

Real time segmentation and tracking of face and hands in VR Applications

José M. Buades, Francisco J. Perales, and Javier Varona

Unidad de Gráficos y Visión por Ordenador
Universitat de les Illes Balears (UIB)
C/Valldemossa Km. 7.5, 07122 - Palma de Mallorca - España
{ josemaria.buades, paco.perales, vdmijvg4 }@uib.es
<http://dmi.uib.es/research/GV/>

Abstract. We describe a robust real-time 3D tracking system of the extreme limbs of the upper human body, i.e., the hands and the face. The goal of the system is that it can be used as a perceptual interface for virtual reality activities in a workbench environment. The whole system includes an input capture and calibration module, a real time color segmentation module, a data association and tracking module and finally a visualization VRML and H-ANIM procedure. The results of our probabilistically skin-color segmentation are skin-color blobs. Then, for each frame of the sequence our algorithm labels the blobs' pixels using a set of object state hypothesis. This set of hypothesis is built from the results of previous frames. The 2D tracking results are used for the 3D reconstruction of limbs position in order to obtain the H-ANIM visualization results. Several results are presented to show the algorithm performance.

1 Introduction

In actual computer systems the interaction is going to a non-contact devices, by means of perceptual and multimodal user interfaces (PUIs and MUIs). That's means that the system allows the user to interact without physical contact with the machine; this communication can be carried out with voice or user gesticulation capture. We are especially interested in visual information, so recognize the human presence in color video images. Exist many precedent work developed in this hot topic: in [3] including finger detection with multi-scale color approach, in [8] in real time for virtual reality applications but in 2D, in [9] to human grasping but using infrared images including a hand model with 15 joints (20 D.O.F.). For our purposes, we would like to define a general, robust and efficient system that can be used with non-expensive analogical or digital cameras (using day light spectrum) and can recover 3D hands and face pose of the human person in real time. Capture is carried out from digital IEEE1394 color cameras. The process is applied to stereo cameras to recover 3D positions. In the presented paper is used a high quality workbench but the system is adaptable to others less expensive systems.

The global process must detect a new user entering the system and analyze him/her to determine parameters such as hair color and clothes. Once the user who is going to interact with the machine has been detected, the system starts to track interesting regions such as the head, hands, body and joints, using information obtained in the user detection task. The input data for the gesture interpretation process are the position and orientation of these regions. This process will determine which gesture the user has carried out. Next, these gesture data are sent to the execution process, which ends the process by performing the action that has been specified, and so completing the feedback process. This is a very complex and challenging task, so we isolate sub problems to make it more tractable. Then we present in this paper the face and hands segmentation and tracking.

Our previous work, define a segmentation algorithm based in functional minimization procedure [10, 11] but this robust method is time consuming and non applicable to virtual reality application where the delay time must be very short (less 40 ms). So in this paper we propose a more simple segmentation algorithm and heuristic classification and tracking procedure based in a set of rules.

In the following section, we explain briefly the tracking method proposed including the skin color pixel detection and training. Also we present our hypothesis generation and applications and finally we conclude with some examples and visualization results in VRML format and H-Anim [12] avatar compliant. This work is an adapted version from a more computational cost version of [5, 10] improving the computational efficiency to apply in virtual reality environments.

2 Tracking of Face and Hands

The tracking algorithm is divided in two parts: skin-color pixel detection and data association. For each frame, first, we segment the skin-color pixels using a previous learned probabilistical skin-color model and a blob analysis. Next, we use a hypothesis based data association algorithm to label the skin-color detected pixels. This algorithm uses a simple prediction step to the state obtained in the previous frame to calculate the new hypothesis.

2.1. Skin-color pixel detection

As we have explained before, we have another proposed method based in more solid mathematical background but are very expensive in computational cost. So we would like to define new systems that can be run in real time and oriented to virtual reality applications, in particular interactive workbench activities with no contact devices.

The assumption that color can be used as a cue to detect faces and hands has been proved in several publications [1,2,10,11]. Usually, it is necessary to model the actor's skin-color in a previous step. In our work we use only one image of the actor to build this model. The user selects manually the regions of the image that contains skin-color pixels, also this procedure can be done in a automatic way, we are working in this process. Next, we transform these pixels from the RGB-space to HSL-space to

take the Hue and the Saturation values for each pixel, that is, the chroma information. These values for the selected pixels are the data samples used to learn the skin-color model:

$$\mathbf{x} = (x_1, \dots, x_n),$$

where n is the number of samples and $x_i = (h_i, s_i)$. We have proved several statistical models and finally the best results has been obtained using a Gaussian model:

$$\boldsymbol{\mu} = \frac{1}{n} \sum_i x_i, \quad \boldsymbol{\Sigma} = \frac{1}{n} \sum_i (x_i - \boldsymbol{\mu}) \cdot (x_i - \boldsymbol{\mu})^T \quad (1)$$

Once the skin-color model is built we can calculate the probability that a pixel is skin colored:

$$P(x) = \frac{1}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma}|}} e^{\frac{1}{2}(x-\boldsymbol{\mu})(x-\boldsymbol{\mu})^T}. \quad (2)$$

Then, in order to detect the skin-color pixels in each frame of the sequence we compute the probability for all the image pixels. In Fig. 1 can be show some results of this process, it can be seen how this model performs well in several environmental conditions.

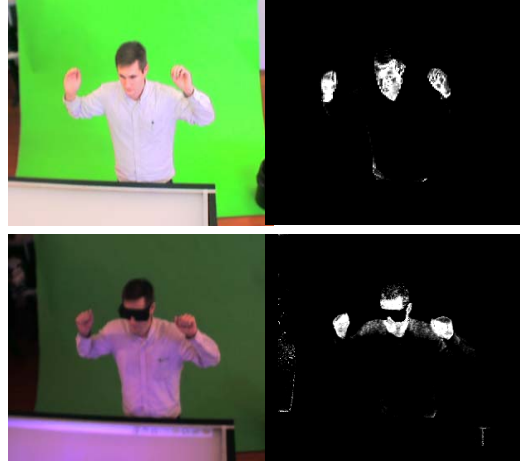


Fig. 1. Examples of skin-color probabilities using the model proposed. In the first row we show one frame for a sequence with normal illumination, and in second row with a proper illumination to correctly see the workbench screens.

However, it is necessary to refine the results of pixel probabilities to obtain an image with only the pixels belonging to the interesting blobs: the face and the hands. Therefore, we define a posterior hysteresis process to obtain the skin-color blobs. All

the image pixels with probability $P > T_{max}$ are considered as being skin colored. These pixels constitute the seeds of potential blobs. More specifically, image pixels with probability $P > T_{min}$, where $T_{min} < T_{max}$, that are immediate neighbours of skin-color pixels are recursively added to each blob. We actually use $T_{min} = 0.05$ and $T_{max} = 0.15$. In Fig. 2 we show the final results of skin-color detection.

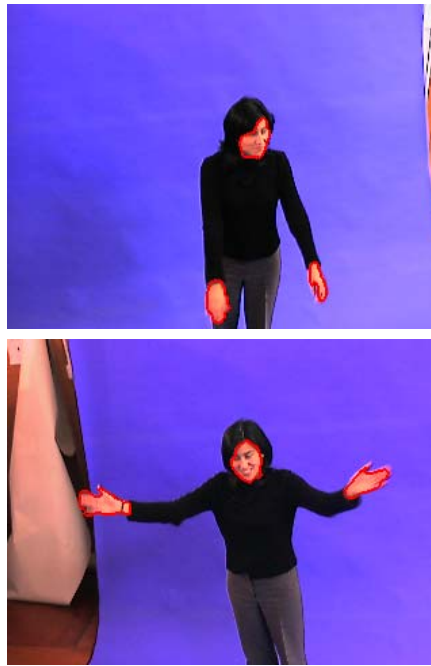


Fig. 2. Contours of skin color regions after the hysteresis process.

The final step in the skin-color pixel detection is the computation of the attributes of each blob (centroid and size) that will be used in the tracking process to represent the state of actor's face and hands.

2.2. Data Association

At each frame, we represent the state of the face and hands using a set of *hypothesis*. Each hypothesis is composed by the position and size of the actor's face or hand (from now, limb). The proposed method to track a limb operates as follows: for the frame of time t the aim is to associate the skin-color blobs with the limb hypothesis in time $t-1$. The objective of this association is to propagate in time the labels that represent the face, the left hand and the right hand, and to detect when a limb appears in the scene and disappears from the field of view.

We assume that at time t , M blobs have been detected using the skin-color model,

$$B = \{b_1, \dots, b_j, \dots, b_M\}.$$

Each blob b_j , corresponds to a set of connected skin-color pixels. Note that a blob may correspond to one of many limbs. As an example, two crossing hands are two different limbs that appear as one blob when one occludes the other. Let N the number of visible limbs present in the viewed scene at time t ($N < 4$). We assume that an ellipse can approximate the spatial distribution of the pixels.

Then, the state of a limb i can be represented by:

$$h_i = (c_{x_i}, c_{y_i}, \alpha_i, \beta_i),$$

where (c_{x_i}, c_{y_i}) is its center and (α_i, β_i) are the lengths of semiaxis. The union of limb states is denoted by:

$$H = \{h_i\}, \quad i \leq 3.$$

Tracking amounts to determining the relation between limb states (h_i) and observations (b_j) in time. The first subproblem can be present when a new limb appears in the field of view. In order to cope with this problem, we define the distance $D(p, h)$ of a pixel $p = (x, y)$ from a ellipse (i.e., limb state) as follows:

$$D(p, h) = \sqrt{\mathbf{v} \cdot \mathbf{v}^T}, \quad (3)$$

where

$$\mathbf{v} = \left(\frac{x - c_x}{\alpha}, \frac{y - c_y}{\beta} \right). \quad (4)$$

From the definition of $D(p, h)$ it turns out that its value is less than 1 if p is inside ellipse h , and greater than 1 if it is outside. Considering a state h and a point p belonging to a blob b , if distance is less than 1, we conclude that the blob b support the existence of the part hypothesis h and that part hypothesis h predicts blob b . Now, if a new part appears in the scene implies that none of the existing hypothesis predicts the existence of the corresponding blob b :

$$\forall p \in b, \quad \min_{h \in H} \{D(p, h)\} > 1. \quad (5)$$

The Eq.5 describes a blob with empty intersection with all ellipses of the existing limb hypothesis. Algorithmically, at each time t , all detected blobs are tested against the criterion of Eq.5. If a blob not belongs to any hypothesis, a new limb hypothesis is created and their corresponding parameters correspond to the blob parameters.

After appearing limbs have been detected, all the remaining blobs must support the existence of limb hypothesis. The main task of the tracking algorithm is to associate blobs to limb hypothesis. We use two rules to make this association:

- **Rule 1:** if a pixel p of a blob is inside of a limb hypothesis then this pixel is labeled with the hypothesis number. Formally

$$\mathbf{R}_1 = \{p \in B \mid D(p, h) < 1\}. \quad (6)$$

- **Rule 2:** if a pixel p of a blob is outside of all the ellipses, then it is labeled to the hypothesis number that is closer to it. Formally

$$\mathbf{R}_2 = \left\{ p \in B \mid D(p, h) = \min_{k \in H} \{D(p, k)\} \right\}. \quad (7)$$

These two rules permit the treatment of limbs occlusion if pixels belonging to a blob can be labeled with more than one hypothesis number using the rule 2, see Fig. 3.



Fig. 3. Occlusion treatment using multiple labeling and rule 2. Object hypothesis are depicted using ellipses.

Another interesting case can happen when finish the occlusion and then a hypothesis is supported by more than one blob (split case). In this case, the hypothesis is assigned to the blob with which it shares the largest number of pixels. Finally, a limb hypothesis should be removed either when the limb moves out of scene. Following our algorithm, a hypothesis is removed when:

$$\forall p \in B, D(p, h) > 1. \quad (8)$$

To conclude, once all the above case have been treated for a frame, the resulting limb hypothesis are used to maintain the limb states, that is to know the position and size of head and hands. Once the state has been estimated, they are propagated in time to the next frame using a linear scheme of prediction:

$$\begin{aligned}\hat{C}_i(t) &= C_i(t) + \Delta C_i(t), \\ \Delta C_i(t) &= C_i(t) - C_i(t-1),\end{aligned}\tag{9}$$

where $C_i(t) = (c_{x_i}, c_{y_i})$. The above equations say that a part will maintain the same velocity on the image plane. Some experimental results showed in Fig. 4.

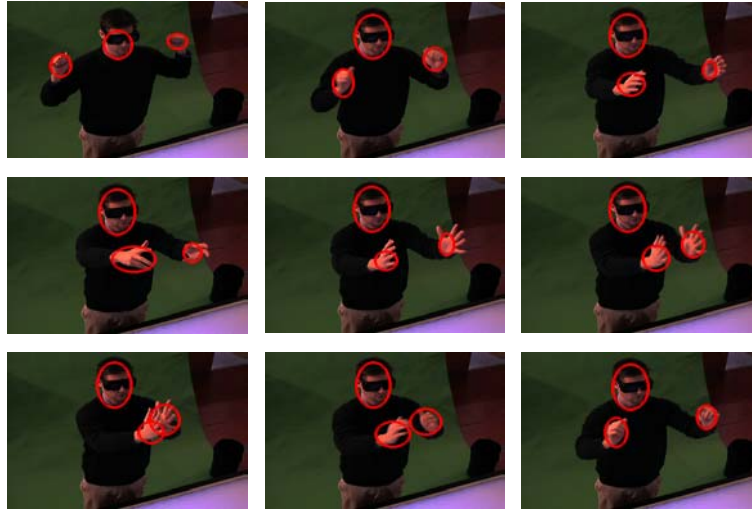


Fig. 4. Experimental results of 2D-Tracking of head and hands.

2.3. 3D position estimation.

At the initialization step a calibration process is carried out using OpenCV library functions and a check board as calibrator object. The calibration process is showed in Fig. 5.

Finally, we use the calibration parameters computed and the state of each limb, see Fig. 6, in order to estimate their 3D positions.

The complete procedure can be seen in Fig. 7.

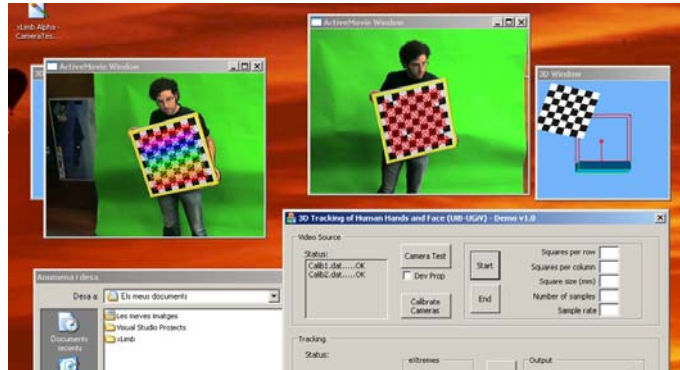


Fig. 5. Calibration process in the initialization phase.



Fig. 6. Extreme limbs states used for 3D estimation in a workbench session.

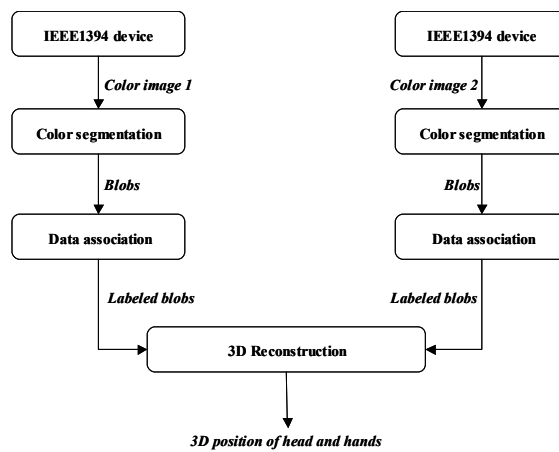


Fig. 7. Complete procedure: color segmentation, data association and 3D reconstruction.

3. Visualization using VRML and H-Anim avatar on Workbench

Hands and head tracking system obtain 2D coordinates from two stereo cameras. Due to a previous calibration process, is able to compute 3D position. This information is needed to display the data in real time on a Barco Consul Workbench. That's mean that a human virtual avatar can be reproduced in a remote workbench. The final objective of this process is to recover in high precision the 3D position and orientation of the hands and face in the interactive applications. In the human model representation we use the H-Anim standard that's mean that we can collaborate with standard VRML models.

At the moment we use IEEE1394 digital videocameras and high quality workbench but the system must be portable in near future to low cost systems (web cams, and domestic virtual reality environments).

3D position is computed for every blob, projecting the centroid of the blob on each image to infinity and compute 3D coordinate as the nearest point to this two lines.

Hands and head are modeled with a VRML file, extracted from a H-Anim humanoid, see two different reconstructed views in Fig. 8.

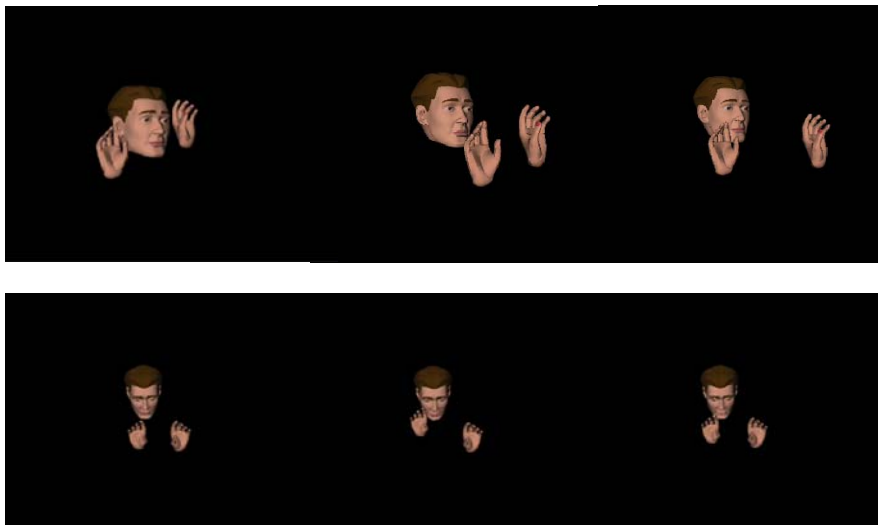


Fig 8: 3D Reconstruction in VRML.

In the web site <http://dmi.uib.es/research/GV/amdoResults2004> you can view and download some video examples of captured data and VRML demos. Also we can combine the synthetic avatar with real human in workbench in real time, as we can see in Fig 9.

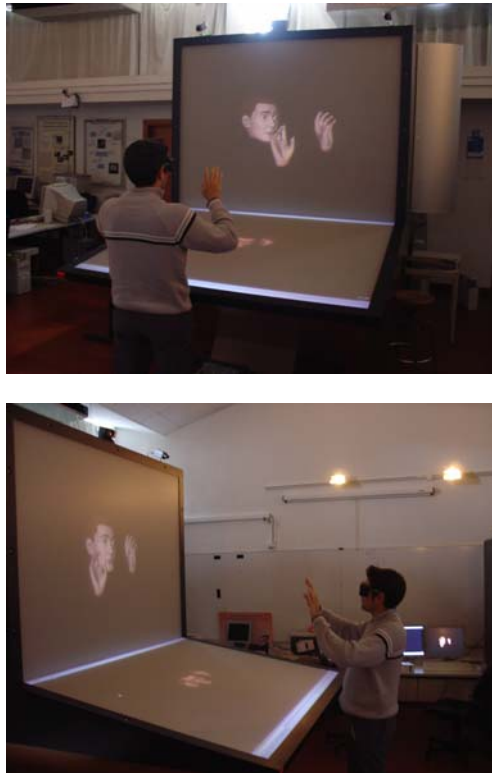


Fig 9: An example of HCI in virtual reality applications

At the moment only 3d position is really computed, so we must to consider the 3D orientation and check the precision of this reconstruction against commercial tracking systems based in ultrasound technology. This comparison must be evaluated in relation commercial cost/precision results.

Also we plan to reconstruct the fingers positions to understand the actions that the user plans to do in relation with the interactive process in workbench. So a more carefully analysis must be done [9].

4. Conclusions and future work

In this paper we have proposed a new system for 3D tracking of human extreme limbs, hands and face, for HCI in real time. Moreover, it analyses the user to determine parameters that will be useful for tracking process that's include an heuristic method based in a well define set of rules. The pixel segmentation process is based on the colour pixel classification hypothesis based data association algorithm. Besides, the process is carried out in real time. The software implementation is efficient and OOP. The result of this process is the tracking and reconstruction of face and hands in 3D to be able a human computer interaction system for virtual reality environments.

It remains as future work to do tracking of interesting body parts and to interpret movements in order to carry out action recognition that the user is performing. So we must to improve the precision results and more precise primitives must be defined to track fingers and face parts to know exactly eyes directions and gaze orientation.

Acknowledgements

The projects TIC2003-0931 and TIC2002-10743-E of MCYT Spanish Government and European Project HUMODAN 2001-32202 from UE V Program-IST have subsidized this work. Also we would like to express our gratitude to all members of Computer Graphics and Vision group at UIB by their collaboration.

References

1. Bradski G.R.; "Computer video face tracking for use in a perceptual user interface". Intel Technology Journal, Q2'98, 1998
2. Comaniciu, D.; Ramesh, V.; "Robust detection and tracking of human faces with an active camera". Proceedings of the Third IEEE International Workshop on Visual Surveillance, 2000; Page(s): 11 -18.
3. Lars Bretzner¹, 2 Ivan Laptev,¹ Tony Lindeberg¹. Hand Gesture Recognition using Multi-Scale Colour Features, Hierarchical Models and Particle Filtering, Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR.02), 0-7695-1602-5/02 © 2002 IEEE
4. H.D. Cheng, X.H. Jiang, Y. Sun, JinGli Wang "Color Image Segmentation: Advances and Prospects", Journal of Pattern Recognition 34, (2001), pp. 2259-2281
5. M. Gonzalez "Segmentación de imágenes en Color por método variacional". Proc. Del XIV C.E.D.Y.A. y IV C.M.A. pp 287-288, 1995.
6. I. Haritaoglu, "W4: Real-Time Surveillance of People and Their Activities" IEEE. Transactions on Pattern Analysis and Machine Intelligence, vol 22 No8, pp 809-830, 2000
7. H. Sidenbladh, M.J. Black and D.J. Fleet "Stochastic Tracking of 3D Human Figures Using 2D Image Motion" ECCV 2000.
8. Kenji Okay Yoichi Satoy Hideki Koikez, Real-time Tracking of Multiple Fingertips and Gesture Recognition for Augmented Desk Interface Systems, Proceedings of the Fifth IEEE

- International Conference on Automatic Face and Gesture Recognition (FGR.02), 0-7695-1602-5/02, 2002 IEEE
9. Koichi Ogawara, Kentaro Hashimoto, Jun Takamatsu, Katsushi Ikeuchi, Grasp Recognition using a 3D Articulated Model and Infrared Images, Institute of Industrial Science, Univ. of Tokyo, Tokyo, Japan, Institute of Industrial Science, Univ. of Tokyo, Tokyo, Japan, Fuji Xerox Information Systems Co.,Ltd. Tokyo, 150-0031, Japan
 10. J.M. Buades, M. Gonzalez, F.J. Perales. "A New Method for Detection and Initial Pose Estimation based on Mumford-Shah Segmentation Functional". IbPRIA 2003. Port d'Andratx. Spain. June 2003. pp 117-125, LNCS 2652.
 11. J.M. Buades, M. Gonzalez, F.J. Perales. "Face and Hands Segmentation in Color Images and Initial Matching" International Workshop on Computer Vision and Image Analysis. Palmas de Gran Canaria. Dec. 2003. pp 43-48.
 12. HANIM 1.1 Compliant VRML97. <http://ece.uwaterloo.ca/~h-anim/index.html>