A Hybrid Technique for Face Detection in Color Images

Saman Cooray and Noel O'Connor

Centre for Digital Video Processing
Dublin City University, Ireland
{coorays,oconnorn}@eeng.dcu.ie

Abstract

In this paper, a hybrid technique for face detection in color images is presented. The proposed technique combines three analysis models, namely skin detection, automatic eye localization, and appearance-based face/non-face classification. Using a robust histogram-based skin detection model, skin-like pixels are first identified in the RGB color space. Based on this, face bounding-boxes are extracted from the image. On detecting a face bounding-box, approximate positions of the candidate mouth feature points are identified using the redness property of image pixels. A region-based eye localization step, based on the detected mouth feature points, is then applied to face bounding-boxes to locate possible eye feature points in the image. Based on the distance between the detected eye feature points, face/non-face classification is performed over a normalized search area using the Bayesian Discriminating Feature (BDF) analysis method. Some subjective evaluation results are presented on images taken using digital cameras and a webcam, representing both indoor and outdoor scenes.

1. Introduction

Detecting human faces in images and video sequences is an important task for many applications. Enhanced video surveillance and security related applications, in particular, are closely related with human face identification where the task of face detection becomes a major requisite to deliver efficient and robust performance under challenging conditions. Face recognition leading to law-enforcement applications using mugshot databases also requires to accurately detect faces in images of varying quality. Many important applications in Human Computer Interaction (HCI) consider the human face as the main entity with which computers can be manipulated. Furthermore, image and video retrieval applications can also benefit from using human faces as a semantic object available for indexing the content. Due to this need for efficient face detection techniques, a large number of algorithms have been put forward by researchers. A detailed survey of existing face detection algorithms, though not up to date, can be found in [1].

Existing face detection algorithms can be divided into the following categories: (i) feature-based methods; (ii) template matching methods; and (iii) appearance-based methods. Feature-based methods exploit the properties of facial features and their structural relationships in facial images. Algorithms based on color images in this category generally use the skin detection step as an initial phase of face localization before subsequent analysis is performed for further verification. It is then followed by face verification methods, which are mostly based on criteria such as the shape of the detected skin blobs, the presence of facial features such as eyes and mouth inside the detected blobs [2], or the combination of the skin blob's shape and facial features [3]. Template matching algorithms, on the other hand, are usually applied to grey-scale images where predefined or flexible templates of faces are compared with image blocks for matching [4]. In appearance-based methods, face models are derived using a large amount of training data, thereby incorporating the possible variations of human faces in real-life images. The recent methods proposed under this category perform face detection using Bayesian Discriminating Feature analysis [5], Neural Networks (NN) methods [6], Support Vector Machines (SVM) [7], etc.

In this paper, we propose a frontal face detection technique combining skin detection, facial eye localization, and appearance-based face classification. The first phase of our algorithm is similar to the approach proposed by Hsu *et al.* [3] where a facial feature extraction step is employed on face bounding-boxes that are derived through skin detection. In their approach, face localization is performed using a lighting-compensated skin detection model, however, we carry out this task using a statistical histogram-based skin detection model. They exploit facial features such as eyes, mouth and head contours for face verification, our approach uses only mouth and eye facial features which collectively facilitate the subsequent use of an appearance-based face classification method. One objective of our research is to investigate how a fast and robust face detection technique

0-7803-9385-6/05/$20.00 ©2005 IEEE

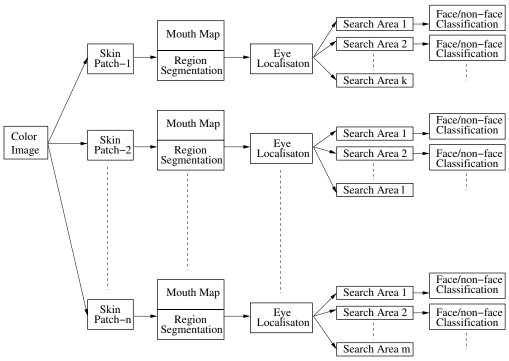


Figure 1: System block diagram.

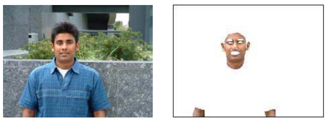
could be developed using the combination of skin detection, facial features, and appearance–based classification. The use of skin pixels facilitates localizing computation to certain parts of the image only. Detecting eye feature points is useful to limit the search space as well as avoiding computationally costly image rotation and processing steps for detecting rotated faces. Finally, a face/non–face classification model is used to determine if the candidate face is a correct face or not.

2. Overall Approach

Our approach to face detection comprises several classification stages. First, it detects skin patches in the image using the statistical skin detection model described in section 3. These skin patches, hereafter called face bounding–boxes, are then processed to verify if they contain faces or not. The first verification process is based on two steps: (i) mouth detection; (ii) eye localization. While the detection of mouth candidates is carried out using the redness property of pixels around lips, eye localization is facilitated by a region–based segmentation algorithm along with the detected mouth feature points. Upon detecting eye feature points within a face bounding–box, a normalized search space (scaled and rotated) is derived relative to the distance between the eye feature points. We then apply an appearance–based face/non–face classification technique to the search space, in order to verify that these candidate face bounding–boxes do actually contain faces. Fig. 1 shows the block diagram of our face detection system.

3. Histogram-Based Skin Segmentation

Many face detection techniques utilizing the skin segmentation step were dependent on simple methods, such as direct threshold methods [8], single or mixture Gaussian models [9]. Direct threshold type skin segmentation methods rely on the fact that human skin colors fall into a small region in the chosen color space, and hence robust skin segmentation becomes problematic against the illumination. On the other hand, Gaussian models despite being a more effective skin modeling method suffer from slow performance, making them less effective in real–time applications. Thus, a more effective methodology to skin segmentation in static images is to use a histogram–based skin segmentation model despite requiring large sets of skin and non–skin training data for modeling the human skin color distribution [10][11].



(a) (b)

Figure 2: Example skin and non–skin training data used (a) Original color image containing both skin and non–skin data (b) Manually skin–segmented image.

Table 1: Description of the Skin/non–skin Training Data

Total no. of skin pixels	66,167,894
No. of unique non–skin pixels	3,172,649
Total no. of non–skin pixels	386,402,186
No. of skin and non–skin overlapping pixels	671,531
No. of unique skin pixels	86,268
No. of unassigned pixels	12,846,768

Considering the two classes, ω_s and ω_n, representing skin and non–skin data distribution respectively, a given pixel X can be classified to be skin if

$$\frac{p(X/\omega_s)}{p(X/\omega_n)} \geq TH \qquad (1)$$

where $p(X/\omega_s)$ and $p(X/\omega_n)$ are the conditional probability density functions of skin and non–skin data while TH is a threshold, which will govern the level of the $false\ detection$ rate and the $false\ rejection$ rate.

3.1. Skin Model: Training Data

Our skin segmentation model was trained using a total number of 2300 color images each of which were manually segmented into skin and non–skin regions. Fig. 2 shows an example of a manually skin–segmented image from which both skin and non–skin data are extracted for training the model. Note that non–skin data is extracted from the complements of the skin segmented images.

The set of 2300 images comprises about 950 images from the ECU skin image database [11], about 100 indoor and outdoor scenes taken using digital cameras, about 70

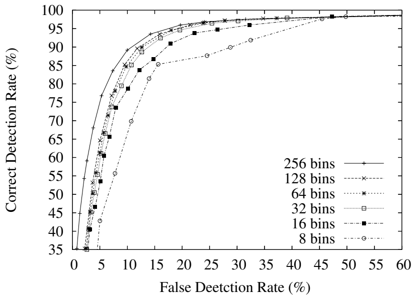


Figure 3: Segmentation performance for different histogram quantization levels.

image frames decoded from broadcast TV, about 400 images taken from the standard face databases such as HHI MPEG-7, FERET, Altkom and Champion, and the rest downloaded from the web. This set, when manually segmented (see Fig. 2b), accounted for a total number of over 66 million skin pixels. The complements of the skin labeled images accounted for about 386.4 million pixels which corresponds to the set of non-skin data. A statistical description of the training data used in this model is given in Table 1.

3.2. Skin Model: Segmentation Performance

The performance of the skin detection model was measured over a test data set comprising 400 color images. The first 200 images of this set was taken from broadcast video and downloaded from the web while the next 200 images were from the FERET database containing 65 black skin images, 65 dark skin images, and 70 white skin images.

The performance analysis of the histogram model is given in Fig. 3 in terms of correct detections vs false detections. This graph illustrates the behavior of this skin detection model for 6 histogram quantization levels, i.e. 256, 128, 64, 32, 16 and 8 bins per channel. The analysis indicates that the quantization level of 256 bins/color corresponds to the best accuracy, an observation which agrees with Phung *et al.* but contradicts Jones and Rehg [11]. The coarser quantization levels, however, indicate that up to 32 bins/color would be bearable given the accuarcy/speed compromise many applications require.

In our experiments, we use the 256 bins/color skin segmentation model defined at 20% false detection rate. Upon detecting skin pixels, face bounding-boxes are derived using the spatial coherence information of the skin bitmap clusters followed by a simple morphological hole-filling technique. A lower limit on the spatial size of such detected candidates is also assigned, in order to remove small objects which are unlikely to be faces.

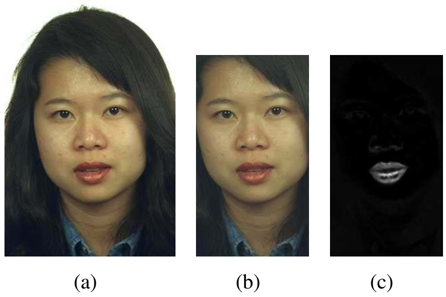

(a) (b) (c)

Figure 4: Mouth detection (a) original color image (b) face bounding-box obtained from skin segmentation (c) mouth-map image.

4. Automatic Eye Localization

While there are several possible facial features being utilized in face detection, we consider only mouth and eye feature points to be the prominent features in our face detection approach. We employ the same method proposed by Hsu *et al.* [3] for detecting mouth feature points based on redness characteristics of mouth pixels. Although this criterion does not classify mouth features available in a given image explicitly, it allows the detection of candidate mouth regions. In our approach, the eye localization task is carried out relative to the position of the mouth candidates.

4.1. Mouth Detection

The mouth-like regions are detected using the mouth map criterion proposed by Hsu *et al.* [3]. Fig. 4(c) shows a mouth-map when (2) is applied to a face bounding-box shown in Fig. 4(b). Finally, the positions of the mouth feature points are obtained by applying a threshold criterion and a hole-filling technique to the mouth-map image.

$$MouthMap = Cr^2.(Cr^2 - \eta.Cr/Cb)^2 \tag{2}$$

where

$$\eta = 0.95 \times \frac{(1/N)\sum Cr^2}{(1/N)\sum(Cr/Cb)} \tag{3}$$

where N represents the spatial size of the face bounding-box.

4.2. Recursive Shortest Spanning Tree Algorithm

A region-based color segmentation algorithm, called Recursive Shortest Spanning Tree (RSST), is used in our face detection system as a means of detecting eye feature points. RSST, in general, can be efficiently used to segment a given image into a desired number of regions. In the conventional algorithm, the region merging sequence is defined using both the luminance and chrominance properties of regions.

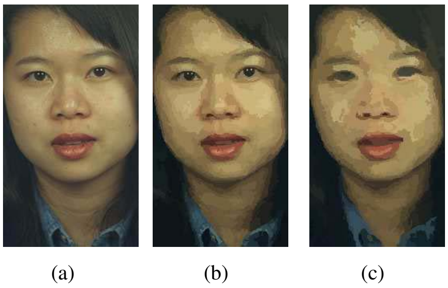


Figure 5: RSST Segmentation (a) face-bounding box (b) segmentation from luminance and chrominance merging (c) segmentation from chrominance merging.

However, the presence of bright and dark pixels in eyes drives us using a slightly different merging distance criterion, which considers only the chrominance components in the image. The original and modified merging distances are defined by (4) and (5). The distance $d(R1, R2)$ represents the merging distance between two regions $R1$ and $R2$ with their mean luminance, chrominance and spatial size represented by $Y(R)$, $Cb(R)$, $Cr(R)$ and $N(R)$ respectively. Fig. 5(b) and Fig. 5(c) illustrate the different types of regions obtained from these two distance measures. It can be noted that when chrominance only merging is used eye regions appear as grey blobs, and hence can be distinctively separated from other image regions.

$$\begin{aligned} d(R1,R2) &= [Y(R1)-Y(R2)]^2+[Cb(R1)-Cb(R2)]^2 \\ &+ [Cr(R1)-Cr(R2)]^2 \times \frac{N(R1)\times N(R2)}{N(R1)+N(R2)} \quad (4) \\ d(R1,R2) &= [Cb(R1)-Cb(R2)]^2 \\ &+ [Cr(R1)-Cr(R2)]^2 \times \frac{N(R1)\times N(R2)}{N(R1)+N(R2)} \quad (5) \end{aligned}$$

4.3. Eye Detection

An important statistical property of eye image regions is that they signal high intensity variance due to the fact that human eyes generally contain both black (near black) and white (near white) regions. This feature can be captured by using a chrominance-based region merging criterion in the RSST segmentation algorithm. Fig. 6 shows an example of a set of regions illustrating the intensity variance distributions against their 2D positions in the original image. In this example, a total number of 170 regions were present in the segmentation shown in Fig. 5(c). This shows an important statistical property of regions that there are only a few regions of high intensity variance.

The above variance measure combined with some heuristic rules based on regions' geometrical and structural properties in frontal face images is used for eye detection. These heuristic rules are:

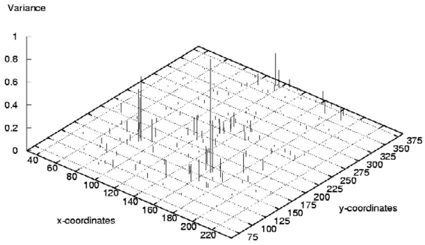


Figure 6: Variance distribution of regions.

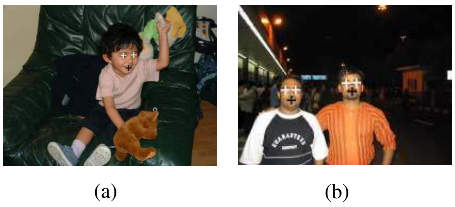


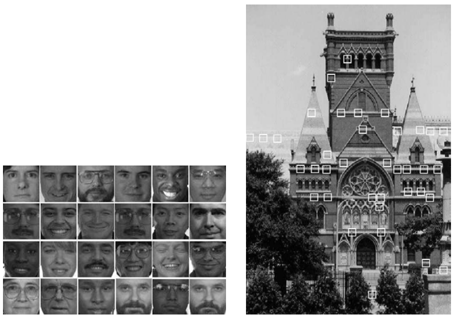
Figure 7: Detected eye/mouth feature points.

- Eye region should be at least 10 pixels above the mouth level;
- Width/height ratio of eye regions should be at least 0.4;
- Distance from the mouth to the left and right eyes should be within a pre-defined range;
- Angle between the mouth and the eyes should be between 35 degrees and 80 degrees;
- The x-coordinate of the mouth feature point should be located in between the x-coordinates of eye feature points;
- Eye region should correspond to a dark blob in the image [12].

Fig. 7 shows some example results from the facial feature extraction technique used in this face detection system.

5. Face/Non-face Classification

The process of training the face and non-face class models is carried out using 2400 faces and 4500 non-faces in our system. We chose frontal and near frontal view 16×16 normalized face images as the training images of faces from the FERET database and the ECU database [11] (see Fig. 8(a)). Some non-face images were initially identified by applying the face detection algorithm only with the face class error criterion to grey-scale images that do not contain human



(a) Example face samples (b) Creating non-face images

Figure 8: Face and non-face images used in training.

faces. By subsequently applying the face detection algorithm using both face and non-face models to several grey-scale images in a bootstrap manner, further non-face candidates were obtained. An example of this process is shown in Fig. 8(b).

5.1. Bayesian Discriminating Feature Method

In our approach, the face classification task is performed using the BDF method, which was originally proposed by Liu for face detection in grey-scale images [5]. We employ the BDF classifier in our system based on locations of the eye feature points detected using the automatic eye localization technique described in section 4.

In this method, both face class ω_f and non-face class ω_n are modeled as a multivariate normal distribution defined by (6) and (7).

$$\begin{aligned} p(Y|\omega_f) & \\ &= \tfrac{1}{2\pi^{N/2}|\Sigma_f|^{1/2}} \exp\{-\tfrac{1}{2}(Y-M_f)^t\Sigma_f^{-1}(Y-M_f)\} \quad (6) \\ p(Y|\omega_n) & \\ &= \tfrac{1}{2\pi^{N/2}|\Sigma_n|^{1/2}} \exp\{-\tfrac{1}{2}(Y-M_n)^t\Sigma_n^{-1}(Y-M_n)\} \quad (7) \end{aligned}$$

where mean feature vectors of face/non-face class are denoted by M_f/M_n and covariance matrices of face/non-face class are denoted by Σ_f/Σ_n respectively.

Based on this representation, Liu [5] derives error terms δ_f and δ_n for face and non-face classes respectively, which are defined by (8) and (11).

$$\begin{aligned} \delta_f &= \sum_{i=1}^{M}\frac{z_i^2}{\lambda_i} + \frac{\|Y-M_f\|^2 - \sum_{i=1}^{M} z_i^2}{\rho} \\ &+ ln(\prod_{i=1}^{M}\lambda_i) + (N-M)ln\rho \qquad (8) \end{aligned}$$

where z_i are the principal components of Z defined by equation 10 below, ρ is the average sum of the remaining eigenvalues defined by equation 9 below, and λ_i are the eigenvalues of face class.

$$\rho = \frac{1}{N-M}\sum_{k=M+1}^{N}\lambda_k \qquad (9)$$

$$Z = \phi_f^t(Y-M_f) \qquad (10)$$

$$\begin{aligned} \delta_n &= \sum_{i=1}^{M}\frac{u_i^2}{\lambda_i^{(n)}} + \frac{\|Y-M_n\|^2 - \sum_{i=1}^{M} u_i^2}{\epsilon} \\ &+ ln(\prod_{i=1}^{M}\lambda_i^{(n)}) + (N-M)ln\epsilon \qquad (11) \end{aligned}$$

where u_i are the principal components of U defined by equation 13 below, ϵ is the average sum of the remaining eigenvalues defined by equation 12 below, and $\lambda_i^{(n)}$ are the eigenvalues of non-face class.

$$\epsilon = \frac{1}{N-M}\sum_{k=M+1}^{N}\lambda_k^{(n)} \qquad (12)$$

$$U = \phi_n^t(Y-M_n) \qquad (13)$$

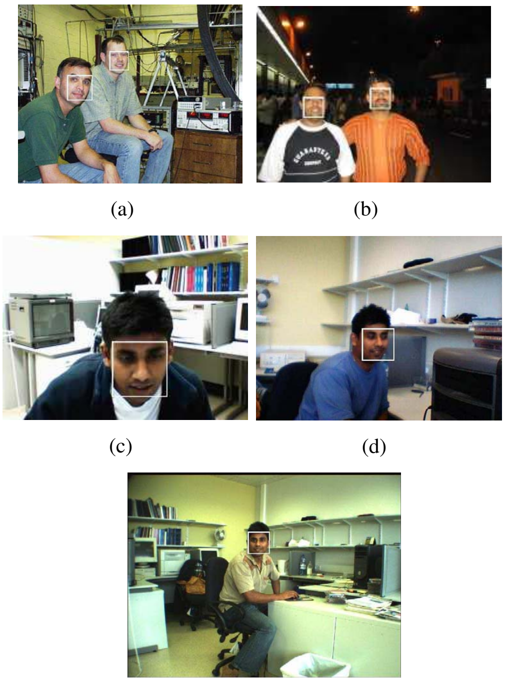
Using the Bayesian decision rule, the current subimage of the test image is classified to be a face if the following condition is satisfied [5].

$$\delta_f < \theta \quad and \quad \delta_f < (\delta_n - \tau) \qquad (14)$$

where θ and τ are empirically found parameters. In our approach, the parameters θ and τ are set to 550 and 100 respectively.

5.2. Combining Eye Features with Face Classification

The final stage of face detection is carried out by applying the BDF classifier to a normalized search space. Depending on the detected locations of the eye feature points, we scale and rotate the image so that the final eye-eye distance becomes 8 pixels and both eyes lie in a horizontal line. Rotation is carried out only if the angle between eye feature points is greater than 5 $degrees$. Hence, this process allows us to find a smaller search space in the image, which is defined by $(3d \times 3d)$ where d is the distance between the eye feature points in the training images, i.e. 8 pixels. We then apply the BDF classifier using (14) to decide if there is a valid face inside the current face bounding-box.



(a) (b)

(c) (d)

(e)

Figure 9: Face detection results (a) indoor image taken using a digital camera (b) outdoor image taken using a digital camera (c)-(d) two indoor images taken using a webcam (e) indoor image taken using a Digiclops stereo camera.

6. Results

Fig. 9(a) and Fig. 9(b) show the detected faces marked with white squares on the original images taken using ordinary digital cameras. Fig. 9(c) and Fig. 9(d) show the detected faces of two slightly out-of-plane rotated images taken using a Creative webcam, and Fig. 9(e) shows the results on an image taken using a Digiclops trinocular stereo camera. Our algorithm is capable of detecting frontal, near-frontal and slightly rotated (both in-plane and out-of-plane) faces, however, the images taken from webcams/stereo cameras were found more difficult to be dealt with. These difficulties were mainly due to the lack of eye details present in the images: an effect primarily caused by indoor lighting conditions.

It is envisaged that upon creating image/video databases using webcam/surveillance cameras, in both indoor and outdoor environments, an objective evaluation of this algorithm will be carried out in order to validate its performance.

7. Conclusion

We have presented a hybrid technique for detecting frontal faces in color images using skin segmentation, facial eye localization, and appearance-based face/non-face classification. The objective of our research is to develop an efficient face detection algorithm to address the challenging issues related with content-based image retrieval in surveillance types of applications. However, the current challenges lie in detecting eye/mouth feature points robustly in partially occluded faces and poor quality images. Skin detection can also be erroneous in some cases. These issues are very challenging and will be the focus of our future research.

Acknowledgments

Part of this material is based on the MediAssist project funded by the Enterprise Ireland. The support by Science Foundation Ireland under Grant Nos. 03/IN.3/I361 is also acknowledged.

References

[1] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Tran. on PAMI*, vol. 24, pp. 34–58, Jan. 2002.

[2] H. Wu, Q. Chen, and M. Yachida, "Face detection from color images using a fuzzy pattern matching method," *IEEE Tran. on PAMI*, vol. 21, pp. 557–563, June 1999.

[3] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detction in color images," *IEEE Tran. on PAMI*, vol. 24, pp. 696–706, May 2002.

[4] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic face identification system using flexible appearance models," *Image and Vision Computing*, vol. 13, pp. 393–401, June 1995.

[5] C. Liu, "A bayesian discriminating features method for face detection," *IEEE Tran. on PAMI*, vol. 25, pp. 725–740, June 2003.

[6] R. Feraud, O. J. Bernier, J.-E. Vialett, and M. Collobert, "A fast and accurate face detector based on neural networks," *IEEE Tran. on PAMI*, vol. 23, pp. 42–52, January 2001.

[7] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *Computer Vision and Pattern Recognition*, pp. 130–137, 1997.

[8] H. Wang and S.-F. Chang, "A highly efficient system for automatic face region detection in mpeg video," *IEEE Tran. on Circuits and Systems for Video Technology*, vol. 7, pp. 615–628, August 1997.

[9] H. Greenspan, J. Goldberger, and I. Eshet, "Mixture model for face-color modeling and segmentation," *Pattern Recognition Letters*, vol. 22, pp. 1525–1536, December 2001.

[10] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detetion," *International Journal of Computer Vision*, vol. 46, pp. 81–96, January 2002.

[11] S. L. Phung, A. Bouzerdoum, and D. Chai, "Skin segmentation using color pixel classification: Analysis and comparison," *IEEE Tran. on PAMI*, vol. 27, pp. 148–154, Jan. 2005.

[12] C.-C. Lin and W.-C. Lin, "Extracting facial features by an inhibitory mechanism based on gradient distributions," *Pattern Recognition*, vol. 29, no. 12, pp. 2079–2101, 1996.