Extraction of Hand Gestures with Adaptive Skin Color Models and its Applications to Meeting Analysis

Yingen Xiong, Bing Fang, and Francis Quek
Center for Human Computer Interaction
Virginia Polytechnic Institute and State University
660 McBryde Hall, MC 0106, Blacksburg, VA 24061
yxiong@cs.vt.edu

Abstract

We present an adaptive skin color model for hand gesture tracking which is applied to cross-modal analysis of planning meetings. We build a skin color model with the Gaussian distribution and a skin color filter for each participant in meetings. By combining with the Vector Coherence Mapping (VCM) algorithm, we track hand motion and obtain 3D trajectories. The hand gesture stream is extracted from hand motion trajectories. Different skin color models are created for different people to handle the differences of skin color. We update each model dynamically to adapt changes of environments. A parallel system has been implemented to track and extract hand motion trajectories. Examples of hand motion gesture tracking in meeting environments are provided. The applications of the adaptive skin color model can increase the speed of hand tracking.

1 Introduction

Meetings are gatherings of humans for the purpose of communication. In video-based multimodal analysis of planning meetings, one of the important objectives is detection, recognition, and understanding of video events associated with the meetings. The understanding of human multimodal communicative behavior, and how witting and unwitting visual displays relate to such communication is a key to any approach to reach this objective. To attack this challenge problem must necessarily combine the psycholinguistics of multimodal human language, signal and language processing, and computer vision.

In video-based meeting analysis, meeting events are recorded by cameras [4]. The extraction of meaningful meeting events is very important. Especially, we are interested in the events related to speakers. This paper addresses the aspect of hand gesture tracking for all participants.

Human hand gesture in standard video data poses several processing challenges. First, one cannot assume contiguity of motion in the video. Dense optical flow methods cannot be used. Second, because of the speed of motion, there is considerable motion blur. Third, the hands tend to merge, separate, and occlude each other. Fourth, hand shapes are highly deformable. Finally, temporal resolution is extremely critical since we need to correlate the initiation and cessation of motion phases with speech units (e.g. a syllable). In this paper, we extend the original VCM algorithm [1] to perform hand motion tracking and extract hand gestural trajectories. We create skin color models and filters to build color constraints for VCM, so that we do not need to track motion in the whole image frame, which will increase processing speed. We update the skin color models and filters dynamically on the fly to fit the changes of environments. A parallel procedure is created to implement the process of hand gesture tracking. We demonstrate the efficiency of the algorithm with examples and applications in cross-modal analysis of planning meetings.

2 Summary of Our Approach

We create adaptive skin color models combining with the original VCM algorithm for hand tracking. For video sequences, we find interest points in skin color regions segmented with skin color filters. We compute optical flow for these interest points between frames with constraints of boundary conditions, spatial coherence, and temporal coherence. The optical flow of all interest points comprises motion vector fields. By clustering the motion vectors, we obtain 2D hand motion trajectories.

The main work of this paper is to build adaptive skin color models and filters to segment skin color regions to obtain correct interest points. The algorithm can be divided into three parts. The first part is to build skin color models. We manually collect samples from the initial image frame. After removing non-skin color points by a clustering approach, we apply the Maximum Likelihood Estimation (MLE) approach to estimate skin color distribution parameters. With these parameters we build initial skin color models with the Gaussian distribution and create a sample database. In the meantime, these samples are stored in the sample database. We create initial skin color filters with the

Proceedings of the Eighth IEEE International Symposium on Multimedia (ISM'06)
0-7695-2746-9/06 $20.00 © 2006 IEEE

initial color models. Skin color and non-skin color regions are segmented with the skin color filters. In the meantime, we collect sample points in skin color regions. After removing non-skin color points, we update the sample database. The skin color models and filters are updated with the updated sample database. In this way, while each frame image in image sequences is processed, the sample database, skin color models and filters are updated. The sample database stores samples from previous three frames to keep the continuity of the change of model parameters. The skin color models are updated dynamically on the fly to fit changes of skin color caused by the changes of environments such as background and illumination.

3 VCM Algorithm

VCM is a parallel algorithm for hand motion tracking. This approach incorporates the various local smoothness, spatial and temporal coherence constraints transparently by the application of fuzzy image processing techniques to compute an optical flow field (vector field) from a video image sequence. A weighted voting process in local vector space is applied to obtain the vector field. By clustering the vector field, the hand motion trajectories can be obtained.

The VCM algorithm tracks motion in the whole image frame. Wherever motions exist, it will compute motion vectors. However, for the application of hand motion tracking, we only need to compute motion vectors on hand and apply these vectors to estimate hand motion trajectories.

4 Adaptive Skin Color Model and Filter

In this section, we build a skin color constraint for VCM, so that it does not need to track the motion on the whole frame. Instead, only skin color points can be chosen as interesting points for motion tracking. With skin color constraint, the speed of VCM can be increased. Two main problems need to be solved in this section: build and train a skin color model; update the model dynamically.

4.1 Skin Color Model

The skin color model theory is established by Yang and other researchers [2, 5]. A survey on skin color detection techniques is available in [3]. According to skin color theory, under certain lighting conditions, a skin color distribution can be characterized by a multivariate Gaussian distribution in the normalized color space. We can model human face and hand with different color appearances. By computing the probability of a pixel in skin color Gaussian distribution we can segment the skin color and non-skin color regions.

We build the skin color model in RGB space. Usually the RGB space in original color image includes luminance component, which makes it difficult to characterize skin color because lighting effects change the appearance of the skin. In order to reduce lighting effects, we convert original color images to chromatic color images. Suppose $\mathrm{x}(R, G, B)$ and $\mathrm{x}'(R_n, G_n, B_n)$ are pixels in the original color image and chromatic color image respectively,

$$R_n = \frac{R}{R+G+B}, B_n = \frac{B}{R+G+B}, G_n = \frac{G}{R+G+B}. \qquad (1)$$

In above, as $R_n + B_n + G_n = 1$, there are only two independent components, so we omit the third component. For each pixel, we have a color vector $\mathrm{x} = (R_n\ B_n)^T$. The two dimensional Gaussian distribution model is expressed as $N(\mu, \Sigma)$ i.e.

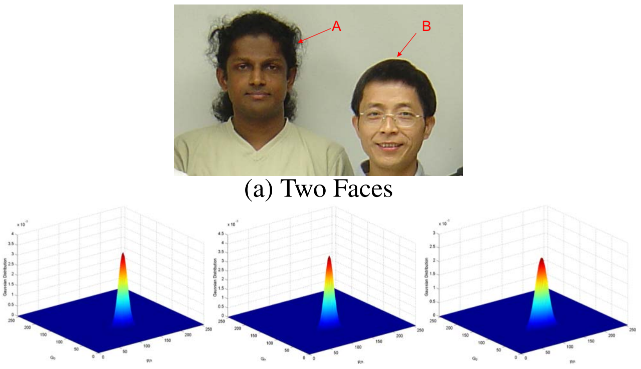
$$p(\mathrm{x}) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp[-\frac{1}{2}(\mathrm{x}-\mu)^T \Sigma^{-1} (\mathrm{x}-\mu)] \qquad (2)$$

$$\text{with} \quad \begin{cases} \mu = E\{\mathrm{x}\} \\ \Sigma = E\{(\mathrm{x}-\mu)(\mathrm{x}-\mu)^T\} \end{cases} \qquad (3)$$

where, μ is the mean vector and Σ is the covariance matrix.

Before we can use this model, we need to create samples to estimate the parameters μ and Σ. One of the classical parameter estimation approaches is the Maximum Likelihood Estimation (MLE).

$$\begin{cases} \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{x}_k \\ \hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} (\mathrm{x}_k - \hat{\mu})(\mathrm{x}_k - \hat{\mu})^T \end{cases} \qquad (4)$$



(a) Two Faces

(b) Model A (c) Model B (d) Model C

Figure 1. Skin Color Gaussian Model

In meeting analysis, we have more than one people in the meeting and we need to track their hand motion. We build a skin color model for each subject so that different models can handle different skin color. With skin color samples, we apply the MLE approach to estimate parameters. Skin color Gaussian models can be built with these parameters. Figure 1 shows two faces and three skin color Gaussian models. Models A, B, and C are built with samples from face A, face B, face A and face B respectively. The parameters for Model A are:

$$\hat{\mu}_A = \begin{pmatrix} 123.64 \\ 84.09 \end{pmatrix}, \quad \hat{\Sigma}_A = \begin{pmatrix} 79.48 & -18.12 \\ -18.12 & 27.71 \end{pmatrix}.$$

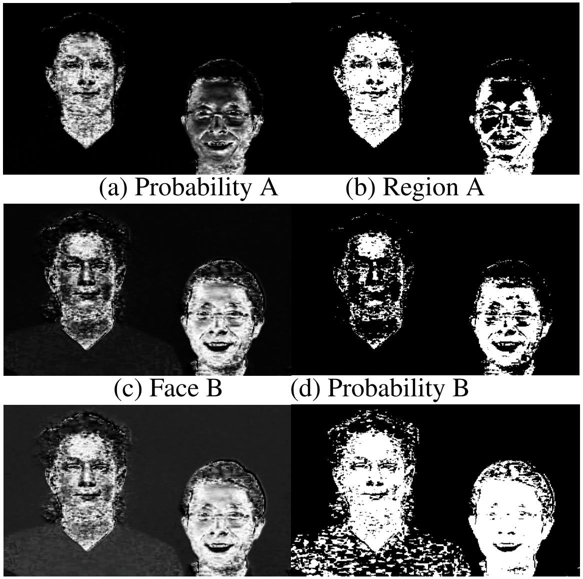
for Model B:

$$\hat{\mu}_B = \begin{pmatrix} 113.26 \\ 85.95 \end{pmatrix}, \quad \hat{\Sigma}_B = \begin{pmatrix} 88.91 & -14.58 \\ -14.58 & 19.98 \end{pmatrix}.$$

and for Model C:

$$\hat{\mu}_C = \begin{pmatrix} 114.72 \\ 85.59 \end{pmatrix}, \quad \hat{\Sigma}_C = \begin{pmatrix} 173.71 & -26.90 \\ -26.90 & 26.06 \end{pmatrix}.$$

From these models we can see that people with different skin colors have different parameters of models. From eigenvalues of covariance matrices we can see the ranges of distributions. The eigenvalues for these three models are 21.10, 85.19 for Model A, 17.02, 91.87 for Model B, and 21.31, 178.46 for Model C. The distribution range of the Model C is much larger than that of previous two models, which will cause some errors in skin color region segmentations. For this reason, we will build a skin color model for each participant in our meeting analysis.


(a) Probability A (b) Region A

(c) Face B (d) Probability B

(e) Face B (f) Probability B

Figure 2. Face, Skin Color Probability, Skin Color Region vs Non-Skin Color Region

4.2 Skin Color Filter

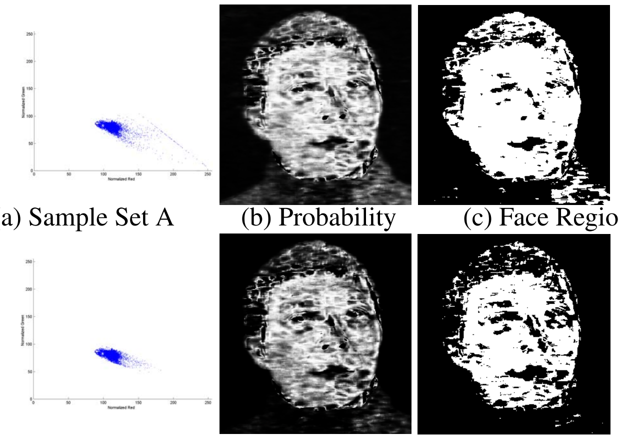
With a skin color model, we can create a skin color filter to segment the skin color and non-skin color regions. In order to build a skin color filter, we compute the probability of each pixel in an image being skin color. The probability of a pixel with color x is

$$P(\mathrm{x}) = \exp\left\{-\frac{1}{2}(\mathrm{x} - \hat{\mu})^T \hat{\Sigma}^{-1} (\mathrm{x} - \hat{\mu})\right\} \quad (5)$$

Given a probability threshold, we can create a skin color filter. As skin colors do vary between each individual subject, we shall find best threshold values for different subjects under different applications (background, illumination, etc.). We can use samples to train the system to learn thresholds for different participants in the meeting.

In implementation, we scale the skin color probability of every pixel in an image from [0, 1] to [0, 255], thus we can create a probability gray scale image. Furthermore, by thresholding, we create a black/white binary image to represent the non-skin color and skin color regions.

Figure 2 shows different skin color filters for different people who have different skin color. Figure 2 (a) and (b) shows the results by the skin color filter created with Model A shown in Figure 1 (b). From the figure we can see that the model built for person A does not fit person B well, because they have different skin color. In the similar way, Figure 2 (c) and (d) shows the results by the skin color filter created with Model B shown in Figure 1 (c). We also can see that the model built for person B does not fit person A well. Figure 2 (e) and (f) shows the results by the skin color filter created with Model B shown in Figure 1 (d). Because the model is built with samples combining person A with person B, the distribution range of the model is much larger. From the results we can see that the color region segmentations include not only skin color regions but also others such as hair and clothes areas. We decide to build a skin color model for each subject with his/her own skin color samples to solve this problem.



(a) Sample Set A (b) Probability (c) Face Region

(d) Sample Set B (e) Probability (f) Face Region

Figure 3. Sample Clustering

4.3 Skin Color Samples

Skin color samples are very important. We use these samples to build a skin color model for each subject. When we use mouse to select regions for sampling, we may include some non-skin color pixels in these samples, which will effect the accuracy of the skin color model. Figure 3 (a) shows the case which the samples contains a large part of skin color pixels (dense points) and a small part of non-skin color pixels (sparse points). The skin color Gaussian model built with these samples is

$$\hat{\mu}_A = \begin{pmatrix} 111.25 \\ 79.89 \end{pmatrix}, \quad \hat{\Sigma}_A = \begin{pmatrix} 80.38 & -29.19 \\ -29.19 & 26.78 \end{pmatrix}.$$

With this model, we compute the probability image shown in Figure 3 (b). From this figure we can see that more errors are made on areas of hair and clothes. After filter the probability image to a skin color region image shown in 3 (c) , we also can see the errors.

Before we apply these samples to build a skin color model, we apply some techniques to remove these non-skin color pixels from the samples. We know the fact that the skin color points of a person should fall in a cluster in skin color space. We apply clustering techniques to keep these skin color points in the samples and remove these non-skin color points. With these skin color ponits, we build a more accurate model. Figure 3 (d) shows the samples after clustering. We rebuild this skin color model as below,

$$\hat{\mu}_B = \begin{pmatrix} 110.68 \\ 80.17 \end{pmatrix}, \quad \hat{\Sigma}_B = \begin{pmatrix} 34.07 & -9.10 \\ -9.10 & 13.87 \end{pmatrix}.$$

By comparing current and previous models, we can see that both mean vector and covariance matrix are changed. The eigenvalues of the covariance matrix in previous and current models are 13.95, 93.21, and 10.38, 37.56. The parameters of the current model are smaller than that of the previous one, which means that the range of the current model is smaller than that of previous one. Figure 3 (e) shows the probability image computed with the current skin color model. We can see that the errors in areas of hair and clothes are reduced much (almost no error in the area of clothes).

Our conclusion for sampling is that apply clustering techniques to remove non-skin color points from samples and then build skin color models and filters with the clustered samples.

4.4 Update Procedure

Updating the skin color model is very important for our hand tracking system to deal with changes of environments. For example, if light conditions change, the skin color appearance will be affected. The skin color model should be updated, otherwise, the system will find wrong interesting points to track.

We present an adaptive skin color model to fit these changes. We update our model dynamically on the fly for each frame according to the color appearance changes.

We create an update procedure including these steps. First, we build an initial skin color model and filter by manually collected skin color samples, and create a sample database to store these samples. Second, with the initial model and filter, we segment next frame into skin color and non-skin color regions and obtain skin color points. Next, we update the sample database with these new skin color points. In the mean time, we update the skin color model and filter with the updated sample database. So for each frame, we apply the current model and filter to perform segmentation, use skin color points to update sample database, and use the updated sample database to update the skin color model and filter. In this way, we improve our model and filter dynamically to adapt the changes of skin color appearance. In order to keep the continuity of the update process, we let the sample database keep sample points of current and previous two frames. In the mean time, the skin color points segmented by the filter in each frame are used in the next step of hand tracking process.

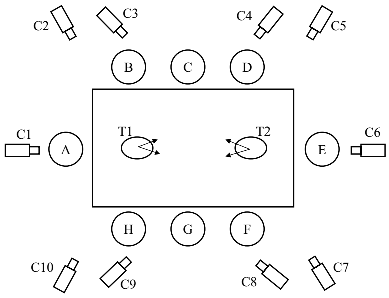
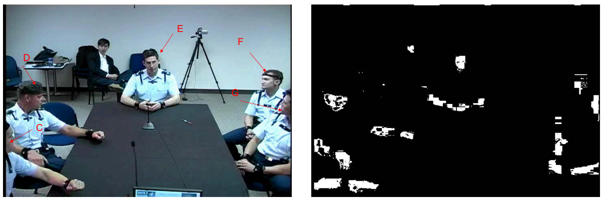


Figure 4. Meeting Room Configuration



(a) Video Taken by Camera C1 (b)Color Segmentation

Figure 5. Experimental Video and Color Segmentation

5 Experiments and Applications

5.1 Experimental Setup and Meeting Room Configuration

Figure 4 shows the original experimental setup and configuration of the meeting room. In our planning meeting experiments, there are eight participants labeled A B C D E F G H in the meeting. Ten movie cameras labeled C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 are installed to record the meeting events. T1 and T2 are two table microphones. Each camera is installed in a fixed position on the ceiling of the meeting room, so that each camera can see certain participants at the same time. We can set camera pairs to capture 3D data. With this configuration, we capture whole data for multimodal analysis of planning meetings. In this experiment, there are only five people (C D E F G) in the meeting shown in Figure 5 (a).

5.2 Results for a Meeting Dataset

This dataset comprised 74,805 frames of video captured five people in a meeting shown in Figure 5 (a). The five participants are labeled C, D, E, F, and G. Subject E is a leader of this meeting. We set three synchronized stereo with calibrated camera pairs to capture the whole meeting events, so that we can obtain three dimensional motion data. These camera pairs are C9 and C3 for subject E, C7 and C10 for subjects C and D, C2 and C5 for subjects F and G. By applying our hand tracking system with adaptive skin color models and filters to process this dataset and obtain hand 3D motion trajectories.

As we mentioned before, in order to handle the differences of skin color from different people, we build a skin color model for each participant. We collect skin color samples from their face and hand areas and apply them to train the models to obtain parameters. For subject E, the model parameters are

$$\hat{\mu}_E = \begin{pmatrix} 95.00 \\ 75.27 \end{pmatrix}, \quad \hat{\Sigma}_E = \begin{pmatrix} 33.98 & -10.06 \\ -10.06 & 96.67 \end{pmatrix}.$$

We create a skin color filter corresponding to the model for subject E. The threshold of the skin color filter is 150. In the same way, we build skin color models and filters to other participants, for participant C:

$$\hat{\mu}_C = \begin{pmatrix} 100.90 \\ 71.56 \end{pmatrix}, \quad \hat{\Sigma}_C = \begin{pmatrix} 92.52 & -24.88 \\ -24.88 & 137.44 \end{pmatrix}.$$

and the threshold 150 , for participant D:

$$\hat{\mu}_D = \begin{pmatrix} 96.12 \\ 76.35 \end{pmatrix}, \quad \hat{\Sigma}_D = \begin{pmatrix} 24.25 & -8.42 \\ -8.42 & 43.38 \end{pmatrix}.$$

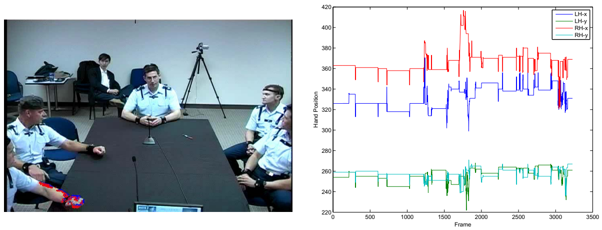
and threshold 150, for participant F:

$$\hat{\mu}_F = \begin{pmatrix} 97.58 \\ 69.70 \end{pmatrix}, \quad \hat{\Sigma}_F = \begin{pmatrix} 14.45 & 2.02 \\ 2.02 & 376.48 \end{pmatrix}.$$

and threshold 150, for participant G:

$$\hat{\mu}_G = \begin{pmatrix} 92.32 \\ 73.67 \end{pmatrix}, \quad \hat{\Sigma}_G = \begin{pmatrix} 13.53 & 4.08 \\ 4.08 & 238.26 \end{pmatrix}.$$

and threshold 150.



(a) Vector Computation (b)Hand Motion Trajectories

Figure 6. Results for the Meeting Dataset

With these skin color models and filters, we segment skin color regions and non-skin color regions for all subjects. Figure 5 (b) shows the result of the segments. The figure is an overlap of five figures which are obtained by the five skin color models mentioned above. From the results we can see that the skin color and non-skin color regions are separated well. In the mean time, we re-collect samples from the skin color regions and update skin color models and filters for next segmentations of the new frame. The next step is to select interest points in skin color regions and perform motion tracking with these interest points. With the skin color constraint, we only compute the motion vectors at skin color interesting points between frames. Figure 6 (a) shows one of the frames. We do not need to compute motion vector outside the skin color regions. By this way, we can reduce the computational work and can increase the processing speed. We also can improve the accuracy of motion vector clustering on hands, since no other motion vectors outside skin color regions effect the vectors on hands.

Finally, we cluster these motion vectors to identify the moving hands. Meanwhile, we can obtain hand motion trajectories. Figure 6 (b) shows an example of the results. Since the whole video is too long, in Figure 6 (b), we only display the hand motion trajectories of subject E in a segment including 3228 frames.

6 CONCLUSION

We presented our work on hand gesture tracking with adaptive skin color models and the application to multimodal analysis of planning meetings. We created a skin color model and filter for each participant in the meeting, so that we can handle the differences of skin color. We update the skin color model dynamically on the fly to fit the changes of the environments such as illuminants and backgrounds. By combining the original VCM algorithm with the adaptive skin color model to perform hand motion tracking, we can reduce the computational work and increase the processing speed and accuracy. We applied the improved algorithm to process our meeting experimental dataset. We obtained very satisfying results. In future, we will apply the algorithm to processing all of our 10 meeting experimental datasets and compare with the ground truth data.

7 ACKNOWLEDGMENTS

This research has been supported by the Advanced Research and Development Activity ARDA VACEII grant 665661: From Video to Information: Cross-Model Analysis of Planning Meetings.

References

[1] F. Quek and R. Bryll. Vector coherence mapping: A parallelizable approach to image flow computation. In *ACCV*, volume II, pages 591–598, Hong Kong, Jan. 1998.

[2] J.-C. Terrillon, M. David, and S. Akamatsu. Automatic detection of human faces in natural scene images by use of a skin color model and of invariant moments. In *FG98*, pages 112–117, Nara, Jp, Apr. 1998.

[3] V. Vezhnevets, V. Sazonov, and A. Andreeva. A survey on pixel-based skin color detection techniques. In *Proc. Graphicon-2003*, pages 85–92, Russia, Sep. 2003.

[4] Y. Xiong and F. Quek. Meeting room configuration and multiple camera calibration in meeting analysis. In *ACM ICMI2005*, pp. 37–44, Trento, Italy, Oct. 2005.

[5] J. Yang and A. Waibel. A real-time face tracker. In *Proceedings of the Third Workshop on Applications of Workshop on Computer Vision (WACV'96)*, Sarasota, Florida, 1996.