Stereoscopic System for Human Body Tracking in Natural Scenes

*IRI. Robotics Institute (UPC/CSIC)
**Dep. of Automatic Control
Universitat Politècnica de Catalunya
frigola@esaii.upc.es, www-esaii.upc.es

ABSTRACT

Human body detection and tracking in a scene constitutes a very active working field due to their applicability to many areas, specially as a man-machine interface (MMI) means. The system presented aims to improve the reliability and efficiency of teleoperation. The system is of application to teleoperated manipulation in civil applications such as big robots in shipyards, mines, public works or even cranes. Image segmentation is performed from movement detection. The recognition of moving bodies is verified by means of a simplified articulated cylindrical model, thus allowing to operate with a low computational cost.

1. INTRODUCTION

Passive sensing of human motion is important for a wide range of applications such as human-computer interaction in virtual environments, surveillance, anthropology, robots task programming, telerobotics and novel forms of artistic expression. In many of these applications a non-intrusive sensory method based on vision is preferable over other methods relying on markers attached to the body. This is due to the fact, that the interpretation of gesture orders opens wide possibilities to the control of many processes without the need to use computer terminals, and then, to have available a hands-free operation interface. Another important aspect to consider is that the video-based approach offers the potential capability to describe the human movement related to other objects in the scene. Thus, such systems can be useful in the control of automatic systems operating in environments such as mines; civil works; and robotics in agriculture or also in industry. That is in jobs where there is a requirement for human-robot co-operation. In general, in any application where a human operator has to interact, hands-free, with an information processing system.

One of the most difficult problems for a dynamic vision system is the tracking of non-rigid objects, such as a human body in a cluttered environment. Recently, there has been a great number of papers concerning human motion analysis [2,8,9]. They use models with a great number of degrees of freedom, which are based on articulated tracking techniques, constrained by the kinematic chains. All these methods require a reasonable accurate initialisation, are restricted to small or slow movements and can not deal with long-lasting occlusions. Alternatively, the authors in [11] have been working in robust colour-region tracking and statistically-derived rules to determine body features in the contours of the silhouette. This system is restricted to environments with relatively controlled lighting conditions due to its slow dynamics for recovering the changes of the background.

The present paper faces up the problem of human motion analysis with a vision system that detects and tracks the human moving parts. This system provides some flexibility to overcome the above mentioned problems. Due to its improved segmentation process, the method is suitable for practical applications in outdoor environments without requiring massive computational resources and still run quickly enough to control, in real time, teleoperated systems. Based on image segmentation from motion detection, the system uses a simplified model of the 3D human shape, that copes with the multiple viewpoint images integration problem. We describe a model-based tracking method, in which three-dimensional geometric models are used for pose estimation even in cluttered scenes, poor lighting conditions and large displacements of the moving target.

2. SEGMENTATION FROM MOTION

Current systems usually operate in specific plateaus with high contrast between the target and the background. Since the detection of a human shape in a natural environment is very complex, the detection operation is frequently carried out from movement detection. With the developed system, very satisfactory results have been
obtained from movement detection, not working from successive images subtraction, that produce wide lateral bands in case of quick body movements, but from consecutive gradient images. The segmentation process is improved by using the estimation of the gradient orientation variations rather than performing only a gradient module analysis. This procedure allows to eliminate almost completely the background. Obtaining the gradient orientation requires higher computing power, but, in this system, this requirement does not implies longer computing time since a specific hardware has been designed for image segmentation. The attainable operating speed allows to operate in outdoor environments, even with large and quick displacements of the moving target.

In this way, what is pretended is to obtain a new gradient image of the moving objects, and from this image, the objects’ contours. These contours will constitute the information corresponding to the scene moving objects. The gradient orientation is more discriminating than the gradient module because the variation of the gradient magnitude can be produced not only by a change in the position of an object on the scene but also to changes in lighting conditions. On the contrary, variations in lighting conditions do not produce, in general, variations in the gradient direction. Fig. 1b shows the contour extracted comparing the value of the gradient orientations of two consecutive images, obtained in a time interval of 200 ms.

The procedure is based on comparing the gradient vectors obtained for every pixel (within a 5 x 5 window) from consecutive images. This comparison could be made using as distance function the module of the difference vector,

\[ |\vec{G}_t(x,y) - \vec{G}_{t+1}(x,y)| > K, \]

which is high sensitive to either motion and illumination changes, or by means of gradient orientation comparison which is less sensitive to lighting conditions:

\[ |\text{atan2}(\vec{G}_t(x,y)) - \text{atan2}(\vec{G}_{t+1}(x,y))| > \theta. \]

The gradient is evaluated only when \(|\vec{G}_t(x,y)| \) or \(|\vec{G}_{t+1}(x,y)| \) are significant enough. Otherwise due to noise, the angle calculated by the arctangent function has no sense. Fig. 2 shows a comparison of the classical method of substraction and thresholding images, and the proposed method.

Fig. 1 Movement detection. a) Original image, b) Moving parts extraction from the background using the gradient orientation method

Fig. 2 (a) Original Image, b) Changes due to motion, c) Grey level subraction, d) Argument comparison.
To speed up the processing time a specific processing board operating over the PC bus has been designed (Fig. 3). In a first step the gradient vectors are obtained for every pixel in the image using a 5 x 5 pixels window and the argument is calculated by a look up table. In a second step an arithmetic operator calculates the difference between the arguments of the present image and a delayed one.

![Diagram](attachment:image.png)

**Fig. 3 The specific hardware structure**

To improve the system performance, the time interval $\Delta t_i$ between the images taken to compare the gradient arguments at each instant, adapts itself to the scene dynamics. For this reason, each time interval, $\Delta t_i$, increases or decreases as a function of the quantity of movement estimated by the apparent motion in the image.

The lateral bands of the image difference is a rough estimation of the quantity of movement, but the area of the lateral bands depends on the camera-target distance. In order to reduce this dependence, instead of measuring this area, we measure the local densities of the binarised difference of the actual grey-level image and a previous reference image. A histogram is calculated counting not null pixels, using an L x L pixels window, for every not null pixel of the image difference. A temporal evolution of the normalised histogram $N_t(l)$ of a target with constant velocity is show in Fig. 4.

![Graph](attachment:graph.png)

**Fig. 4 Temporal evolution of local densities and threshold plane.**

The last portion of $N_t(l)$ is very useful to detect significant motion of the target. The hardware module sends an interruption signal to the host processor when some given thresholds of movement sensitivity are surpassed. Fig. 4 shows one such threshold plane.

The use of this hardware makes it possible to have available the map of the scene mobile points, at video rate. These points are obtained at variable time intervals according to the scene dynamics, without requiring additional processing time.

3. MODEL FITTING

A simplified 3D geometrical model of the human body is used. The imposition of some anthropomorphic constraints and the availability of some dimensional measures, make it possible to operate with a geometrical model qualitatively more elemental. Thus, it is possible to reject the shapes that do not fit to an adequate profile. The model of a person is constituted by a set of cylinders that fit to the moving parts profile. The model consists of two coaxial cylinders that are adjusted to the head and body, and also a set of up to four cylindrical surfaces that are adjusted to the body overhanging elements, that can correspond to the arms (Fig. 5).

![Model](attachment:model.png)

**Fig. 5 Polycylindrical simplified geometrical model**

In order to adjust a detected moving object to the defined geometrical model, the system first obtains some singular points from the moving object shape. The singular points considered are the absolute and relative maximum ordinates of these shapes and also their relative minimum. The absolute minimum has not been considered since, as a consequence of ground reflections, it is usually meaningless. These considerations are based on the assumption that the operator is standing or seating, and consequently the direction of exploration is from the head to the feet. The singular points are classified as belonging
to the head by analysing their relative positions with respect to the obtained contour. The algorithm developed to classify, in a simple way, an absolute or relative maximum as a head consists on only counting, in descendent sense and from each singular point detected, the number of body lines that are located over the vertical line passing through this point. These lines have been obtained by limiting them horizontally between the two extreme points of the contour shape obtained (Fig. 6).

**Fig.6 Method for singular points classification**

The point with the highest absolute or relative value, that has been classified as belonging to the head using this criterion, has shown to be coincident enough in the two stereo images. Since it is the most prominent point in the image its detection in the two stereo images presents no ambiguity when there are no occlusions. The head point is the one used to centre the coaxial axis of the two main cylinders of the model. This first model adjustment carried out by stereovision calculations is considered to accomplish the "person" conditions if the corresponding cylinder verifies:

$$H_1 < H < H_2 \quad \text{and} \quad R_1 < R_c < R_2$$

being $H_1$ and $H_2$ the minimum and the maximum height over the reference plane, that enables to detect either sitting or standing persons, as convenient. The cylinder's radii must also be contained between some limits that in the system tested are 10 and 20 cm.

The cylinder corresponding to the body, obtained as the circumscribed surface to the extracted contour, starts 20 cm below the C point. If this contour presents significant enough maxima or minima abscissa points, the skeleton of these protuberances are extracted and their shapes fitted to a maximum of two cylinders each. The measure of these arms' models, considered from their ends, points $B_1$ and/or $B_2$, to the vertical axis passing through the C point, can also be used to validate the person's model that has been shaped.

4. **MODEL VALIDATION**

Each time a moving object is detected the system obtains the three-dimensional co-ordinate of the prominent points. Once the point considered as the upper limit of the head axis C has been identified, as well as the other possible points $B_1$ and $B_2$, the system verifies the coherence of the radii $R_1$ and $R_2$ corresponding to these points $B_1$ and $B_2$ and also their heights: $H_1$ for the head, and $H_1$ and $H_2$ for points $B_1$ and $B_2$. Fig. 7 shows the global structure of the tracking system.

The fact that these points are obtained at variable time intervals, according to the target dynamics, with time values up to 160 ms, allows to perform also an efficient temporal filtering, to obtain the trajectory of the person and those of his arms, which are tracked with a higher resolution.

**Fig. 7 Tracking system structure**

5. **CAMERAS CONFIGURATION AND 3D MEASUREMENTS**

The developed system operates by stereovision, selecting two out of $n$ cameras, those that permit to avoid occlusions.
To obtain a resolution high enough so as the use of a volumetric model is meaningful, the stereoscopic vision system uses a long baseline. The resulting disparity is admissible thanks to the system robustness. The cameras are placed strategically on the working zone. With this long baseline (D > 5 meters) configuration the maximum error obtained, working in environments up to 8 x 6 m, is less than 2 cm.

![Fig. 8 Cameras configuration and base line.](image)

Camera calibration is a basic aspect of any stereoscopic system. Usually, a camera calibration technique consists on identifying a set of extrinsic parameters (the position and orientation of the camera in a given reference frame) and a set of intrinsic parameters (coefficient of distortion, scale factors and pixel offsets) [3,10]. In the present system, the intrinsic parameters are calibrated in the laboratory using a reticular pattern. Unfortunately, the task of calculating in situ the 6 extrinsic parameters for each camera need an excessive number of calibration points. In order to reduce the calibration data points needed by the system, some assumptions are made. These assumptions are based on the fact that the incidence of the camera position errors in stereoscopic systems with a long baseline is much lower than those produced by the orientation errors.

The relative cameras position introduced by the user is considered to be correctly measured (with error less than 1% of D) and only two points far enough in the image are needed to calibrate the orientation of the camera. In these conditions, the net effect of an error in the camera position produces a quasi-constant error in the determination of the object position, which can be compensated as an offset.

If a higher number of calibration points is used, it is possible to improve the robustness in case of occlusions of any of these points (Amat et al., 1998).

6. MULTIPLE BODIES TRACKING

A multiple persons location and tracking module can be very useful by selecting the optimal views of people and providing a foveal area for gesture recognition. Our approach is based on multiple 3D geometrical hypothesis and test. A human is modelled as a 3D cube (with variable and restricted sizes) and the question to solve is which number, configuration and location of these cubes better explain the background-segmented images. Using this simplified model of a person we can not deal with very complex occlusions. Our approach is based on the selection of the two cameras that offers greatest success possibilities in a posterior analysis (Fig. 9).

From the singular points (see section 3), we measure the vertical (vs) and horizontal (hs) distribution of the blob (vs and hs represent a bounding rectangle) and look for a coherent matching in the other images. A definition for a coherent matching between two bounding rectangles is resumed in the next conditions:

- The minimum distance of the rays defined by the respective singular points is less than a proportion of the maximum camera-object distance.
- The position defined by these rays must lie into the working volume.
- The horizontal (HS) and vertical (VS) sizes of the inverse-projection of the respective projected measures in the image plane (hs and vs), must have acceptable width and high.

Given a bounding rectangle in an image, we must decide which is the best matching rectangle in the images obtained from the other cameras. A measure of the adjust of two bounding rectangles, \(i\) and \(j\), could be computed as:

\[
\text{abs} (HS_i - HS_j) + \text{abs} (VS_i - VS_j)
\]

A sketch from a vertical view (Fig. 9) shows an example of the horizontal sizes (HS\(_i\), HS\(_j\)) and four possible matching combinations. The difference in heights (not showed in figure) of the wrong matches will be enough to discriminate them.

![Fig. 9 Matching combinations](image)
7. GESTURE RECOGNITION

The ability to perceive particular body motions as meaningful gestures is therefore essential if computer systems are to interact with human users. Some authors have adopted appearance-based techniques for gesture recognition, where the basis of the representation of human movements is a sequence of pattern images. This representation allows to model complex poses for which no simple 3-D models or recovery methods are available [5,6,7]. But, in any case, there is no explicit information about the geometrical aspects of gestures. Several applications need to measure the head and hand position (and orientation) in order to know where (or what) a user is pointing. The recognised gesture must be a complementary information with geometrical data to subsequent high level interpretation of human actions.

This MMI based on the tracking of a person's gestures has given satisfactory results in applications where the coded orders to transmit from the human operator to the machine are relatively simple. The interpretation of these simple orders is based on the obtention of a new simplified model of the tracked body, extracted from the skeleton of the body polycylindrical model. This dynamic model constituted by the axis of the body-arms cylinders allows decoding orders from the analysis of image sequences using the diagram shown in figure (Fig. 10).

The orders initially interpreted are: "Stop" when the two arms are raised up over the head; "Slow", "Regular" or "Quick" when the two arms move periodically; "Raise-up" or "Go-down" when only one arm moves periodically. And finally "Up" or "Down" when one arm is maintained raised-up or down. The automatic visual interpretation of these orders is carried out quicker and more reliably than using as an interpreter a human operator. This fact is especially true in tasks where the operator that directs a given work is not the same as the one having the control of the machine.

8. CONCLUSIONS AND RESULTS

An application of this interface from the detection of a person and the tracking of his gestures has been developed. The movement detection from the variation of the gradient orientation over the images enables to carry out an efficient extraction of persons from the static background. The use of a geometric simplified model has enabled to keep tracking moving persons with high reliability, being the model updated even when the tracked person moves at relatively high speed. Fig. 11 shows the updating of the model in different obtained images.

Currently, the system is being experimented within the research project Intelligent Room by the Human-Computer Interaction Group of the MIT Artificial Intelligence Lab. The goal of the Intelligent Room is to allow people to interact with computational systems the way they would with other people: via gestures, voice, movement, and context [4]. The Pointing Device is the module able to track human body movements and provides the 3D-position of the head, hand and elbow of a person at a rate of up to 10 times per second. The module is registered in the local area network, and other computers, using the same local area network facilities (TCP/IP and Sockets), can easily read the data provided by the device. Fig. 12 is a taken snap of a PC monitor showing a real-time graphical representation of the data provided by the device.
8. REFERENCES


ACKNOWLEDGEMENTS

The authors would like to express their acknowledgements to the Computer Interaction Group of the MIT Artificial Intelligence Lab, especially to professor Tomas Lozano-Perez.