Proceedings of the 2003 IEEE
International Conference on Robotics & Automation
Taipei, Taiwan, September 14-19, 2003

Visual Human Machine Interface by Gestures

Manel Frigola, Josep Fernández, Joan Aranda
Automatic Control & Computer Engineering Dpt
Universitat Politècnica de Catalunya
Barcelona - SPAIN
{frigola,ruzafa,aranda}@esaii.upc.es

Abstract

Like oral communication, gestures are a natural way to carry out Human Machine Interface. In the early days of robotic systems, human gesture was used to control robot movements by means of a master-slave structure. In spite of the use of robot programming languages, manual control is the most reliable way to carry out complex tasks in unstructured environments. In these situations, a non-contact, passive and remote system can be helpful to control a teleoperated robot by means of human gestures.

In this paper, a vision system able to detect, locate and track the head and hands of a human body is presented. The system uses several calibrated cameras placed around the operator scenario to locate the body parts of a person in 3D. The system combines different computer vision techniques to increase the reliability of the body parts detection: image movement detection, user skin colour segmentation and stereo. The data provided by these modules are fused looking for coherence according to the human body dimensions. With the scheme proposed it is possible to obtain a low-cost real-time system for human computer interfacing based in a natural way of communication (gestures). Civil area such as big robots in shipyards, mines, public works or cranes are some possible applications.

1. Introduction

Human-Machine Interface (HMI) is a key factor in the development of efficient computer assisted control systems. Standard and specific interfaces for control applications have been developed for years, and more recently the speech recognition and voice synthesizing systems enable the development of oral control systems.

Oral computer communication systems can be complemented by other natural communication means, based on gesture interpretation. This interface is oriented to teleoperated tasks that needs a hands-free operation.

First works on body motion analysis use a scene with an homogeneous background. Currently, several segmentation techniques may be useful to discriminate persons in complex scenes. This techniques are movement image detection [1], colour and texture [2], depth from stereo [5], skin colour image extraction [4], thermal images [8] and face detection based on pattern recognition [6]. Also, a increasing number of papers combine different techniques, as in [7] that propose the usage of motion, colour and face detection but no stereo information. In an another work [5] stereo, colour and face detection are combined, but motion information is not used. On the other hand, in [3] motion and stereo is used but the colour information is not analysed.

The present experimental system pretends to capture the human gestures and to be able to work in non-specific environments, without introducing new elements in the scenario or over the own human operator. The vision system uses different techniques as movement, colour, tracking and stereo to extract gesture information in the image sequence.

2. System description

The system is mainly constituted by three parts: movement detection, colour tracking and data fusion (figure1). The movement detection module extracts the human silhouette when the user moves in the scene. If the gesture is enough clear, the body parts (head and arm tips) are detected by means of a first analysis of the

silhouette. From these first localisation in the image, a continuous analysis of the colour distribution of the body parts shows the predominant skins colours. Once the system learned the colour of the user skin, a second process starts to work in parallel with the movement detection.

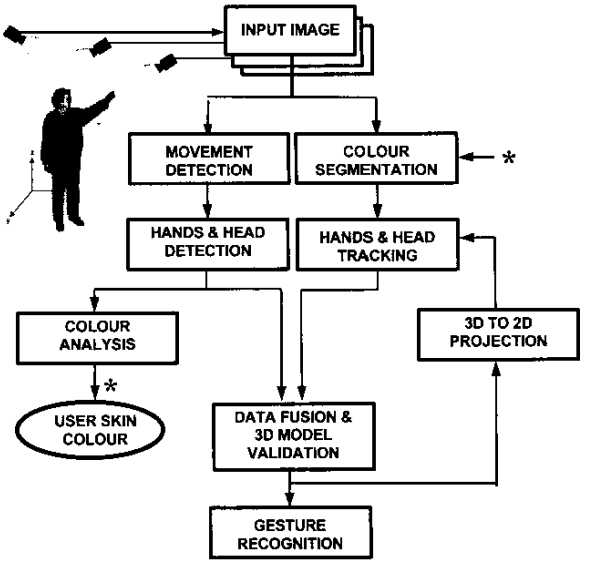


Figure 1. System diagram

This second process carries out a skin colour tracking process to try and guess the position of the hands and face. This process could overcome the drawback of the movement detection part that lost the body parts when the user maintains their hands close to the silhouette of the body. Due to that, this colour segmentation and tracking process uses a low level tracking scheme, the positions of the body parts must be re-established after a period of occlusion. This re-establishment is done by means of the movement detection process that is always working.

The data fusion module collects all proposals from both modules to build a more complete and valid description of the user movements. As the system works with several calibrated cameras, it is enough to detect the same body part in two different cameras to infer its 3D position on the scene. Also, the system infers the position of the body parts in the images from their estimated position in the scene, in order to complete the body tracking loop. This information is the input of the gesture interpretation module.

2.1. Movement detection

Image segmentation is one of the main problems to face up in computer vision. The selection of the adequate segmentation technique greatly depends on the kind of scene to analyse and its environment conditions. When the scenes are complex and it is not able to find features discriminating enough, it is possible to resort on the analysis of image sequences to detect the movements of the user. Image subtraction is one of the most common movement detection technique because its computing simplicity. Nevertheless, this technique is quite sensible to noise and lighting fluctuations that prevent the use of threshold values sensible enough. For this reason image subtraction is used in specific plateau, under controlled lighting conditions and usually with high contrast between the moving object and the background. To improve the performance of the system in 'complex' scenarios the comparison pixel by pixel is carried out from the estimated gradient vector instead of using only its absolute value (figure 2). In this case, images subtraction is performed as follows:

$$| \vec{G}_t(x,y) - \vec{G}_{t-1}(x,y) |$$

(Equation 1)

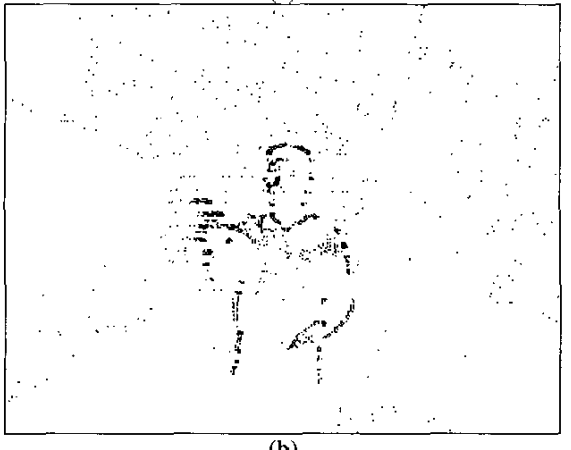
Where $\vec{G}_t(x,y)$ represents the measurement of the gradient vector computed at position (x,y) of the gray image $I(x,y)$, taken at instant t. The movement detection is carried out very fast due to a hardware implementation. The image comparison hardware adapts the time of comparison of two images to the dynamics of the movement to achieve better detection results.

2.2. Hands and head detection

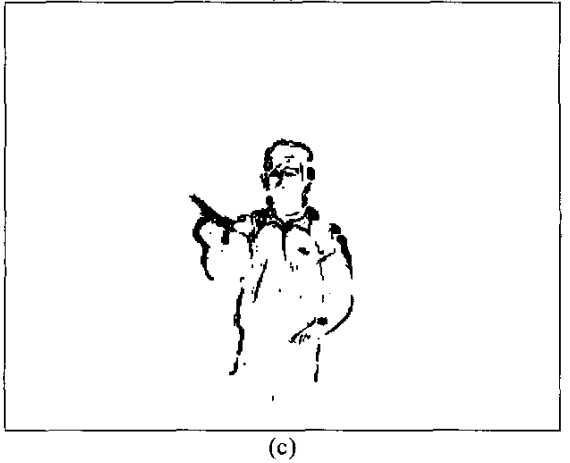
The detection module can compute reference position of the different body parts when they appear clearly in the image for the first time or after a short occlusion period.

Singular points selected are extreme points of the silhouette (figure 3.a.) and the head detection is 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 depending on their relative position and according to the main axis. Figure 3.b shows the hands and head candidates. It can be seen how this process fails when only silhouette information is used and the arm is close to the user body. The problem is solved when colour information is added.

(a)

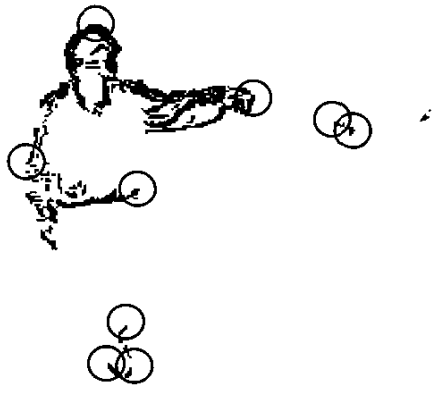


(b)



(c)

Figure 2. Image Movement Detection.
a) Original image,
b) Result using subtraction,
c) Result using gradient vectors difference



(a)

(b)

Figure 3. Results of the detection process:
a) singular points candidates,
b) hand and head candidates

2.3. Colour segmentation and tracking

Skin colour is the basis of the colour segmentation process. In an initial learning step, a colour analysis is performed in areas identified as head or hands by the hands and head detection process (explained in section 2.2). For each area, a colour histogram is computed from pixels inside the extracted silhouette. User skin colour is learned as the main colour that appears inside of these areas (one for the head and, depending of occlusions, two for the hands) and it is similar to one of the predefined racial skin colour. The red/green ratio [9] has been used

to describe the skin colour interval (figure 4.b). The use of red/green ratio avoids the need of a camera chromatic calibration, unlike other skin colour extraction methods. The estimation of the user skin colour is dynamically actualised. When the movement analysis process detects the human body with enough confidence (in clear postures), the colour analysis is carried out, and the user skin colour and skin thresholds are adjusted again with a recursive filter. In this way, illumination changes and colour camera drift.

Once user skin colour is defined, the colour segmentation is carried out in two steps. First of all, pixels with a colour value similar to user skin colour that are inside of the three interest image areas (the expected image position of head & hands) are selected. In a second step, the connected regions are labelled (fig.4.c).

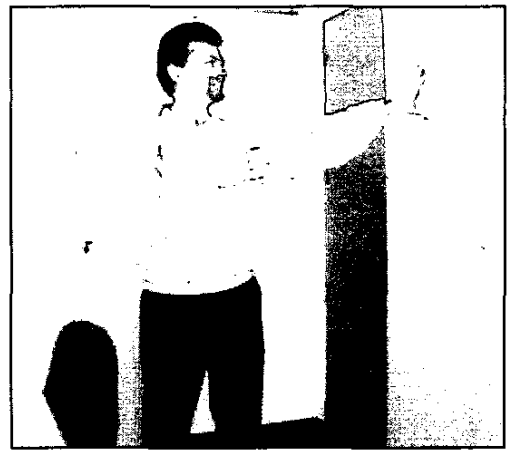
For tracking purposes, each region is characterised by its colour, position and area. The tracking process uses a confidence measure *CM* for a match between region *i* of image *t* and region *j* of image *t+1*, which is obtained by means of:

$$CMij = \alpha \cdot \text{Sigmoid}\,(K_1 \cdot |\text{colour difference}|) + \beta \cdot \text{Sigmoid}\,(K_2 \cdot |\text{position difference}|) + \gamma \cdot \text{Sigmoid}\,(K_3 \cdot |\text{area difference}|) \quad \text{(Equation 2)}$$

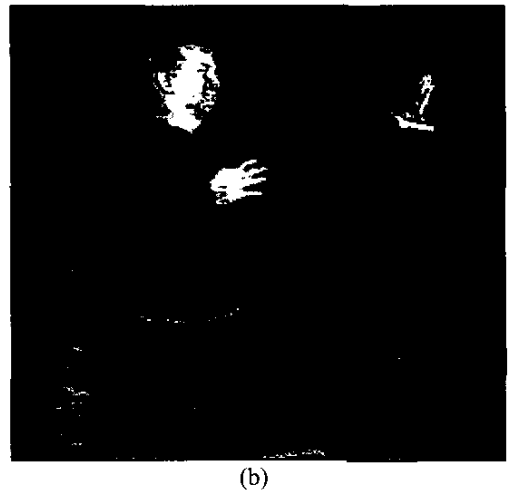
For each camera, the output of the colour segmentation and tracking process are the image co-ordinates of the centre of gravity of the head and hands regions, and the associated *CM* values.

2.4. Data fusion and 3D model validation

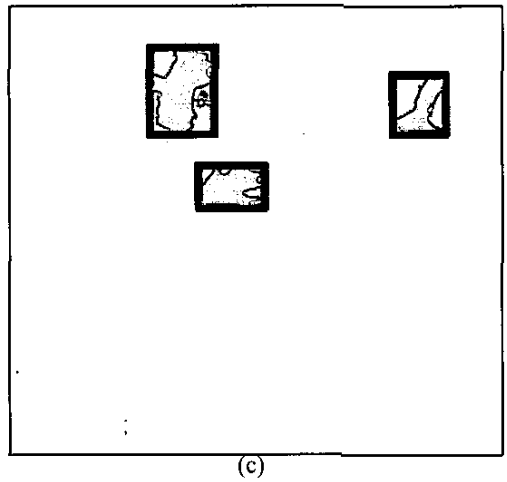
Data fusion is carried out analysing the movement-colour coherence. When both information are available, movement and colour processing are independent, and the skin colour regions segmented outside the movement image area are rejected. If poor movement is detected in the image, the skin colour regions segmented are used to detect head and hands. At first, face and hands from the human body are assumed to be located in the scene by triangulation of the results of the movement detection module. Also their prominent colour has been determined so a feasible colour segmentation of these body parts is possible.



(a)



(b)



(c)

Figure 4. Skin segmentation.
a) original image, b) skin threshold and c) skin areas.

Now, the problem can be exposed as tracking the segmented blobs (corresponding to face, left hand and right hand) in every image by using colour information and 3D movement coherence.

During operation, colour tracking module provides to data fusion module with the position of the segmented blobs and the *CM* value representing the reliability of the tracking. This information must be contrasted with that provided by movement detection module.

Fusion algorithm is divided into two stages. First of all, it validates data coming from every different view. And after, it performs a test of 3D dimensional coherence using all the projections selected in the first stage.

The first stage contemplates four main cases:

Case A. The blob has been reliably tracked (*CM* high).

Case A.1. Movement detection reaffirms that the blob belongs to a body part. This is the most favourable case. Here co-ordinate data of the projection of this blob is taken into account in 3D test of coherence.

Case A.2. No movement is detected in the proximity of the blob. This can be owing to:

a) The body part is temporally stopped.
b) The body part is moving but it rests on the body.
c) Colour tracking has segmented a blob belonging to the background.

Option c) will be find out from others by 3D test of coherence as time goes on. If some fixed object is tracked, quickly it will begin to produce some incoherent position data (in front of others views).

Case B. The blob position is not reliable (*CM* low) or the blob does not appear.

Case B.1. Movement is detected in the proximity of the predicted blob position. If the position of this movement is coherent with tracked blobs from others views then interest colour of the body part of this view must be updated in order to perform a most reliable segmentation.

Case B.2. No movement is detected in the proximity of the predicted blob position. This could be owed to an occlusion of the body part from this view. If the blob can not be tracked from other views the system does a short-term prediction.

After, a test of the coherence of the results of the fusion process is performed. This test is based on the 3D measurements that are obtained by triangulation. For data validation the elemental anthropomorphic constraints are used. The dimensions of the user are taken from the first measurements of the system, when the user begins to gesticulate. Fig. 5 is an example image in which the system shows the results of a tracking sequence of the head and hands of a user pointing at an object.

Figure 5. Hands and head detection and tracking results. System output of an image sequence for gesture interpretation.

3. Conclusions

This Human Machine Interface based on the tracking of a person's gestures has given satisfactory results in applications where the orders to transmit from the human operator to the machine are relatively simple, as up-down, stop, turn, approach or go.

The integration of colour segmentation and movement detection benefits the robustness of the system. The movement detection hardware adapts the time of image comparison to the dynamics of the movement to achieve better detection results. The stereo information is used at the finals steps of the system when the data volume is more reduced and has superior information quality than in the low-level steps. Colour is a useful information that allows to identify the segmented regions and facilitate a further processing.

Tue use of a dedicated hardware to obtain the movement image and the colour segmented image allows to operate at a rate of up to 4 Hz, using a PC computer.

With the framework presented, it is possible to have a low-cost, fast enough, and robust system that can be used as an input module of a gesture recognition system [1].

References

[1] Amat J., Casals A., Frigola F., Pages J.. Possibilities of Man-Machine interaction through the perception of human gestures. In journal Contributions to Science 1(2):159-173, Institut d'Estudis Catalans, Barcelona , 1999-2000.

[2] Azarbayejani A. and Pentland A. Real-time self-calibrating stereo person tracking using 3-D shape estimation from blobs features. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition, Vienna, 1996.

[3] Azoz, Y.; Devi, L.; Sharma, R. Reliable tracking of human arm dynamics by multiple cue integration and constraint fusion. In proc. of IEEE Computer Vision and Pattern Recognition, pp. 905–910, 1998.

[4] Brand J. et ali. A comparative assessment of three approaches to pixel-level human skin-detection. In proc. of IEEE International Conference on Pattern Recognition 2000, 15(1), pp. 1056-1059

[5] Darrell, T.; Gordon, G.; Harville, M.; Woodfill, J. Integrated person tracking using stereo, color, and pattern detection. In proc. of IEEE Computer Vision and Pattern Recognition, pp. 601-608, 1998.

[6] Feraund, R.; Bernier, O.J.; Viallet, J.-E.; Collobert, M. A fast and accurate face detector based on neural networks. In IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 23 Issue: 1, pp. 42 -53, 2001.

[7] Graf, H.P.; Cosatto, E.; Gibbon, D.; Kocheisen, M.; Petajan, E. Multi-modal system for locating heads and faces. In proc. of the Second International Conference on Automatic Face and Gesture Recognition, pp. 88 –93, 1996.

[8] Ohya J., Miyasato T., Nakatsu R. Virtual Reality Technologies for Multimedia Communicatiosn. In Mixed Reality – Merging Real and Virtual Worlds. Ed. Y. Ohta, H. Takamura. Springer-Verlag, NY, 1999.

[9] Ogihara A., Shintani A., Takamatsu S. and Igawa S. Speech recognition based on the fusion of visual and auditory information using full-frame colour image. IECE Trans. Fundamental, pp. 1836-1840, 1996.