A NEW SKIN COLOUR ESTIMATION METHOD BASED ON CHANGE DETECTION AND CLUSTER ANALYSIS

Stefano Masneri, Oliver Schreer

Fraunhofer Heinrich-Hertz-Institute, Berlin, Germany
Email: [stefano.masneri,oliver.schreer]@hhi.fraunhofer.de

ABSTRACT

Skin colour is a useful and robust cue for face and hand detection and tracking and it has been widely exploited in many different applications such as human motion analysis or image and video retrieval. The presented work is a result of the joint Fraunhofer/Max-Planck research project AVATecH (Advancing Video/Audio Technology in Humanities Research), which deals with automatic annotation of large video corpora in humanities research for psycholinguistics. One of the tasks in this project is the development of a robust skin colour estimation algorithm necessary for the realization of tools for hands and head tracking. Because of the peculiarities of the large video corpora, there are no skin colour models applicable to detect the correct skin colour in each video. Therefore an algorithm was developed that estimates skin colour independently for each video and more important, it works reliably without the need of a training set. The algorithm uses both a change detection tool, to select the most suitable frames for skin colour estimation, and an iterative clustering algorithm to select the range in the YUV domain that best represents skin colour.

1. INTRODUCTION

Detecting human skin colour is a task of utmost importance in many different applications, such as face and gesture recognition, human-computer interaction, surveillance, image indexing and retrieval. In the last few years, the problem of human skin colour detection has received considerable attention and many different approaches were used to model skin colours. An incomprehensive list includes: the definition of explicit cluster boundaries in the colour space domain [1], histogram-based normalized lookup tables [2], Bayes classifiers [3], self-organizing maps [4], single Gaussians [5], mixture of Gaussians [6], elliptic boundary model [7]. Other approaches also use face detection algorithms [8], [9]. A survey outlining and comparing these methods can be found in [10]. Furthermore, different algorithms use different colour spaces to perform the estimation. In [11], different colour spaces are analyzed and the conclusion is drawn that the separation in skin and no skin classes is independent of the selected colour space.

What all these approach have in common, though, is that they work in two phases: the training phase, which allows the algorithm to learn the parameters of the skin classifier based on a database of skin patches from different images, and a detection phase, which uses the results of the previous step to classify each pixel in the image.

The approach described here is different: first of all the algorithm presented in this paper estimates skin colour parameters for video data, while other algorithms work on image datasets. Usually these algorithms estimate skin parameters only once and use them to detect skin pixels in the whole database, while the algorithm proposed estimates different parameters for each video. Secondly, due to the huge size of the database (more than 50 Terabytes) the estimation algorithm must work in a fully automatic way, without any kind of user interaction in the process, approaches that require input from the user are not applicable in this case. Other techniques relying on a face detector to identify skin pixel regions can't be applied either, because in many videos in the database there's no clear view of the person's face (see Figure 4). Finally, since there is no ground truth data available for video datasets, no training phase is feasible and therefore the algorithm has no *a-priori* information that can be used to improve the estimation results. The algorithm described in this paper estimates for each shot of the input video its skin colour parameters, optimized for the special characteristics of that shot (lighting, camera settings, recording quality). The parameter estimation is performed in two steps. At first the algorithms picks a small set of video frames (in practice just one frame is needed) which are suitable for the analysis performed in the subsequent step, where an iterative process is used for the skin colour parameter estimation. The paper is organised as follows: after a short description of the constraints given by the dataset under investigation, the previously mentioned two steps of the algorithm are described in detail in section 3 and 4. In section 5, experimental results are presented. The paper ends with a conclusion.

2. AVATECH DATASET

One of the major requirements of the skin detection tool described in this paper was the ability to work with the whole AVATecH dataset. This posed a few challenges, since the dataset is extremely heterogeneous. Even if all the videos are related to gesture or sign language analysis, the content is very variegated: there are videos with one or more people, videos with or without monochromatic background, people can be filmed in close-up, in a waist shot or in full-figure, the lighting conditions are hardly ever uniform. Furthermore, different videos have different resolutions, frame rates, coding artefacts due to different bitrates.

The video content is so heterogeneous, in fact, that skin colour parameters retrieved empirically for some videos are completely wrong for other videos and there isn't a unique set of parameters which works well enough for most of the dataset. For this reason the algorithm presented in this paper estimates appropriate skin colour parameters for each video.

The algorithm is using the YUV colour space, because it allows the separation between luminance and colour components. In most of the cases the colour information is sufficient for the detection of skin in an image, the luminance component being important only in particular conditions. Furthermore, experimental studies [12] show that, regardless of the ethnicity, the skin of different people has similar colour components (at least in the UV domain) and the differences are mainly caused by the luminance component.

3. CHANGE DETECTION

The first step of the algorithm is about the selection of frames suitable for the estimation of skin colour parameters. Since the skin colour estimation is performed studying clusters inside of the image and analyzing their compactness and circularity, the best frames are the ones where there is no overlay between the arms and the head. The idea is to use a change detection algorithm, applied to the luminance component of the image. For each frame of the video a binary image I_C representing change is computed. Given the intensity components F_n and F_{n-1} of two consecutive frames, I_C is defined as

$$I_C(x,y) = \begin{cases} 1 \ if \ |F_n(x,y) - F_{n-1}(x,y)| \geq \alpha \\ 0 \ otherwise \end{cases}$$

Where α is a threshold determined experimentally. I_C is then split in blocks of size NxN (the value of N depending on the image resolution) and for each block the number of ones in I_C is counted. In this way a new image H_C is defined as

$$H_C(x,y) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} I_C(Nx+i; Ny+j)$$

Following this definition, H_C is N^2 times smaller than I_C and each pixel of H_C represents the sum of the values of a portion of the image I_C. High values of H_C are related to significant amounts of change in the corresponding region of the frame. Figure 1 shows an example frame with the change image overlaid (top), and its corresponding H_C image (bottom).

H_C is then processed in order to remove outliers and to cluster the data. The clustering procedure simply groups together all the connected pixels of H_C with value greater than zero. Ideally, each cluster represents a body part moving in the current frame.

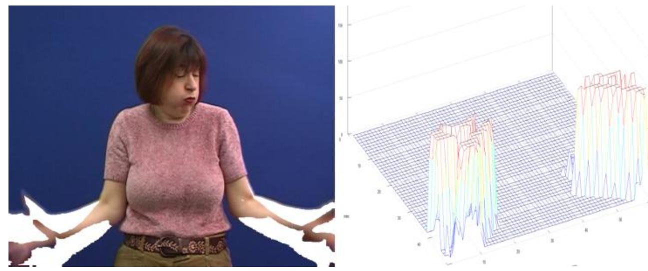

Figure 1: Change detection in a video frame and corresponding H_C image.

Information regarding size, position, compactness is recorded for each cluster found in H_C. After that all the information is passed to a cost function, which assigns a score to the current frame based on the properties of the clusters. The clusters' properties taken into account are:

- The size of the biggest cluster, with respect to the size of the frame;
- The ratio between the size of the biggest and the smallest cluster (if there is no overlapping between arms and head the clusters' dimensions are similar);
- The average and minimum distances between clusters (overlaps are less likely to occur if the distance between clusters is high);
- Position of the clusters (usually no arm/head overlap occurs in the lowest part of the frame).

Furthermore, frames are discarded (i.e. they are assigned a score zero) if the number of cluster found is too high (because the videos in the AVATecH database never involve more than three people at the same time).

In order to work with the whole dataset the cost function has to satisfy certain conditions, such as providing meaningful results independently from the video resolution, the number of people in the scene and their position. As depicted in the given example in Figure 1, the algorithm also has to cope with motion blur due to fast moving parts in the video. An empirical analysis on a large variety of testing videos proved that suitable frames for further processing can be obtained, i.e. the body parts are not overlaying and there is no excessive blurring in the frame due to fast moving body parts. The three frames getting the highest score are then selected to perform the skin colour estimation.

4. ITERATIVE SKIN COLOUR ESTIMATION

Once the most suitable frames in the video have been selected, the skin colour is estimated applying an iterative approach. To successfully estimate the skin colour the algorithm retrieves six parameters, the mean value and the size of the range for the luminance component Y and the colour components U and V. These parameters define a subset (a parallelepiped) in the YUV colour space, and all the pixels in the image inside this subset are then marked as skin. To estimate the mean values and ranges for the three components an iterative approach has been developed. At each iteration two parameters are estimated while keeping the other four fixed and the results from one iteration are used in the following ones. The skin colour estimation is performed in 4 steps:

1. Coarse estimation of U and V mean values
2. Estimation of U and V ranges
3. Fine estimation of U and V mean values
4. Estimation of Y mean and range

During each iteration the input frame is segmented into different binary images B_i. Given the Y, U and V components of the input frame, the mean values $\bar{Y}, \bar{U}, \bar{V}$ and the ranges $\Delta Y, \Delta U, \Delta V$, the segmented images are defined using the following rule (in the case of the first iteration of the algorithm):

$$B_i = \begin{cases} 1 \; if \; Y(x,y) \epsilon [\bar{Y} - \Delta Y, \bar{Y} + \Delta Y] \\ \quad \& \; U(x,y) \epsilon [\bar{U_l} - \Delta U, \bar{U_l} + \Delta U] \\ \quad \& \; V(x,y) \epsilon [\bar{V_l} - \Delta V, \bar{V_l} + \Delta V] \\ 0 \; otherwise \end{cases}$$

Using the first iteration step as an example, 60 binary images B_i are created, each one with different U and V mean values. Figure 2 shows (light blue, overlaid on the input frame) the results of the segmentation using four different $\bar{U}$ and $\bar{V}$ values, while Figure 3 shows the block diagram of the different steps of the iterative algorithm.

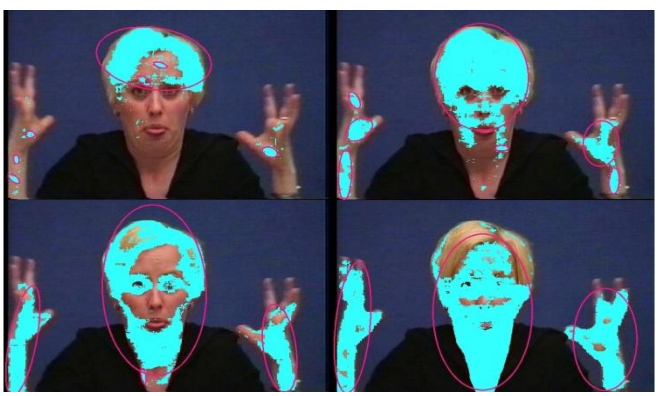

Figure 2: Segmentation using different UV parameters

On every binary image, similarly to what was done during the change detection step, a cluster analysis is performed to decide the values of $\bar{U}$ and $\bar{V}$ that are most likely to represent human skin colour. The decision of the best values is based on the following properties:

- the compactness of the cluster (defined as the ratio between the number of segmented pixels and the area of the ellipse that best approximates the shape of the segmented skin region);
- the size of the cluster relative to the size of the frame;
- the number of clusters in the image and their position;
- the distances between the clusters in the images B_i and H_C.

Based on the results of the cluster analysis the best $\bar{U}$ and $\bar{V}$ values are selected. The second iteration performs then the same operations, but keeping fixed the values of $\bar{U}$ and $\bar{V}$ (the values estimated at step 1 are used) while changing the ranges ΔU and ΔV. The third iteration is performed to increase the accuracy of the estimation of $\bar{U}$ and $\bar{V}$: this step is essentially the same as the first one but it is performed using ΔU and ΔV values found during iteration 2 and using $\bar{U}$ and $\bar{V}$ values in the neighbourhood of the values estimated at iteration 1. Finally, $\bar{Y}$ and ΔY are estimated using the same approach.

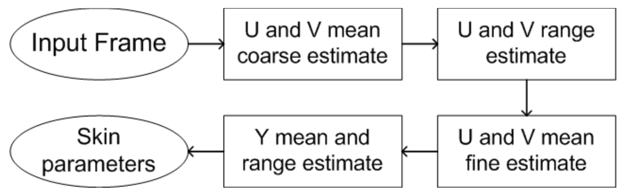

Figure 3: Block diagram of the iterative skin colour estimation

To increase the robustness of the algorithm the estimation of the skin color parameters is performed on the three frames selected during the change detection step. The skin parameters obtained for these frames are then used to perform skin detection on the whole video shot.

5. EXPERIMENTAL RESULTS

Many skin detection algorithms use existing database of skin labelled images, such as the Compaq dataset [16] to test their performance. Unfortunately there is no such ground truth data for videos, so it is not possible to compute precision and recall, which are standard measures in classification. Moreover, the aim of the algorithm is to perform a successful and stable tracking of hands and heads on the video dataset therefore some errors are tolerated if they do not affect the overall performance. Hence, the ground truth has been estimated on a large number of testing videos from the AVATecH dataset [19] by selecting manually the best parameters (the mean values $\bar{Y}, \bar{U}, \bar{V}$ and the ranges $\Delta Y, \Delta U, \Delta V$) for successful segmentation and tracking of hands and head. These parameters have then been compared to the fully automatic approach described in this paper. The test videos represent fairly well the AVATecH dataset in its entirety, since different videos have different quality, resolution, bit rate and all the possible use cases are represented. The parameters estimated by using this algorithm allow very accurate skin detection in most of the test videos and the correct identification of hands, arms and heads. In some cases the algorithm did not perform accurate skin colour estimation due to the fact that the video quality is extremely bad and also the colour of the clothes is similar to skin colour. Table 1 shows the resulting UV interval obtained by the algorithm in 9 test videos and compares them with the optimal intervals retrieved manually, while Figure 4 shows frames from 4 different videos with the results of skin detection using the algorithm.

	Ground truth U	Ground truth V	Retrieved U	Retrieved V
SSL_JI_int_b	[76, 104]	[160, 186]	[75, 105]	[160, 188]
NGT_AH_fab3_f	[104, 124]	[144, 174]	[105, 125]	[145, 175]
NGT_AH_fab3_b	[104, 124]	[144, 174]	[107, 127]	[141, 176]
BSL_PS_fab3_b	[100, 120]	[136, 170]	[104, 128]	[135, 170]
NGT_JR_fab3_b	[110, 130]	[130, 160]	[115, 135]	[125, 155]
NGT_JR_fab3_f	[110, 130]	[130, 160]	[110, 135]	[130, 160]
BSL_CN_int_b	[60, 88]	[146, 178]	[75, 125]	[146, 181]
SSL_JI_fab3_b	[80, 120]	[141, 175]	[81, 121]	[135, 170]
NGT_BB_fab3_b	[65, 115]	[150, 174]	[75, 110]	[150, 180]

Table 1: Skin detection results on 9 test videos.

6. CONCLUSION

This paper presented a new algorithm for fully automatic skin colour estimation in videos. The main difference to previous approaches is that no training datasets are required. It works completely automatic without any hard constraints on the video material and detects different skin colour parameters for each video. The algorithm is somehow constrained with respect to the number of persons present in the scene and assumes that the illumination conditions in each shot remain constant. In the database under investigations the approach works successfully for scenarios up to three persons. The algorithm has been validated compared to manual ground truth estimation of the skin colour parameters. Future developments aim to increase the robustness of the algorithm and its speed.

Figure 4: Skin detection results.

7. ACKNOWLEDGEMENT

We would like to thank the collaborating Max-Planck Institute in Nijmegen, Netherlands for the provision of the database used in the AVATecH project.

8. REFERENCES

[1] P. Peer, J. Kovac, and F. Solina, “Human skin colour clustering for face detection”. In *International Conference on Computer as a Tool, EUROCON*, 2003.

[2] Q. Chen, H. Wu and M. Yachida, “Face detection by fuzzy pattern matching”, in *Proceedings of the Fifth International Conference on Computer Vision*, pp. 591–597.

[3] M.J. Jones and J.M. Rehg, “Statistical color models with application to skin detection”, in *Proceedings of the CVPR*, 1999, vol. I, pp. 274–280.

[4] D. Brown, I. Craw and J. Lewthwaite, “A som based approach to skin detection with application in realtime systems”, in *Proceedings of the British Machine Vision Conference*, 2001.

[5] R.L. Hsu, M. Abdel-Mottaleb and A.K. Jain, “Face detection in color images”, in *IEEE Trans. Pattern Analysis and Machine Intelligence 24*, vol. V, pp. 696–706.

[6] M.H. Yang and N. Ahuja, “Detecting human faces in color images”, in *International Conference on Image Processing (ICIP)*, 1998, vol. I, pp. 127–130.

[7] J.Y. Lee and S.I. Yoo, “An elliptical boundary model for skin color detection”, in *Proceedings of the 2002 International Conference on Imaging Science, Systems, and Technology*, 2002.

[8] J. Fritsch et al., “Improving adaptive skin color segmentation by incorporating results from face detection”, *Proceedings of Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, 2002, pages 337–343.

[9] J.F. Lichtenauer, M.J.T. Reinders & E.A. Hendriks, "A Self-calibrating chrominance model applied to skin color detection", *Proceedings of the Second International Conference on Computer Vision Theory and Applications*, 2007, Vol. I, pp. 115-12

[10] V. Vezhnevets, V. Sazonov and A. Andreeva, “A Survey on Pixel-Based Skin Color Detection Techniques”, in *Proceedings of the GraphiCon 2003*, 2003, pp. 85-92.

[11] M. Jones and J. Rehg, Compaq Skin Database, *http://www.crl.research.digital.com/publications/techreports/abstracts/98-11.htm*

[12] Abbas Cheddad, Joan Condell, Kevin Curran and Paul Mc Kevitt, “A Skin Tone Detection Algorithm for an Adaptive Approach to Steganography”, *Journal of Signal Processing*, 2009, vol. 89, Issue 12.

[13] AVATecH web site: *http://www.mpi.nl/avatech*