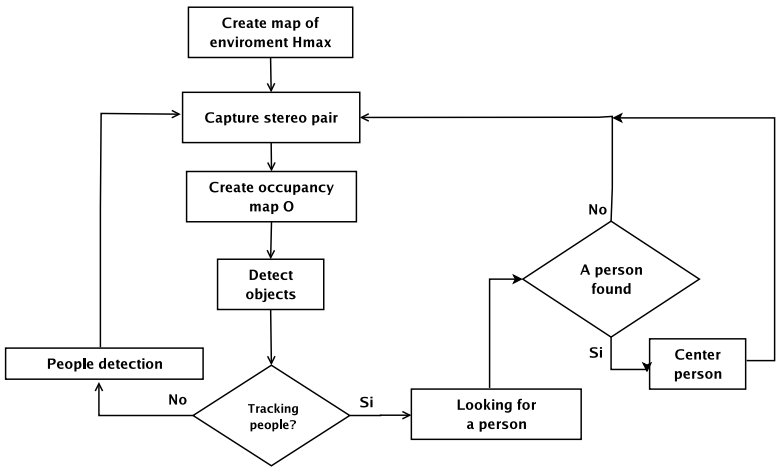


Fig. 1. Machine of states of the process.

Initially, a map of the environment is created (let us denote it by $\mathcal{H}_{max}$) 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.

Once the environment has been registered, the system begins a continuous process to capture images in order to create 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 people detector [12] over the color image of the scene. The false positives generated by the face detector will be rejected thanks to the integration of the information of

Fig. 2. Creation of the height map. Upper row (a,b,c,d) shows the images in instants $1, 4, 10$ and 13. Central row (e,f,g,h) shows the instantaneous height maps $\mathcal{H}^t_{max}$ for each one of the upper images. Lower row (i,j,k,l) shows the evolution of the height map $\mathcal{H}_{max}$ created as the median of the height maps $\mathcal{H}^t_{max}$ created until that moment

the disparity image and $\mathcal{O}$. If finally, some of the objects detected in $\mathcal{O}$ turn out to be people, the agent will begin to track the closest one.

The tracking is also an iterative process that creates on every moment a occupancy map $\mathcal{O}$ to track on it the target person. To perform the tracking the Kalman filter has been used. If the target person is located, it is determined if the PTU needs to be moved in order to center the image and in this manner have them always on sight. The objective of the centering process is to keep on the image, as long as feasible, the head and arms of the person. In the following sections the more relevant processes previously mentioned will be elaborated in more detail.

3.1 Creation of the map of the environment $\mathcal{H}_{max}$

Firstly, to the detection process, the environment is registered. This process aims to register the structure and motionless objects in it. This environment model will assist when separating the objects that are not part of it (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. 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}.

Because of stereoscopic system is subject to error, instead of only projecting the height of the point detected on a cell, it is projected the whole uncertainty area of that point. For that purpose we have used the error model of the stereoscopic system with the parameters provided by the manufacturer.

The creation of $\mathcal{H}_{max}$ by only using a unique disparity image is subject to problems. On one hand, possible objects that do not belong to the environment (for example people passing by) could be incorrectly included as part of the environment. Also, the correlation algorithms for the stereo detection are subject to error that cause that not all the scene points are detected. Due to these reasons, instead of creating $\mathcal{H}_{max}$ from a unique disparity image, it will be done by taking several images on different time sets t. For each one of this images, an instantaneous height map is created $\mathcal{H}^t_{max}$. Finally, the different $\mathcal{H}^t_{max}$ created are used to calculate $\mathcal{H}_{max}$ through a robust estimator such as the median. For that purpose, each cell of $\mathcal{H}_{max}$ will have as a maximum height value the median of all values observed on the different $\mathcal{H}^t_{max}$ for that particular cell using Equation 2.

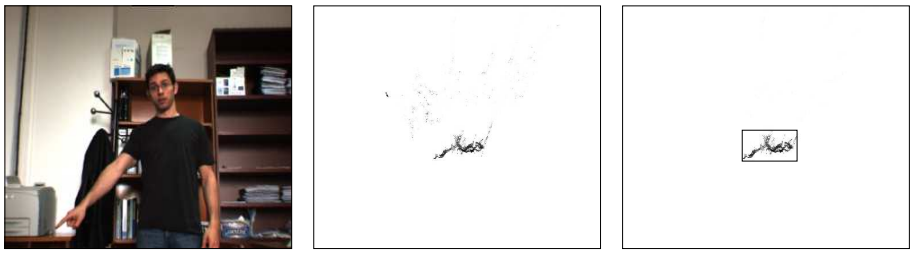
$$\mathcal{H}_{max}(x,y) = median(\mathcal{H}^{t=1}_{max}(x,y), ..., \mathcal{H}^{t=n}_{max}(x,y)) \tag{2}$$

On Figure 2 we can observe the creation of the height map of an environment. The map has been created by using a sequence of 13 images. Figure 2 only shows the images of the instants $t = \{1, 4, 10, 13\}$. On the upper row (Figures 2 (a-d)) are shown, from left to right, the images of the moments previously mentioned. On the middle row (Figures 2(e-h)), we can see the instantaneous height maps ${\mathcal{H}^{\sqcup}}_{max}$ of the upper row images. The dark areas represent the highest zones and the white areas represent the lowest ones h_{min}. On the lower row (Figures 2(i-l)), it is shown the evolution of the height map $\mathcal{H}_{max}$ until the instant t. We can observe that for $t = 1$, $\mathcal{H}^{t=1}_{max} = \mathcal{H}_{max}$. But as we continue using more images, the height map tends to truly represent the environment. To create these maps 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.

In order to create a complete map of the environment the camera will need to turn so it can capture information from different directions. For that purpose, the process previously described will be repeated for different values of the ϕ angle until it covers all the visible environment by the visual system. Due to space reasons, image of a complete map is not included.

3.2 Creation of the occupancy map $\mathcal{O}$

Once the height map $\mathcal{H}_{max}$ has been created, the people detection can begin. The first step, is to create a occupancy map $\mathcal{O}$, which will indicate on each cell the



(a) (b) (c)

Fig. 3. (a) Image of the environment (b) Occupancy map $\mathcal{O}$ (c) Detected objects in $\mathcal{O}$

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 closer to the camera correspond to small surfaces and vice versa. Therefore, the farthest points will increment the value of the equivalent cell by a higher quantity than closer ones, using the same equation that in [6]. If we used the same increment for all the cells, the same object had a lower sum of the areas the farther it is located from the camera. This scale on the increment value will compensate the difference on size of the objects observed according to their distance from the camera.

Once $\mathcal{O}$ is created, we will analyze it 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.

Figure 3(b) we can observe the occupancy map $\mathcal{O}$ of the environment on 3(a) using a height map $\mathcal{H}_{max}$ from 2(l). The darker values represent the areas with higher occupancy density. The image has been manually retouched to make visible the occupied areas. As it can be observed, on the upper area of Figure 3(b) there are small dark dots that represent the noise of the stereo process. On 3(c) we can see in a frame the only object detected after the closing process, grouping and thresholding.

3.3 Face Detection

If after the creation and analysis of $\mathcal{O}$, an object that has entered the environment has been detected, we proceed to use a face detector to determine which one could be a person's face. As face detector method we have used the one initially

Fig. 4. Face detection. (a) Registered image and face detected (b) Disparity image and face detected (c) Face region belonging to the person.

proposed by Viola and Jones [12]. This method consists on a general object detector based on the utilization of multiple simple classificators arranged in cascade [8]. The method takes as input an image in gray levels and selects the areas of the image where a face is detected. (Figure 4(a)).

Due that the face detector tends to identify false positives, it is important to verify that detected object is indeed a person's face. As a verification mechanism the points detected that could be part of a face should not be spread out among different objects in $\mathcal{O}$. For that reason, we will analyze the area of the face on the disparity image to verify if the points are part of one only object in $\mathcal{O}$. However, this area could have points that represent the background or other object even when the detector indeed identifies a face. On Figure 4(a) we can see how a face is identified and how in the same area, there are points that belong to face and to the wall on the background. For this reason, it is very important to precisely identify which points, on the area identified by the detector, are truly part of a face and which are not. For this mean we run a process that consists on calculating the median of the disparity values from the frame indicated by the face detector. The points with such disparity value are used as seed to perform a region growing. On 4(c) we can observe the selected region by this method for the disparity image 4(b). If after this analysis the system would identify more than one person on this environment, the system would start tracking the closest one.

3.4 Tracking

The tracking process is interactive and begins when we take a stereo pair and create its map of occupancy $\mathcal{O}$. After identifying the people present in $\mathcal{O}$ (as it has been previously explained), we need to determine which one will be tracked. To merge the available information taken in previous moments as well as the information processed on the current moment the Kalman filter has been used. This tool will allow us to merge in a proper manner the position that predicts the model with the information gathered during our search process. If the person is detected, active vision is properly used directing the visual system so the target is always centered on the image. The centering process aims to keep the subject

visible on the image placing it on the best possible image position to analyze their gestures. If the subject is standing up on normal position, the goal is to capture the head and torso. If the subject raises their arms to point to any object or if he bends to pick up something from the floor, it is desirable to be able to register the action. On this work, we have contemplated the possibility of the movement sideways as well as the movements that imply changes in height (bending or sitting down). We have experimentally proven that the best option to achieve this is to keep the highest visible zone of the subject in the upper area of the image. To determine the movement that the PTU needs to perform in order to center the subject, we have used a system based on fuzzy rules that have been designed with expert knowledge and that have been adjusted according the our experimentation.

4 Experimentation

During the explanation of the model we have shown examples of its performance. A broader experimentation has been done that we are unable to show with images due to space reasons, but we will briefly explain. This experimentation refers to the detection and tracking of different people under different illumination conditions and different distances from the vision system. To perform the stereo process we have used images size $320x240$ using sub-pixel interpolation to enhance the precision.

The use of the proposed height map enables us to model the whole environment rotating the stereoscopic system on all directions. The creation method proposed (to use the median of the heights) allows to create the map even when there are people moving around in the room just as shown on Section 3.1.

We have proven the accurate performance of the people detection method by satisfactorily eliminating the false positives produced by the face detector. The more adequate distances to detect people vary within $0,5$ and $2,5$ meters. However, once the person to be tracked is selected among the others, the tracking can take place in longer distances.

The time to compute is different on the detection process than on the tracking one, although the stereo process consumes most of the time (120 ms). On the tracking process, the face detector is the toughest task (81 ms), reaching ratios of $2,5$ fps for the whole detection process (including the stereo processing). However, on the tracking process we reached ratios up to 5 fps and we have proven that it is enough in our case. These values could substantially increase if it were feasible to optimize the code of the stereo process or with the use of specific hardware, for depth computation.

5 Conclusions and future work

On this paper we have presented an agent capable of people detection and that uses active vision to track them. For that reason we have used a stereoscopic system installed on a robotic head. The agent initially creates a height map

of the environment that registers the motionless characteristics of it. 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 agent is capable of tracking them by using a robotic head that enables the stereo system to rotate. In order to achieve robust tracking process we have used the Kalman filter. The agent keeps the person located at all times by framing their arms and head in the image. This task could be used by other agents that might need to analyze gestures and expressions of potential users in human-machine applications.

As future projects we visualize the update of the map of the environment. For that reason, the agent should be capable of adding to the environment map those objects, that are not people and stay motionless for a long period of time and were not there when the map was initially created. Other aspect to consider is the use of the particles filter for the tracking task.

References

1. T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated Person Tracking Using Stereo, Color, and Pattern Detection. *Int. Journ. Computer Vision*, 37:175–185, 2000.
2. T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42, 2003.
3. D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.
4. D. Grest and R. Koch. Realtime multi-camera person tracking for immersive environments. In *IEEE 6th Workshop on Multimedia Signal Processing*, pages 387–390, 2004.
5. I. Haritaoglu, D. Beymer, and M. Flickner. Ghost 3d: detecting body posture and parts using stereo. In *Workshop on Motion and Video Computing*, pages 175 – 180, 2002.
6. Michael Harville. Stereo person tracking with adaptive plan-view templates of height and occupancy statistics. *Image and Vision Computing*, 2:127–142, 2004.
7. 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*, pages 681–686, 2004.
8. Intel. OpenCV: Open source Computer Vision library. *http://www.intel.com/research/mrl/opencv/*.
9. W. Liang, H. Weiming, and L. Tieniu. Recent developments in human motion analysis. *Pattern Recognition*, 36:585–601, 2003.
10. L. Snidaro, C. Micheloni, and C. Chiavedale. Video security for ambient intelligence. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 35:133 – 144, 2005.
11. R. Tanawongsuwan. Robust Tracking of People by a Mobile Robotic Agent. Technical Report GIT-GVU-99-19, Georgia Tech University, 1999.
12. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 511–518, 2001.