Hands detection and tracking for interactive multimedia applications

Vincent Girondel, Laurent Bonnaud, Alice Caplier

Vincent Girondel, Laurent Bonnaud, Alice Caplier
Laboratory of Image and Signal (LIS),
46 avenue Félix Viallet, 38031 Grenoble Cedex, France
<firstname.name>@lis.inpg.fr

HANDS DETECTION AND TRACKING FOR INTERACTIVE MULTIMEDIA APPLICATIONS

Abstract

The context of this work is a European project *art.live*[1] which aims at mixing real and virtual worlds for multimedia applications. This paper focuses on an algorithm for the detection and tracking of face and both hands of segmented persons standing in front of a camera. The first step consists in the detection of skin pixels based on skin colour: the HSI and YCbCr colour spaces are compared. The colour space that allows both fast detection and accurate results is selected. The second step is the identification of face and both hands among all detected skin patches. This involves spatial criteria related to human morphology and temporal tracking. The third step consists in parameter adaptation of the skin detection algorithm. Several results show the efficiency of the method. It has been integrated and validated in a global real time interactive multimedia system.

1 INTRODUCTION

art.live is a project which aims at developing an innovative authoring tool that enables artists and users to easily create mixed real and virtual narrative spaces and disseminate them in real-time to the public through the Internet (or any TCP/IP network). Globally speaking, the objective of *art.live* is to investigate some tracks in the huge field of mixed reality, that provokes the encounter of real and virtual worlds.

This can result in visual ambiances where people and objects entering a camera field of view are placed into a virtual environment (see figure 1) on large screens and/or on the Internet. These persons are then offered to interact with the story and with other people using another (possibly remote) instance of the system. For instance in the case of figure 1, the player aims at catching virtual butterflies. When all butterflies have been caught, the player himself is transformed into a butterfly. More complex scenarii have been implemented where several persons in front of several cameras are simultaneously extracted and put into the same virtual background. A complete interactive system with two cameras has been implemented and tested during public demonstrations. For example school children enjoyed the system during an exhibition in Arc et Senans, France.

In order to implement such scenarii, computer vision tools are necessary: person segmentation for the incrustation, face and hands detection and tracking for triggering specific actions (for example butterfly capture or person transformation into a butterfly) in the scenario. In this paper we are focusing on face and hands detection and tracking.

As we work on interactive scenarii, the processing rate must be as close as possible to the video acquisition rate (at least 12 images per second). As a consequence processing complexity is a crucial criterion in algorithmic choices. Other systems with a high processing rate (30+ images per second) have been developed [1, 2] but they operate on small images and only faces are detected.

[1] *art.live*: IST project 10942, ARchitecture and authoring Tools prototypes for Living Images and Video Experiments

Fig. 1. Example of a scenario with a real person trying to catch virtual butterflies. Image is copyright Casterman, F. Place and the *art.live* project.

A lot of work has already been done on face detection and tracking. Two main approaches have been tried: colour based detection [1, 3] and facial features extraction [4, 5]. For our project, we have the following requirements:

- we aim at detecting both face and hands, contrary to most papers which only deal with face(s)
- the same method has to be used for detection of both face and hands (for computational cost reasons)
- features would be very complex to define for hands

Therefore a method based on colour is better suited to our project. When the background has a colour similar to the skin, this kind of method is perhaps less robust that a method based on body modelling. However, results of section 5 show that the proposed method works on a wide range of backgrounds.

The first section describes the detection of skin pixels and the second section proposes a method in order to identify relevant skin patches (face and hands). The last section presents some results focusing on processing rate and the precision of face and hands positions.

2 SKIN DETECTION

Before the detection of skin pixels, each connected components coming from a background removal algorithm is delimited by a rectangular bounding box (see figure 6). This algorithm performs an image difference with a background reference image. Several detection algorithms have been tested and compared during the *art.live* project (see for instance [6, 7]). But the detection of people is not the scope of the paper.

The skin detection is based on colour information. We have tested and compared several colour spaces, but only the most appropriate for skin detection (YCbCr and HSI) are reported here.

2.1 Colour spaces selection for skin detection

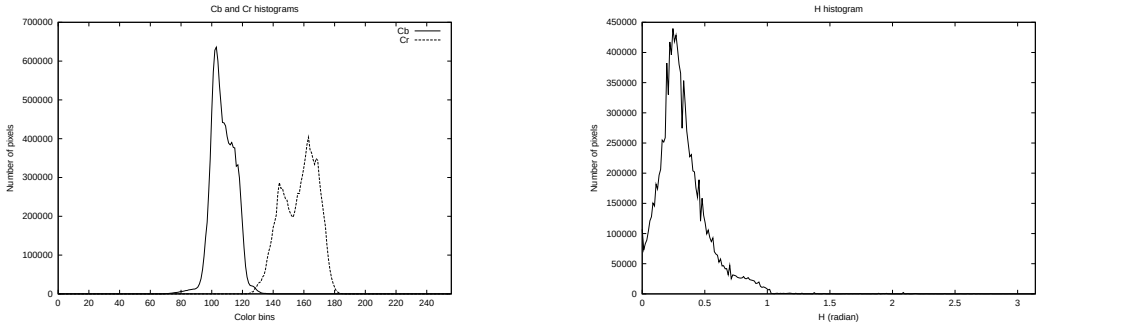
In order to compare the discriminative power of YCbCr and HSI colour spaces, a skin samples database has been built. It consists in the Von Luschan skin samples image (made of 36 skin samples of different colours ranging from pale white to deep black) and 20 skin samples acquired with the camera and frame grabber we use in the *art.live* project (in order to take into account the white balance and the noise of the acquisition system).

Numerous papers show that H or (Cb,Cr) are well discriminative for skin detection [3, 8]. Histograms of these components have been computed on the database (see figure 2). Figure 3 is a plot of all pixels from the database on the CbCr plan with an average value of Y. It exhibits two lobes: the left one corresponds to the Von Luschan skin samples and the right one to the other 20 skin samples.

Criteria used to select the most appropriate colour space are the following:

Concentration of skin samples: in the HSI space all pixels are grouped along the H axis (see figure 2b), in the YCbCr colour space pixels are grouped in a small region of the CbCr plan (see figure 2a).

Quality of detection: detection using the H component and detection in the CbCr plan give similar results (see section 5).



(a) (b)

Fig. 2 . Distributions of skin samples from the database (a) CbCr histograms (b) H histogram.

Computation speed: as images are acquired in the YCbCr colour space, it offers very good efficiency even though 4 thresholding operations are necessary instead of 2. On the contrary, the computation of H is very costly, even if the $\arccos$ function computation can be avoided.

Required memory: one idea to reduce the computational cost of H is to pre-compute the YCbCr $\rightarrow$ H transformation in a look-up table. However this would require 16 or 32 MBytes (depending on the needed precision) of memory which is too expensive for our application.

Finally, according to those 4 criteria, the selected colour space is YCbCr.

2.2 Fast skin detection method

As a result of previous analysis, the CbCr plan is divided into 2 connected and complementary areas: skin area and non-skin area. In order to test if a pixel belongs to the skin area, a thresholding is performed. To approximate the repartition of skin samples in the (Cb,Cr) plan, complex shape models such as triangles or ellipses have been considered. But a simpler rectangular model offers similar detection performances with a lower computational cost. It limits necessary computations to a double thresholding for each Cb and Cr variable. A rectangle containing most of our skin samples is defined by $Cb \in [90; 130]$ and $Cr \in [130; 180]$ (big rectangle of figure 3). The rectangle is centred on the lobe corresponding to our images to adjust the detection to our acquisition system. The right lobe is not completely included in the rectangle in order to avoid too much false detection. In [9] considered thresholds are slightly different ($Cb \in [77; 127]$ and $Cr \in [133; 173]$) which justify the tuning of parameters to the acquisition system and conditions. As an example, the small rectangle of figure 3 (defined by $Cb \in [90; 112]$ and $Cr \in [142; 170]$) only contains skin samples from a particular person in a particular image sequence. Therefore it will be necessary to adapt thresholds to each person and lighting conditions (see section 4).

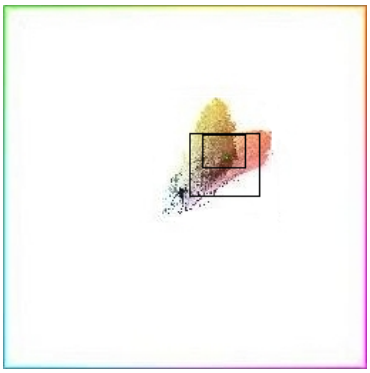


Fig. 3. Distribution of skin samples from the database in the CbCr colour plane.

In order to further reduce the computational cost, the skin/non-skin decision is performed for 2×2 pixels blocks. The considered observations for the decision are the average values of (Cb,Cr) colours components computed on a 4×4 pixels block centred on the 2×2 block. Those values are then compared with the 4 thresholds defined in the previous paragraph.

Since Cb and Cr components are already sub-sampled by a factor of 2, each 2×2 block is aligned on odd pixel coordinates so that the average computed on a 4×4 block is in fact the average of a 2×2 (Cb,Cr) pixels block.

As a bounding box around each person is computed, skin detection is restricted to this box.

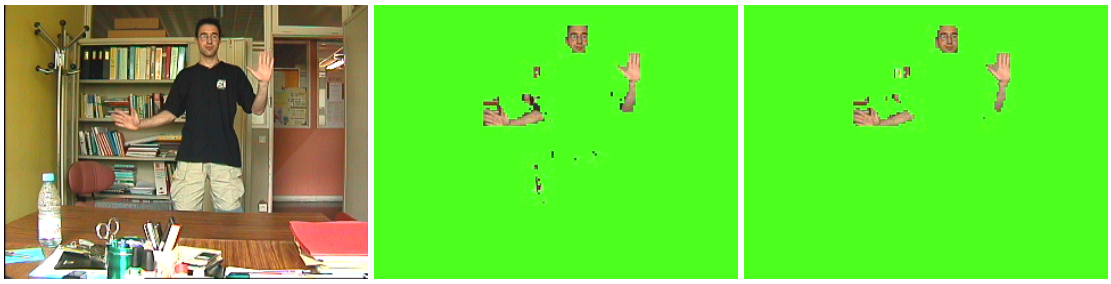


Fig. 4 . From left to right: original image, skin detection in the H and CbCr colour spaces.

3 FACE AND HANDS IDENTIFICATION

Pixels detected as skin are first labelled into connected components that are either skin patches or detection noise. Among all detected connected components, for each bounding box, skin patches (if present) have to be extracted from noise and identified as face or hands. To reach this goal several criteria are used. Detected components inside a given bounding box are sorted in decreasing order in lists according to each criterion.

Size and position criteria are:

- List of biggest components (Lb) : face is generally the biggest skin patch followed by hands and other smaller patches are generally detection noise
- List of leftmost components (Ll) : useful to detect the left hand
- List of rightmost components (Lr) : useful to detect the right hand
- List of high-most components (Lh) : useful to detect the face

Temporal tracking criteria are:

- List of closest components to the previous face position (Lcf)
- List of closest components to the previous left hand position (Lcl)
- List of closest components to the previous right hand position (Lcr)

Selection of the face involves (Lb,Lh,Lcf), selection of the left hand involves (Lb,Ll,Lcl) and selection of the right hand involves (Lb,Lr,Lcr). The first top elements of each list are considered as likely candidates. When the same element is not at the top of all lists, the next components in the lists are considered and one list has more influence than the others. This selection is guided by heuristics related to human morphology.

Several heuristics are used for the face identification: for instance, the face is supposed to be the biggest, the high-most skin patch and the closest to the previous face position. In many cases there is a connected component that is at the top of those 3 lists. But other cases are possible for instance when:

- the face appears smaller than one hand in the image (for instance bearded persons or hand closer to the camera)
- one hand is above the head
- one or both hands come into contact with the head in the image

In those cases, Lcf (tracking information) is the dominant list because head motion is slow and steady. The maximal rank considered in other lists is limited in order to avoid unlikely situations.

4 THRESHOLDS ADAPTATION

Adaptation of the skin detection thresholds is useful to take into account colour changes coming from inter-individual skin color variations, light sources and image acquisition systems. Several papers use colour models, for instance Gaussian p.d.f in the H or (r,g) color space [1] and perform thresholds adaptation by adjusting model parameters.

Cb and Cr histograms on figure 2 do not seem like Gaussian distributions. This has been confirmed by quantitative tests. An evaluation of Gaussianity of Cb and Cr distributions has been performed on

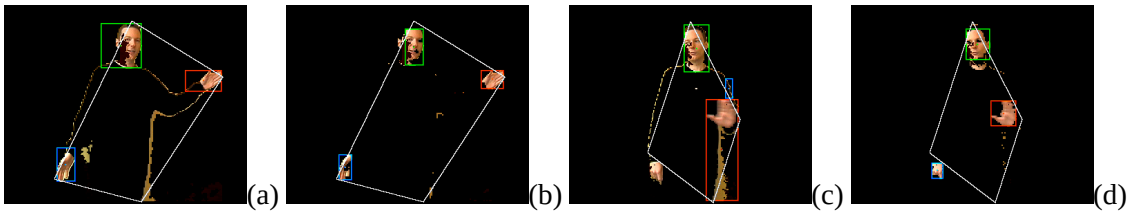


Fig. 5. Skin detection results with face and hands identification. (a) and (c): without thresholds adaptation. (b) and (d): with thresholds adaptation.

skin pixels of several long image sequences. For each image a Kolmogorov-Smirnov test has been performed. As a result, about half of distributions cannot be reliably represented by a Gaussian distribution. Therefore we chose to directly adapt thresholds without considering any model.

Skin detection thresholds are initialised with (Cb,Cr) values defined by the "big rectangle" of section 2.2. In order to adapt this initial rectangle, 3 transformations are considered (they apply separately to both dimensions Cb and Cr) :

- The rectangle is translated towards the observed mean of skin pixels belonging to the identified face skin patch. Using the previous face identification result prevents the adaptation to be biased by detected noise. In order to avoid too sharp transitions a gradual adaptation is performed: the translation is limited to only one colour unit per frame. The translated rectangle is constrained to remain inside the initial big rectangle.
- The rectangle is gradually reduced (by one colour unit per frame). Either the low threshold is incremented or the high threshold is decremented so that the reduced rectangle is closer the observed mean of skin pixels belonging to the face skin patch. Reduction is not performed if the rectangle reaches a minimal size (15 colour units).
- The rectangle is reinitialised to the initial values if the adapted thresholds lead to no skin patch detection.

Those transformations are applied once for each image of the sequence. As a result detection should improve over time. The adaptation needs about 30 images to reach a stable state.

5 RESULTS

Figure 4 shows a comparison between detection in the H and CbCr colour spaces. The detection interval for H is $[0.3; 0.4]$ and for CbCr, the small rectangle is used (as explained in section 2.2). Results are similar: in both cases face and hands are correctly detected, the CbCr colour space leading to slightly more accurate boundaries (see the arm on the right). Moreover the H result shows more detection noise.

Figure 5 shows that thresholds adaptation improves skin detection quality. Skin detection in this sequence is difficult because the background is yellow (close to skin colour). Threshold adaptation reduces false detections which can lead to bad face and hands identification. For example results in figure 5(d) are better than results in figure 5(c): without adaptation the left hand bounding box is too large and the right hand is not correctly identified.

Figure 6 shows results of face and hands identification and tracking for 2 sequences (4 non-consecutive frames have been extracted from the 500 frames of the sequence). Skin detection is performed inside the bounding box of the person (shown in white). Even if the background removal is not perfect, face and hands are correctly identified and tracked as shown by the colour boxes.

The processing rate of the whole system (including background removal) is 24 CIF images (352×288 pixels) per second on a low-end PC (700MHz processor). Skin detection, face and hands identification and tracking represents only 20% of the total computational cost. Furthermore this part is implemented in unoptimised C++ whereas the rest of the system has been optimised with MMX assembly.

6 CONCLUSION

The proposed method yields results with a precision suitable for our application that consists in touching virtual objects. It is fast enough for a responsive system that offers pleasing human interaction. Both results have been validated during the public demonstrations of the project thanks to a questionnaire filled by people trying the system.

In future applications, it may be interesting to track several persons forming a group instead of having only one person in front of each camera. This would allow scenarii where people could interact in the real world in addition to interaction in the virtual world (through the system and a projection screen). In this context, reliable face detection and tracking could be used in order to distinguish several persons.

REFERENCES

[1] Jie Wang, Weier Lu, and Alex Waibel, "Skin-color modeling and adaptation," in *Proc. ACCV*, Hong Kong, 1998, vol. 2, pp. 687–694, http://www.is.cs.cmu.edu/js/.

[2] Gary R. Bradski, "Computer vison face tracking for use in a perceptual user interface," *Intel Technology Journal*, 1998.

[3] J.C. Terrillon, M. N. Shirazi, H. Fukamachi, and S. Akamatsu, "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images," in *Proc. IEEE Int. Conf. on Face and Gesture Recognition*, Grenoble, France, Mar. 2000, pp. 54–61.

[4] R. Brunelli and T. Poggio, "Face recognition: features versus templates," *IEEE-T-PAMI*, vol. 15, no. 10, pp. 1042–1052, Oct. 1993.

[5] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspace for face recognition," in *IEEE-C-CVPR*, Seattle, WA, USA, 1994, pp. 84–91.

[6] A. Caplier, L. Bonnaud, and J.-M. Chassery, "Robust fast extraction of video objects combining frame differences and adaptive reference image," in *Proc. IEEE Int. Conf. Image Processing*, Thessaloniki, Greece, Sept. 2001.

[7] A. Cavallaro and T. Ebrahimi, "Video objects extraction based on adaptative background and statistical change detection.," in *SPIE Electronic Imaging*, San Jose, California, USA, Jan. 2001.

[8] K. Sobottka and I. Pitas, "A novel method for automatic face segmentation, facial features extraction and tracking," *Signal Process. : Image Commun.*, vol. 12, no. 3, pp. 263–281, June 1998.

[9] D. Chai and K.N. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, June 1999.

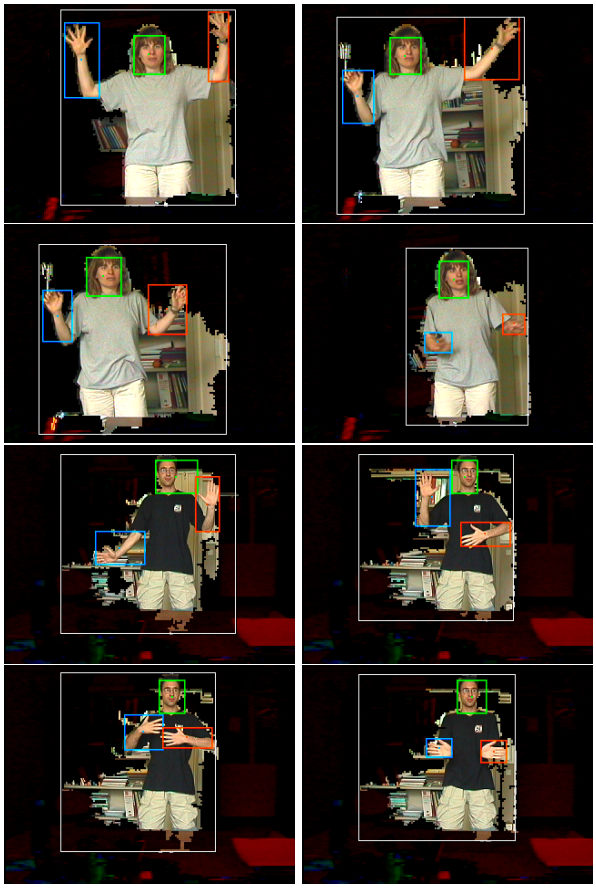


Fig. 6. Face and hands identification and tracking. Segmented person superimposed on a dark background. Big white rectangle: bounding box of the person. 3 small rectangles: face (green), right hand (blue), left hand (red).