A Specialized Vision System for Control by means of Gestures

Josep Amat, Alicia Casals and Manel Frigola
Dep. of Automatic Control
Universitat Politècnica de Catalunya
Pau Gargallo n. 5, 08028 Barcelona SPAIN
phone: 34-3-4016971, fax: 34-3-4017045, email: casals@esaii.upc.es

Abstract

Persons detection and tracking in a scene constitutes a very active working field due to their applicability to many areas and specially in MMI applications. The system presented in this paper is the result of the work carried out in the last years to improve the reliability and efficiency of operation in natural complex environments. Image segmentation is performed from movement detection. The recognition of moving objects such as persons, is performed by means of a very simplified multicylindrical geometrical 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, anthropometry, 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 bodies of human subjects. This is due to the fact, that interpretation of gesture orders open wide possibilities to the control of many processes without the need to use computer terminals, and then to have available a free hands operation interface. Thus, such systems can be useful in the control of automatic systems operating in environments such as mines; civil works; robotics in agriculture or also in industry, in jobs where there is a requirement for human-robot cooperation. 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 in a cluttered environment. Recently, there have been a great number of papers concerning human motion analysis, but no hardware-oriented techniques exists to automatically acquire a concise 2D model of the human body. This paper contributes to the problem of human motion analysis with a vision system that detects and tracks the human moving parts, the gestures. Based on image segmentation from motion detection, the system uses a simplified model of the 3D human shape, which copes with the multiple viewpoint 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.

The system described in this paper is related to body-tracking projects such as those from Regh and Kanade [1] that use kinematics models and Metaxas and Terzopoulos [2] that use dynamic models. Recently, Azarbayejani and Pentland [3] have been working in spatial-color-texture statistical analysis of regions. These systems limitation is that they require accurate initialization of the feature model [1], [2], or initial model of the scene (with dynamic recovery of the background), [3] requiring all of them massive computational resources.

Functionally, our system are perhaps more closely related to the works of Huttenlocher [4], Nugroho et al [5], and Batista et al. [6] specially in the low-level target motion detection. This system constitutes a further step in this direction, including a geometrical human simplified model designed to supervise the tracking process with a relatively low computational cost. The presence of motion in the scene is, in Nugroho, perceived by time-varying edge detection and temporal images sequence comparison. The segmentation operation in our system is improved by using the estimation of the gradient orientation variation, rather than performing only a gradient module analysis. This method is suitable for real-time applications in a stationary camera environment.
2. Detection based on a Simplified Model

Since the detection of a human shape in a natural environment usually presents difficulties, frequently, the detection operation is carried out from the movement detection that allows to eliminate almost completely the background.

With the developed system, we have obtained very satisfactory results in the movement detection, not working from successive images subtraction, but from consecutive gradient images. In this way, we pretend to obtain a new gradient image of the moving objects, and from it, the objects’ contours. These contours will constitute the information corresponding to the possible scene moving objects. The resulting image is better if the contours are obtained from the variation from image to image of the gradient orientation, rather than if they would had been obtained from the module of the gradient calculated from the gray level images. The gradient orientation is more discriminating than the gradient module due to the fact that 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 gradient image corresponding to the scene of fig. 1a). Fig. 1c) shows the contour extracted comparing the value of the gradient orientations of two consecutive images, obtained in a time interval of 200 ms.

To improve the system performance, the time interval $\Delta t$ between the images, taken to compare the gradients arguments at every instant, adapts itself to the scene dynamics. For that reason, at each time interval, $\Delta t$, increases or decreases as a function of the quantity of movement, $Q$, obtained in the $\Delta t$, interval. The $Q$ variable used, quantity of movement, corresponds to the maximum local density detected in the image, using a $h\times h$ pixels discretizing window.

With the aim to discriminate with more reliability the scene moving elements, even in scenes with a very complex background that do not allow to obtain good enough results, from the persons to detect and track, a simplified geometrical model of the human body is used. Thus, it is possible to reject the shapes that do not fit to an adequate profile, so as to be considered a

Fig. 1 Moving objects detection. a) Original image, b) Gradients image, c) Extraction from the background.
The developed system operates by stereovision, using two cameras that permit to obtain dimensional measures. The imposition of some anthropomorphic constraints and the availability of dimensional measures, make it possible to operate with a geometrical model qualitatively more elemental.

To obtain a resolution high enough so as the use of a volumetric model is meaningful, the stereoscopic vision system utilized has a long baseline. The cameras are placed one at each side of the working zone. With this configuration the maximum error obtained, within the working experimental environment, a room of 8 x 6 m with two cameras having a 50° aperture angle, is 1,8 cm. The problem of using a long baseline consists, in general, on the fact that the determination of the homologue points is very difficult from intensity images. But, this is not the situation in our system, since the singular points considered are the maxima absolute and relative heights, and these points do not use to be ambiguous.

2.1 The model

The model of a person is constituted by a set of cylinders that fit to the moving parts profile obtained from each camera. 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 correspond to the arms (Fig. 2).

In order to adjust a detected moving object to the defined geometrical model, the system first obtains some singular points from the obtained shape. These singular points are the absolute and relative maxima ordinates of these shapes and also their relative minima. The absolute minimum has not been considered since, as a consequence of ground reflections, they are usually meaningless.

The singular points are classified as head or arms by analyzing their relative positions with respect to the obtained contour. The algorithm developed to classify, at high speed, 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 in 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. 3.

The point with the highest absolute or relative measure that has been classified as a head using this criterion, has shown to be coincident enough in the two stereo images. This point is the one used to center the coaxial axis of the two main cylinders of the model, Fig. 5. This first model adjustment carried out by the stereovision calculations is considered to accomplish the "person" conditions if the corresponding cylinder verifies:

\[ H_s < H_r < H_b \]
\[ R_s < R_r < R_a \]

being \( H_s \) and \( H_r \) 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 had been 10 and 20cm.

Fig. 4 Fitting of a multicylindrical model from the detection of the prominent 3D point of the head (C)

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 abcissa points, the skeleton of these protuberances are extracted and their shapes fitted to a maximum of two cylinders. The measure of these arms' models, considered from their ends, points B₁ and/or B₂, to the vertical axis passing through the C point, can also be used to validate the person model that has been shaped.

2.2 Target Tracking Validation

Each time a moving object is detected the system obtains the tridimensional coordinate 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 B₁ and B₂, the system verifies the coherence of the radii R₁ and R₂ corresponding to these points B₁ and B₂ and also their heights, H₁ of the head, and H₁ and H₂ of these points B₁ and B₂, Fig. 5.

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

2.3 Calibration

Camera calibration is a basic aspect of any stereoscopic system [7][8].

We use four reference points that can be specific elements introduced to the scene with this end, or can be natural elements of the own scene.

In the developed system we use four singular points of a scene element which coordinates have been measured with enough precision. In this case we have used four vertices inside a window, fig. 6. To attain a higher calibration resolution, the tracking of these reference points is performed by correlating a region of the gray level image containing these points with a model. Each model of 16 x 16 pixels contains a singular point obtained in a previous manual calibration points definition step.

The models are continuously updated performing a classical recursive filtering algorithm. The correlation is carried out splitting up this calibration process along the different cycle time, so as not to damage the tracking processing time. Since the incidence of the stationary camera displacement Δₓ, Δᵧ, and Δz errors in stereoscopic
steps pipeline structure also permits to carry out a residual noise filtering, thus making the tracking of a target more robust.

It is expected that with the work being carried out for the improvement of the segmentation and data filtering a higher precision can be obtained.

The system has been successfully experimented in a TV plateau to facilitate the process of weather information. The weather-man indicates the state of the weather and its evolution by means of his gestures. He can either put the symbols corresponding to storms, to sun or to clouds, Fig. 8 a) or can move the anticyclones indications to show their evolution over the isobaric map, Fig. 8 b).

3. Results

The results that have been obtained with the person detection and tracking system are considered as highly satisfactory. The movement detection from the variation of the gradient orientation over the images allow to carry out an efficient static background extraction. The use of this geometric simplified model has enabled to keep tracking moving persons with high reliability, being the model able to be updated even when the tracked person moves at relatively high speed. In Fig. 7 we show the updating of the model in different obtained images.

The use of a dedicated hardware to obtain the gradient module and the orientation of the gray level image, that operates as an image acquisition and preprocessing module of a computer, PC type, allows to operate at a rate of up to 6Hz. The dedicated hardware with a two
Fig. 7 Result of applying the multicylindrical model to different images

4. References


