ENHANCING MODEL-BASED SKIN COLOR DETECTION: FROM LOW-LEVEL RGB FEATURES TO HIGH-LEVEL DISCRIMINATIVE BINARY-CLASS FEATURES

You-Chi Cheng[1], *Zhe Feng*[2], *Fuliang Weng*[2], *and Chin-Hui Lee*[1]

[1]School of ECE, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2]Robert BOSCH Research and Technology Center, Palo Alto, CA 94304, USA

ABSTRACT

We propose two very effective high-level binary-class features to enhance model-based skin color detection. First we find that the log likelihood ratio of the testing data between skin and non-skin RGB models can be a good discriminative feature. We also find that namely the background-foreground correlation provides another complementary feature compared to the conventional low-level RGB feature. Further improvement can be accomplished by Bayesian model adaptation and feature fusion. By jointly considering both schemes of Bayesian model adaptation and feature fusion, we attain the best system performance. Experimental results show that the proposed joint framework improves the 68% to 84% baseline F_1 scores to as high as almost 90% in a wide range of lighting conditions.

Index Terms— Discriminative feature, Bayesian adaptation, score fusion, likelihood ratio, skin color model

1. INTRODUCTION

Extracting interesting body parts in images or video clips provides highly salient information for many human-centric applications, such as sign language recognition [1], biometric authentication [2], and other human computer interaction (HCI) systems [3]. Due to illumination changes, reflections, shadows, and occlusions, it is often challenging to detect the human body parts in a robust manner under adverse operational conditions.

To alleviate such difficulties, enhanced sensors, such as infrared and depth cameras [4, 5], to get 3D information have been utilized. In addition, simultaneously detecting several body parts, such as face and limbs, proved to be beneficial [1, 6]. For detection of individual body parts, however, skin color detection and foreground-background segmentation are two mainstream approaches.

In this paper, we investigate skin color detection to localize hands in videos. In practical situations, skin regions can have dramatically different distributions in different backgrounds and under various illumination conditions [7]. However, it seems that a globally optimal color space representation cannot be easily determined at this moment. Therefore, instead of finding a robust feature set or color space representation, we study the skin color detection problem from a model-based perspective.

One key research challenge in model-based approaches to detection and classification problems is the selection and modeling of features. Several color spaces, such as RGB, normalized RGB, HSV, YCbCr, are all possible candidates for describing the skin color. One choice is to analyze an image frame with a block of pixels and take the average of RGB pixel values to represent the block in a three-dimensional RGB feature space. Unfortunately, other color representations cannot give statistically meaningful improvement in our study. Therefore, in this paper we explore additional high-level features available in the binary-class situations for skin color detection. One such feature is the log likelihood ratio between the densities of skin and non-skin models. Another discriminative feature is the correlation between skin foreground and non-skin background.

The main advantage of utilizing model-based methods is the possibility to feedback the evidences from the observed data into previously trained models so that the adjusted parameters of the trained models can not only better fit the observed data but also still maintain good generalization capability. One commonly used method is Bayesian model adaptation, such as maximum a posteriori (MAP) adaptation [8].

Furthermore, multi-modal information also proves to be beneficial to the detection performance in our study. Motivated by the successful experiences in [9], where scores generated by several modalities serve as features to train a feature fusion model, we fuse the regularized log likelihood ratio (LLR) feature and the foreground-background correlation coefficient into a single feature vector.

Skin and non-skin models can be trained on the above-mentioned feature vector to obtain an improved performance over the low-level RGB feature vector. The detection performance can be further enhanced if we replace the LLR features by MAP-adapted LLR features. Experimental results indicate that the proposed joint framework can effectively alleviate the model parameter mismatch problem and achieve the F_1 score as high as almost 90% across a wide range of lighting conditions.

2. SKIN COLOR MODELING AND FEATURES

As mentioned earlier, average RGB value within a block of size 16×16 is chosen as a feature vector. This low-level RGB feature vector is used to design our baseline detectors by modeling skin and non-skin classes as Gaussian mixture models (GMMs) [9, 10]. Two novel high-level features, log likelihood ratio (LLR) score and background-foreground correlation, available in the current binary-class setting will be proposed in the following subsections. Enhancement of the LLR feature through Bayesian adaptation and fusion of these two features will be presented subsequently.

2.1. Log Likelihood Ratio Features

Likelihood scores from K modalities for class C and $\widetilde{C}$, denoted by $\boldsymbol{f}(x|\boldsymbol{\theta}_C) = [f_1(x|\theta_C^1), ...f_K(x|\theta_C^K)]^T$ and $\boldsymbol{f}(x|\boldsymbol{\theta}_{\widetilde{C}}) = [f_1(x|\theta_{\widetilde{C}}^1), ...f_K(x|\theta_{\widetilde{C}}^K)]^T$, respectively, can also serve as features to design skin color detectors because they have been used in hypothesis testing to design test statistics for verifying the validity of certain events. When $K = 1$, the log likelihood ratio

$$LLR(x|\theta_C, \theta_{\widetilde{C}}) = \log f(x|\theta_C) - \log f(x|\theta_{\widetilde{C}}) \qquad (1)$$

produced by the skin and non-skin models will be used in this study as a high-level, discriminative binary-class feature.

2.2. Background-Foreground Similarity Features

2.2.1. Preliminary Background Identification

When images are given by video clips instead of with individual pictures, we can get some information by comparing the background (non-skin) and foreground (skin) regions. Motivated by voice activity detection (VAD) [11] in the speech community, we designed an energy-based algorithm to identify background frames.

Assume the first frame of each clip is always a background frame. Then all other frames in the same clip are examined using a given threshold η, $0 < \eta < 1$. For the i^{th} frame, if its sample standard deviation σ_i satisfies the following inequality, it is considered as a background,

$$\frac{\sigma_i - \sigma_{min}}{\sigma_{max} - \sigma_{min}} < \eta \tag{2}$$

Contiguous background frames are then averaged to form an averaged background for future comparison.

2.2.2. Background-Foreground Similarity

Having identified background frames in Section 2.2.1, blocks in each non-background frame are compared with their corresponding background blocks to check if they are in the same class.

A straightforward way to test if two blocks in similar images belong to the same class is to examine their sum of squared errors. However, this method can be problematic if the block pixels are under affine transformation, which is similar to the illumination model proposed in [12]. The linear correlation coefficient ρ, however, is more robust to this variation.

$$\rho = \frac{Cov(f,g)}{\sqrt{Var(f)Var(g)}} \tag{3}$$

Under the jointly Gaussian and uncorrelated assumption of the pixel vectors in a local image foreground block f and its corresponding background block g, the test statistic $t = \sqrt{\rho^2(M-2)/1-\rho^2}$ has the t-distribution [13] , where M=16×16 is the number of samples in each block being analyzed. Since $|t|$ is often used in t-test and is also a strictly increasing function of $|\rho|$ in the range of $[0, 1]$, for easy normalization, $|\rho|$ is chosen as a feature to measure the background-foreground similarity at each image block. We believe this is also an effective high-level discriminative binary-class feature to be used for skin color detection.

3. BAYESIAN MODEL ADAPTATION

When training and testing conditions are different, parameters obtained from the training phase may not be able to describe the actual distribution of the testing data. Bayesian model adaptation, such as maximum a posteriori (MAP) adaptation [8], is often adopted under this circumstance in the speech community. MAP adaptation assumes the form of the prior density to be a conjugate prior of the probability density function of the feature vectors, and therefore the posterior density will not only include the observed data information but also have the same form as the prior density [8]. The information of the observed data and the prior density are effectively combined under this framework. Assume the likelihood function for class C is denoted by a K-mixture GMM: $f(x|\theta_c) = \Sigma_{i=1}^{K} w_i \mathcal{N}(x|\mu_i, \Sigma_i)$, as shown in [8], given a set of adaptation data, $x_n, n = 1, ...N, x_n \in \mathcal{R}^D$, parameter adaptation can be done as follows:

$$w_i^{MAP} = \frac{\nu_i - 1 + \sum_{n=1}^{N} c_{in}}{N - K + \sum_{j=1}^{K} \nu_j}, \tag{4}$$

$$\mu_i^{MAP} = \frac{\tau_i m_i + \sum_{n=1}^{N} c_{in} x_n}{\tau_i + \sum_{n=1}^{N} c_{in}}, \tag{5}$$

$$\Sigma_i^{MAP} = \frac{u_i + \tau_i (m_i - \mu_i^{MAP})(m_i - \mu_i^{MAP})^T}{\alpha_i - D + \sum_{n=1}^{N} c_{in}} + \frac{\sum_{n=1}^{N} c_{in}(x_n - \mu_i^{MAP})(x_n - \mu_i^{MAP})^T}{\alpha_i - D + \sum_{n=1}^{N} c_{in}}, \tag{6}$$

where

$$c_{in} = \frac{w_i \mathcal{N}(x_n|\mu_i, \Sigma_i)}{\sum_{j=1}^{K} w_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \tag{7}$$

and ν_i is the i^{th} hyperparameter of the prior density of mixture weights and τ_i, m_i, u_i, α_i are hyperparameters of the prior density for i^{th} Gaussian mixture:

$$p(w_1, ...w_K) \propto \prod_{i=1}^{K} w_i^{\nu_i - 1}, \tag{8}$$

$$p(\mu_i, \Sigma_i^{-1}) \propto |\Sigma_i^{-1}|^{\frac{\alpha_i - D}{2}} exp[-\frac{1}{2} tr(u_i \Sigma_i^{-1})] \times exp[-\frac{\tau_i}{2}(\mu_i - m_i)^T \Sigma_i^{-1}(\mu_i - m_i)]. \tag{9}$$

We will not adapt the covariance matrix to save computational cost. Also, the adaptation is performed in an unsupervised manner: the current model set is first used to decode the test samples to a specific class. Such labels are then treated as supervision information for that class and is then used to perform the above MAP adaptation algorithm. An example of resolving model parameter mismatch by MAP adaptation will be given in Section 5.

4. FUSION OF DISCRIMINATIVE FEATURES

Score fusion has been proved to be useful to integrate information from several modalities or features [9, 14]. It is usually done in two different ways. One is to find a simple transform to combine the outputs of several classifiers online, the other is to treat scores generated by competing classifiers as features to train higher level classifiers. We investigate the latter one in this study.

In this study, we choose the log likelihood ratio of skin vs. non-skin color models defined in Eq. (1) and the linear correlation coefficient defined in Eq. (3) as the elements to be combined. Borrowed from the experience in [9], these two discriminative features can be combined in $\mathcal{R}^2$ to train fusion classifiers modeled by GMMs for skin and non-skin classes, respectively.

To make sure the model will not favor a specific feature, the log likelihood ratio $LLR(x|\theta_C, \theta_{\widetilde{C}})$ is regularized. Here we choose a logistic sigmoid function defined as follows:

$$\gamma(LLR(x|\theta_C, \theta_{\widetilde{C}})) = \frac{1}{1 + exp[-\alpha(LLR(x|\theta_C, \theta_{\widetilde{C}}) - \beta)]} \tag{10}$$

with empirically chosen parameters $(\alpha, \beta) = (0.5, 0)$ to transform LLR scores into the range of $[0, 1]$. Also, as stated in Section 2.2.2,

we choose $|\rho|$ instead of ρ to make it also in the range of $[0, 1]$. Thus the fusion feature for a 16×16 block used to train fusion GMMs is $[\gamma(LLR(x|\theta_C, \theta_{\bar{C}}))\ |\rho|]^T$. This feature will be shown to have enhanced discriminative powers over the RGB features.

Furthermore, we proposed a joint framework which uses MAP adaptation mentioned in Section 3 to replace the LLR score with MAP-adapted LLR score under the previously stated fusion scheme. This joint framework can simultaneously enhance discriminative power and compensate the mismatch between training and testing conditions. As we will discuss soon in Section 5, the proposed joint framework is very effective under highly-mismatched conditions.

5. EXPERIMENTAL RESULTS

Our dataset is recorded using Dragonfly 2 color camera with the setup of 15 fps 1024×768 YUV422 and 10 ms shutter speed to remove blurring artifacts caused by possibly fast hand motion. All frames were resized to 256×192 to save storage space. 16 users were recorded at 3 time slots of a day: morning, noon, and evening, resulting in roughly 4000 clips with about 16,000 frames.

In order to study the performance under different recording conditions, we first used frame-average RGB values to cluster the whole dataset into three categories. After a brief data analysis, we found Categories I, II, and III were mostly recorded in the morning, noon, and evening, respectively. And the ratio of frame numbers in Categories I, II, and III is roughly 4:2:1. We randomly selected 1/3 of the data from each category for manual labeling with 16×16 blocks and randomly divided the labeled data evenly for training and testing. GMMs with mixture numbers determined by [15] were used for modeling. Apparently the category with most number of samples, i.e., Category I, will dominate the training. Based on their different illumination conditions, we respectively marked Categories I, II, and III as well-matched, slightly-mismatched, and highly-mismatched.

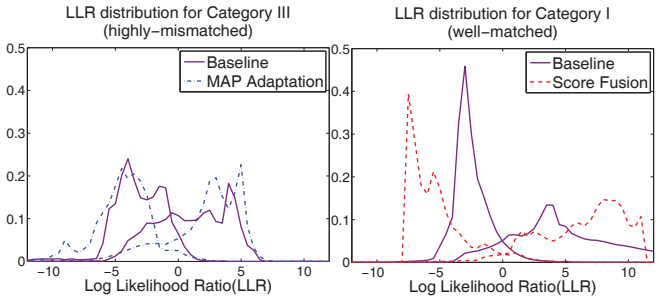
Besides, due to imbalanced number of samples between skin and non-skin blocks, which is 1:4, accuracy ends up being a bad performance index. To fairly evaluate the performance, indices measuring correctly classified skin blocks should be used. To do so, let (TP, TN, FP, FN) denote numbers of correctly classified skin blocks, correctly classified non-skin blocks, non-skin blocks classified as skin blocks, and skin blocks classified as non-skin blocks. The precision, recall, and F_1 are defined as $TP/(TP + FP)$, $TP/(TP+FN)$, and $2\times precision\times recall/(precision+recall)$, which is a harmonic mean of the precision and recall, respectively.

The advantages of applying MAP (Section 3) and fusion of features (Section 4) over the baseline (Section 2) are shown in Figure 1. Note that the misclassification rate can be visualized by the overlapping area under 2 LLR distribution curves. In a highly-mismatched scenario, applying MAP adaptation, as shown in the dash-dot lines in Figure 1(a), the reduction of between-class overlap is clear. In a well-matched condition, applying feature fusion can still enhance the discriminative power, as shown in the dashed lines in Figure 1(b).

F_1 scores with LLR decision threshold 0 are listed in column 3 of Table 1. We can see that MAP adaptation with 5 iterations outperformed the baseline score by 6.4 to 14.1% and applying fusion scheme on baseline score can beat the MAP adaptation only in category II by 0.63% and degraded by 0.31% and 5% in Categories I and III, respectively. Although 0.63% and -0.31% are within the range of statistical error, the 5% degradation in Category III clearly shows the difficulty of the original fusion scheme under mismatched conditions. The proposed joint framework, however, consistently outperformed MAP adaptation and the original fusion scheme by 1.5 to 5 % and 1.8 to 9.7 %, respectively. Next, we check the best F_1 scores for each algorithm, as listed in column 4 of Table 1. We can see that MAP adaptation is better than the baseline by 4.5 to 13.9%, and score fusion outperformed MAP by 1.4 to 2% except in Category III, and the proposed framework, again, outperformed MAP adaptation and the original fusion scheme by 2.2 to 6.0%, and 0.8 to 10.1%, respectively. The proposed method effectively combined the advantages of both methods and obtained an F_1 as high as 89.0 to 92.6%. Especially, it can effectively compensates the mismatch between the dark environment (Category III) and the training condition.

Table 1. The Performance Evaluation for threshold=0

Category I	Precision	Recall	F_1	F_1^{best}
Baseline	79.51	89.01	83.99	85.88
MAP adaptation	94.50	86.55	90.35	90.36
Score fusion	84.83	95.91	90.03	91.77
Proposed framework	90.98	92.85	91.91	92.56
Category II	**Precision**	**Recall**	F_1	F_1^{best}
Baseline	69.16	86.90	77.02	78.69
MAP adaptation	90.98	84.73	87.74	87.90
Score fusion	82.15	95.62	88.37	89.82
Proposed framework	91.79	91.85	91.82	91.98
Category III	**Precision**	**Recall**	F_1	F_1^{best}
Baseline	77.16	61.74	68.59	69.12
MAP adaptation	87.15	78.77	82.75	83.04
Score fusion	71.83	85.95	78.26	78.90
Proposed framework	86.69	89.25	87.95	89.03



(a) MAP adaptation (b) Fusion of features

Fig. 1. Separation: (a) before (solid line) and after (dash-dot line) adaptation for Category III data; and (b) before (solid line) and after (dashed line) score feature fusion for Category I data. The misclassification rates (intersection areas) were halved for both scenarios.

A clearer picture can be found in Figure 2. The curve trend indicates that in well-matched cases, i.e., Category I, score fusion can outperform MAP adaptation in general, and was not good in slightly-mismatched Category II scenarios, and became much worse in highly-mismatched Category III conditions. But the proposed method could outperform all algorithms when the precision exceeds 90%. In highly-mismatched conditions, it can compensate for the mismatch very well.

6. CONCLUSION

In this paper, we explored two high level features available in binary detection and classification problems, namely likelihood ratio of competing skin and non-skin models and correlation between

foreground and background, for skin color detection. We further enhanced the robustness of these feature by embedding MAP adaptation into the original feature fusion scheme. Experiments demonstrated the effectiveness of the proposed method under highly mismatched condition as well as the general ability to achieve high F_1 scores, which is close to 90% in all three categories from bright to dark lighting conditions.

7. REFERENCES

[1] K. Imagawa, S. Lu, and S. Igi, "Color-based hands tracking system for sign language recognition," in *Proc. of FGR 1998*, pp. 462–467.

[2] R.K. Rowe, U. Uludag, M. Demirkus, S. Parthasaradhi, and A.K. Jain, "A multispectral whole-hand biometric authentication system," in *Proc. Biometrics Symposium, 2007*, pp. 1–6.

[3] V.I. Pavlovic, R. Sharma, and T.S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 19, no. 7, pp. 677–695, 1997.

[4] G. Lu, D. Zhang, and K. Wang, "Palmprint recognition using eigenpalms features," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1463–1467, 2003.

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. of CVPR 2011*, pp. 1297–1304.

[6] L. Brethes, P. Menezes, F. Lerasle, and J. Hayet, "Face tracking and hand gesture recognition for human-robot interaction," in *Proc. of ICRA 2004*, vol. 2, pp. 1901–1906.

[7] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognition*, vol. 40, no. 3, pp. 1106–1122, 2007.

[8] J.L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and Audio Processing, IEEE Trans. on*, vol. 2, no. 2, pp. 291–298, 1994.

[9] K. Nandakumar, Y. Chen, S.C. Dass, and A. Jain, "Likelihood ratio-based biometric score fusion," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, pp. 342–347, 2007.

[10] T.S. Jebara and A. Pentland, "Parametrized structure from motion for 3d adaptive feedback tracking of faces," in *Proc. of CVPR 1997*, pp. 144–150.

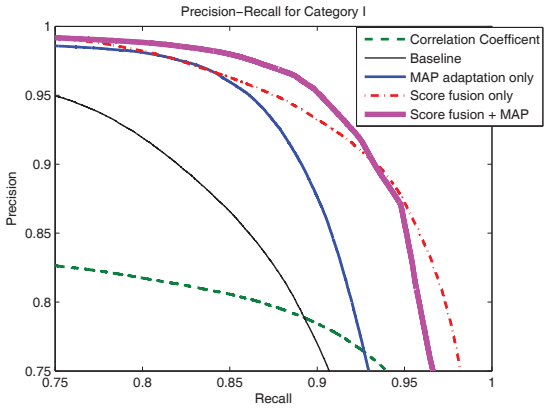
[11] J. Ramirez, JM Górriz, and JC Segura, "Voice activity detection. fundamentals and speech recognition system robustness," *Robust Speech Recognition and Understanding*, pp. 1–22, 2007.

[12] H.W. Haussecker and D.J. Fleet, "Computing optical flow with physical models of brightness variation," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 23, no. 6, pp. 661–673, 2001.

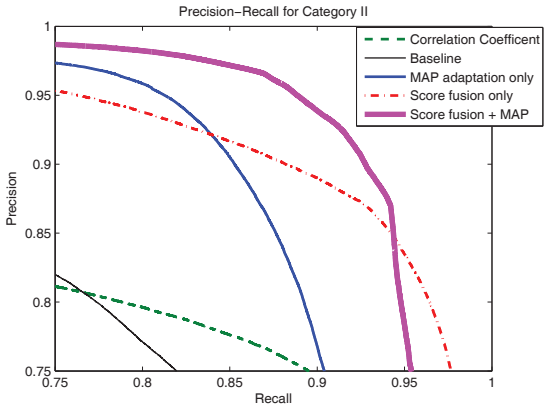
[13] H.M. Walker and J. Lev, *Statistical Inference.*, Henry Holt and Company, 1953.

[14] M. Xu, L.Y. Duan, C.S. Xu, and Q. Tian, "A fusion scheme of visual and auditory modalities for event detection in sports video," in *Proc.of ICASSP 2003*, vol. 3, pp. 189–192.

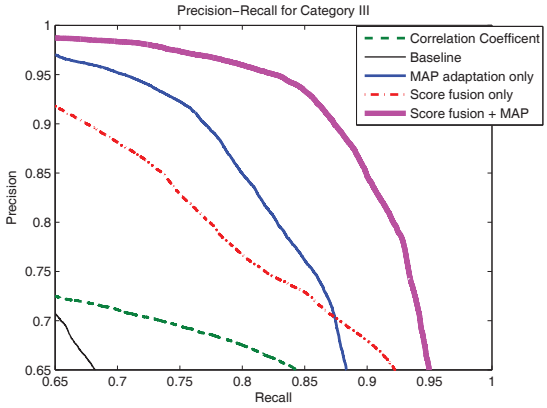
[15] M.A.T. Figueiredo and A.K. Jain, "Unsupervised learning of finite mixture models," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, pp. 381–396, 2002.



(a) Category I: Normal illumination, well matched



(b) Category II: Bright illumination, slightly mismatched



(c) Category III: Dark illumination, highly mismatched

Fig. 2. Precision-recall curves for 3 different testing categories, we can see that the purple lines have the maximum F_1 scores when precision is asked to be more than 90%. Applying high-level feature fusion (also known as score fusion in this work) scheme is in general better than MAP adaptation in well-matched conditions (Category I), but deteriorates in slightly-mismatched conditions (Category II), degrades severely when the condition is highly-mismatched (Category III). The proposed framework can solve this difficulty and make the dark illumination condition (Category III) receives the most improvement.