Face and Hands Segmentation in Color Images and Initial Matching with a Biomechanical Model

Jose Maria Buades Rubio, Manuel González Hidalgo, Francisco Jose Perales
Unidad de Graficos y Vision. http://dmi.uib.es/research/GV
Departamento de Matemáticas e Informatica
Universitat de les Illes Balears
Crta Valldemossa Km 7.5
Spain E-07122, Palma de Mallorca, Baleares
{josemaria.buades, dmimgh0, paco.perales}@uib.es

ABSTRACT

In this paper we describe a robust and efficient procedure to detect skin region with homogeneous color values in monocular indoor images. The mathematical background is a functional segmentation process adapted to color images. The idea is to obtain a system to detect the hands and face in a sequence of monocular or binocular images. The main criteria are to select the best probabilistic skin regions in the previous segmentation process. In particular we are focuses in hands and faces because are visible in any avatar in general. We also use a biomechanical restriction to reach this initial estimation, so the system is to be able to detect hand-face-hand configuration to be used as an initialization procedure to human detection in a more general human recognition system. Several analytical rules are applied to reach these specific configurations.

We introduce the segmentation process and region classification criteria (hands, face, head and upper-torso). Finally, we present some results over a significant statistical number of potential users. Also we implement the algorithm efficiently in order to obtain real time results processing standard video format.

Keywords

Human Computer Interaction. User Detection. Initial Pose Estimation. Color Segmentation.

1. INTRODUCTION

In actual computers systems the interaction is going to a non-contact devices. That's means that allow 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 colour video images. Also we would like to define a general, robust and efficient system that can be used with non-expensive cameras and digitalizing cards. Capture is carried out from colour cameras. The process is over 2D images but we can extend easily to stereo cameras.

The global process must detect a new user entering the system and analyse him/her to determine parameters such as hair colour 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.

In the following section, we explain briefly the mathematical background of the segmentation process. Section 3 introduces the main method to detect the user in front of the camera and carefully explains the analysis process and parameters needed for a future tracking process. Finally, we conclude with some new and extended results including a set of colour images and conclusions, future works and references. This work is a new version of [Bua03] improving the segmentation process and the result ratio in classification process.

2. MULTICHANNEL SEGMENTATION ALGORITHM

Image segmentation is the first step in data extraction for computer vision systems. Achieving good segmentation has turned out to be extremely difficult, and is a complex process. Moreover, it depends on the technique used to detect the uniformity of the characteristics sought between image pixels and to isolate regions of the image that have this uniformity. Multiple techniques have been developed to achieve this goal, such as contour detection, split and merging regions, histogram thresholding, clustering, etc. A Survey can be found in [Che01].

In color image processing, pixel color is usually determined by three values corresponding to R (red), G (green) and B (blue). The distinctive color sets have been employed with different goals, and specific sets have even been designed to be used with specific segmentation techniques [Che01].

We define a color image as a scalar function $g = (g^1, g^2, g^3)$, defined over image domain $\Omega \subseteq \Re^2$ (normally a rectangle), in such a way that $g\colon \Omega \to \Re^3$. The image will be defined for three channels, under the hypothesis that they are good indicators of autosimilarity of regions. A segmentation of image *g* will be a partition of the rectangle in a finite number of regions; each one corresponding to a region of the image where components of *g* are approximately constant. As we will try to explicitly compute the region boundaries and of course control both their regularity and localization, we will employ the principles established in [Koe94, Mor95] to define a good segmentation. To achieve our goals we consider the functional defined by Mumford-Shah in [Mum89] (to segment gray level images) which is expressed as:

$$E(u,B) = \int_{\Omega} \sum_{i=1}^{3} (u^i - g^i)^2 \, d\mu + \lambda \ell(B) \qquad (1)$$

where *B* is the set of boundaries of a homogenous region that define a segmentation and *u* (each u^k) is a mean value, or more generally a regularized version of *g* (of each g^k) in the interior of such areas. The scale parameter λ in the functional (1) can be interpreted as a measure of the amount of boundary contained in the final segmentation *B*: if λ is small, we allow for many boundaries in *B*, if λ is large we allow for few boundaries.

A segmentation *B* of a color image *g* will be a finite set of piecewise affine curves - that is, finite length curves - in such a way that for each set of curves *B,* we are going to consider the corresponding *u* to be completely defined because the value of each u^i coordinate over each connected component of $\Omega \setminus B$ is equal to the mean value of g^i in this connected component. Unless stated otherwise, we shall assume that only one *u* is associated with each *B*. Therefore, we shall write in this case *E*(*B*) instead of *E*(*u*, *B*). A segmentation concept which is easier to compute is defined as follows:

Definition 1. *A segmentation B is called 2-normal if, for every pair of neighboring regions O_i y O_j , the new segmentation B' obtained by merging these regions satisfies E(B') > E(B).*

We shall consider only segmentations where the number of regions is finite, in other words $\Omega \setminus B$ has a finite number of connected components and the regions do not have internal boundaries.

A more detailed explanation of the concepts and their mathematical properties can be consulted in [Koe94, Mor95] and we can see the properties of the functional in [Mum89,Mor95]. The use of multichannel images (eg. color images) can be seen in [Mor95, Gon96]. We shall use a variation of segmentation algorithm by region merging described in [Koe94] adapted to color images.

The concept of 2-normal segmentations synthesizes the concept of optimal segmentation we are looking for, and it lays on the basis of the computational method we use. The 2-normality property is well adapted for the construction of an algorithm based on region growing by merging neighboring regions. Two regions will be merged if this operation reduces the energy. At each step we need to compare the balance of energy if we remove a common boundary $\partial(O_i, O_j)$ of two neighboring regions O_i, O_j. If *B* is 2-normal, one has $E(B) \le E(B - \partial(O_i, O_j))$, which, in the case of a piecewise constant function *u*, implies the balance

$$\lambda \ell(\partial(O_{i,} O_j)) \le \frac{|O_i| \cdot |O_j|}{|O_i| + |O_j|} \left(\sum_{k=1}^{3} (u_i^k - u_j^k)^2 \right) \qquad (2)$$

where $|\cdot|$ is the area measure and u_i, is the approximation of *g* on O_i to compute the data for evaluating the balance for each region O_i we associate its area $|O_i|$ and we can compute $u_i^k = \int_{O_i} g^k / |O_i|$ for k=1, 2, 3.

We call equation (2) the merging criterium. We decide to remove the common boundary $\partial(O_i, O_j)$ of O_i and O_j if this equation is not satisfied. By repeating this step, that is, by comparing the balance energy for deciding to join any two neighboring regions, we finally obtain a 2-normal segmentation for the scale parameter λ, a segmentation, i.e., where no further elimination improves the energy. Then, we have implemented a multiscalar algorithm and data structure similar to that used in [Koe94] but adapted to color images and real time processing.

The algorithm used the RGB components, because the segmentations obtained are very accurate to our goal. But the system is able to use another color space or color descriptor as we can see in [Che01]. Moreover, if it is needed it can weigh the channels used in order to obtain the segmentation.

3. USER DETECTION AND INITIAL POSE

The image is captured and segmented with the algorithm explained in the previous section and is then analyzed to determine whether it is a user or not, as we can see below in a work related with this topic [Har00]. If a user has been detected, the system studies him and obtains some parameters that will be useful in the tracking and analysis process [Sid00]. By applying this process directly to segmented images without using information from previous frames, the system is robust to background changing and variable illumination. The parameters obtained from the segmentation task are fixed in order to user interactions with upper torso (body, arms, hands and head). The system obtains the upper torso configuration: shirt, hair, hands and face. User detection process is waiting for a user located opposite the camera, with hands separated and at the same height that head, then it recognizes and later analyzes user configuration.

This module receives a segmentation of the captured image, analyzes every region and marks as skin region if its RGB medium value is in a characteristic color range of skin. To achieve more homogenous regions, neighboring skin regions are merged. This merging is carried out to avoid detecting a hand or the face in two neighboring regions, following the merging criteria:

$$\begin{gathered} \forall O_i, O_j / Neighbour(O_i, O_j) \wedge \\ Skin(O_i) \wedge Skin(O_j) \Rightarrow O_i \cup O_j \end{gathered} \qquad (3)$$

where $Neighbor(O_i, O_j)$ means that two regions are neighbors and $Skin(O_i)$ means that is a skin region.

After this, we obtain a skin region set, called β, where any pair of skin regions are separated.

For all ordered set of three regions included in β, we identify each one as face Z, left hand Y_1 and right hand Y_2, then we evaluate a criteria to determine whether this configuration is correct.

$$\underset{i,j,k}{Max}\left\{\varphi(O_i, O_j, O_k) : \forall O_i, O_j, O_k \in \beta\right\} \geq \alpha \qquad (4)$$

where α is a threshold probability and we call φ the user detection function. In this function we take into account the following:

- Central region, face, must be the biggest.
- Lateral regions, hands, have a similar area.
- Face region area A(Z) must be between a minimum Z^- and a maximum Z^+
- Hands area A(Y_1) and A(Y_2) must be between a minimum Y^- and a maximum Y^+
- Vertical position Y_1 and Y_2 should be similar and nearest possible to Z

The user detection function returns a value between zero and one that measures the probability that a user has been detected. From all possible combinations of Z, Y_1 and Y_2 the one with the greatest value, greater than a reference minimum value α, is chosen as the best configuration.

In order to apply the above algorithm, we need to fix the following values: a color range of skin to detect hand and face regions, a threshold probability α to discriminate non expected initial positions. To avoid high differences of hands we include an area similarity criterion, a maximum size of hand area is also necessary. All these parameters are used in order to discriminate bad detections.

All threshold values are established in relation with camera to user distance and image resolution. This distance is predefined by initial application setup.

After a user has been detected, the same image is analyzed to determine hair and shirt color. Region proposed as hair, X is the upper neighboring region of Z if A(X) / A(Z) relation is greater than a threshold, hair is discarded and is considered that it is a bald user.

To analyze shirt, the following algorithm is applied. Initially, shirt region W is the greatest region whose upper boundary is included in the boundary of Z (see Figure 1). Afterwards, neighboring regions of W are joined until Z is connected with Y_1 and Y_2 through W. A candidate region T_i chosen at every step i to be joined to W is in relation with: color space distance between mean color of T and W, and distance in pixels from T to Y_1 and Y_2.

With this process, the system detects a user and obtains useful data for the tracking system.

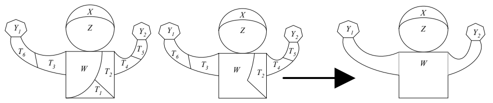

Figure 1. Shirt region detection

4. RESULTS

We have implemented the above algorithm in C++. It has been tested in 320x240 resolutions (Figure 2) and 640x480 standard video resolution (Figure 3). We initialize the multichannel segmentation algorithm with an initial segmentation wich is a grid of size T_x x T_y on the image, usually we take $T_x = T_y = 1$, 2 or 4. From this initial segmentation, the algorithm determines a 2-normal segmentation following the merging criterion described in (2) and the specifications of the algorithm described at the end of section 2. The stopping criterion can be: if the last level $\lambda = 2^n$ has been reached or if there is just one region left or if the desired number of regions is reached. In our displayed experiments the stopping criterion is to achieve a fixed number of regions. Then, we apply the algorithm described in section 3 where the selected parameters are detailed: Skin range color in HLS ([0-10], [20-230], [62-255])

In the two sequences of pictures we can see in Green the boundaries of hair region. The color Red is used for boundaries of hand and face regions, the centroid of these regions is visualized with a solid red square. In Pink we display the upper-torso boundary and finally we use Black and White for other regions detected for the segmentation algorithm.

In the first sequence we take a 2x2 initial segmentation and the system runs at 5 frames/second in a P4 1.6GHz. We display several different initial positions and cloth configuration; and we can see how the proposed method detects the interesting regions. In the second sequence, Figure 3, we display from a population of 30 persons, where 29 persons have been perfectly detected.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a new system for user detecting for HCI that does not use background substraction, therefore the system is robust to environment and illumination changes. Moreover, it analyzes the user to determine parameters that will be useful for a future tracking process. The region segmentation process based on the Koepfler-Morel-Solimini algorithm adapted to multichannel images is sufficiently good and beneficial for our aims. Besides, the process is carried out in real time. The software implementation is efficient and OOP. The result of this process is the input of a tracking and reconstruction of an intelligent human computer interaction system. 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. At the moment, we are working on particle filter tracking with a biomechanical model to reduce the search space solutions. Moreover, a stereo version is proposed to improve final results.

This paper is subsidized by the project IST-2001-32202 HUMODAN and CICYT TIC2001-0931.

6. REFERENCES

[Che01] 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

[Koe94] G. Koepfler, J.M. Morel, and S. Solimini, "Segmentation by minimizing a functional and the merging methods", SIAM J. on Numerical Analysis, Vol 31, No 1, Feb. 1994

[Mum89] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and variational problems", Communications on Pure and Applied Mathematics, XLII(4), 1989

[Mor95] J.M. Morel and S. Solimini. "Variational Methods for Image Segmentation", Birkhauser Verlag. 1995

[Gon96] 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.

[Har00] 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

[Sid00] H. Sidenbladh, M.J. Black and D.J. Fleet "Stochastic Tracking of 3D Human Figures Using 2D Image Motion" ECCV 2000.

[Bua03] 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

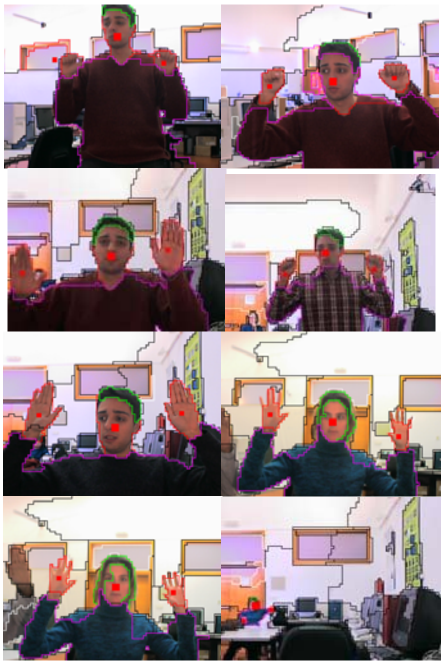

Figure 2. Some results obtained in real time with a Sony VFW-V500 camera. Images are 320x240 resolution in RGB color.

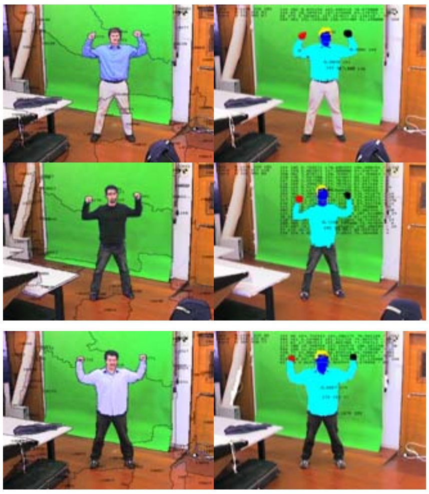

Figure 3. Some results from 30 persons test.