Proceedings of International Conference on Image
Processing and Character Recognition

Human Face Detection in Color Images[1]

Chichyang Chen[2] and Shiu-Ping Chiang
Department of Information Engineering, Feng Chia University
Taichung, Taiwan 407, ROC
Fax: (04)451-6101, email: cychen@fcusqnt.fcu.edu.tw

Abstract

Color characteristics are proposed to be utilized to detect human faces in color images. The proposed face detection method first uses a neural network to segment out the candidate face regions. Then, an energy thresholding method that can take the shape, the color, and the edge characteristics of the face features into the extraction process is devised to extract the lip. Finally, three shape descriptors of the lip feature are used to further verify the existence of the face in the candidate face regions. Our experimental results show that this method can detect faces in the images from different sources in an accurate and efficient manner.

I. Introduction

Video and image signals are very important key elements in a multimedia system. Ideally, the multimedia system should provide the users with the ability to query and edit video and image signals conveniently. To accomplish this goal, the video and image signals have to be manipulated according to their contents [1]. In [1], Smoliar and Zhang called this kind of video manipulation as content-based video indexing and retrieval. Before the video and image signals can be stored or retrieved according to their contents, the objects in the image frames of the video signal have to be recognized.

The recognition of human faces has been an intensive research topic recently [2][3]. According to the review by Samal and Iyengar in [2], there are basically two approaches to the problem of human face detection. Model based vision techniques that determine the faces as a whole unit is the first approach, for example, Hough transform [4], template matching[5]. In the second approach, the face is located by first locating some important features of the face. Eyes [6] and lips [5] are the most commonly used features. However, these approaches suffer either from their low performance because of their complicated algorithms or from the restricted applications due to the simple model adopted.

Human faces are key elements in the image frames of video signals in a multimedia system. The detection of human faces in the video is therefore an important step toward the goal of content-based video indexing and

[1] This work is supported by National Science Council, ROC, under grant no. NSC85-2213-E035-024.
[2] To whom all correspondences should be addressed.

retrieval. However, this step is a very difficult task. The difficulty comes mainly from the fact that there are a large volume of image frames in a common video signal. In addition, the background of the image is usually complex. Recently, a new method for face detection in complex background has been proposed [7]. In [7], Yang and Huang developed some simple rules based on the contrast between the face and background to locate the candidate areas of the faces. Although their rules are simpler than those techniques applied to detect face features, they are still time consuming. From their experimental results, it takes about 60 to 120 seconds on a SUN-4 station for locating the face in a picture of 512 $\times$ 512 pixels. Furthermore, the detection of faces is not very reliable, only 83 percent of faces are successfully detected.

In a multimedia system, the video and image signals are usually colored. In this research, we explore the techniques that can detect and locate human faces in the color image frames of video signals by using the color characteristics of the human faces. The number and the size of the faces may not be known a priori. For simplicity, we assume that the face is not occluded and the facial hair or the make-up and the accessory do not distort the face, and the face is a front face without distortion.

The color spectrums of the face skin and face features tend to be clustered in different connected regions in color coordinate space. By using this property, our proposed detection method is divided into three processing stages. In the first stage, the color image under test is segmented into skin and non-skin regions. Each connected region with skin color is then considered as a candidate face region. The second stage is to extract the face features from the candidate face regions by using their color characteristics. A method, called "energy thresholding method", is proposed to segment the lip in the candidate face region. By taking the color, shape, and edge characteristics of the lip feature together into the thresholding process, this method can accurately and efficiently extract the area of the lip. Finally, in the third stage, the lip area extracted from the second stage is further verified by using its three shape descriptors.

The organization of this paper is described as follows. In Section II, color classification technique that is used to segment the candidate face regions in the image is presented. Section III describes the energy thresholding method for extracting the face features from the candidate face regions. Section IV discusses how to identify and locate the face regions by using the shape of the extracted face features. Results of the experiments for testing the proposed face detection approach are shown in Section V. Finally, Section VI concludes this research.

II. Color classification

The goal of color classification is to classify the pixel colors in the image into skin color and non-skin color. A three-layer neural network is used to perform this classification. The training pixels with skin color are obtained from one hundred faces in different images. After training, the neural network acts as a two-dimensional classifier that classifies the color space into two different areas. In the following, color classification method is first explained, then segmentation and filtering techniques for obtaining the candidate face regions are described.

II.A Neural network for color classification

The architecture of the neural network used is a feed forward three-layer network. The input layer has two

nodes, and the output layer has only one node. The hidden layer has three nodes. According to [8], there must be at least three layers in the network in order to form a closed region in the space to be classified.

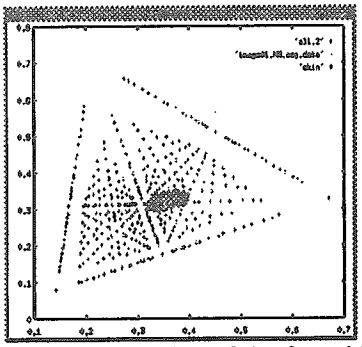
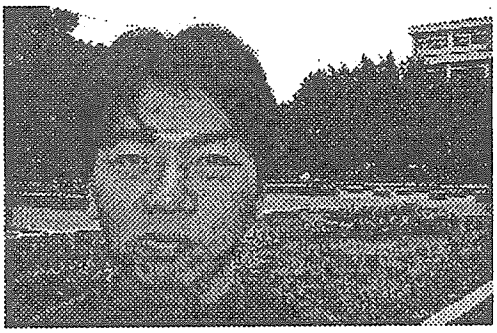
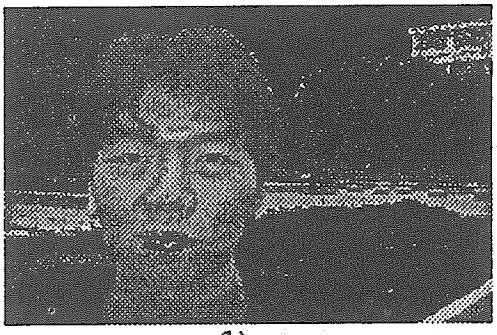


Figure 1: Color distribution of the face skin colors from neural network training.



(a)



(b)

Figure 2: (a) An image with one face. (b) The image after color classification.

The training method adopted for this network is supervised and back-error-propagation learning [8]. The training vectors used in the training process are obtained from the face regions in the training images. The RGB coordinates of the pixels in these face regions are first transformed into CIE-xyz coordinates. The training vectors include the CIE x and y coordinates only. After training, the region of the face-skin color in CIE-xy coordinate space is a nonlinear and connected region, which is shown in Fig. 1.

To test whether a pixel is a skin pixel or not, the xy coordinates of the pixel is sent to the input of the network. Then the output value of the network y is examined. If y is larger than or equal to 0.5, the pixel is considered a skin pixel. Otherwise, it is a non-skin pixel. An example image with a face is shown in Fig. 2(a). After color classification by using the network, the face region has been successfully extracted out, as shown in Fig. 2(b).

II.B Segmentation and Filtering

The primary goal of the processing in this stage is to segment the candidate face regions out from their background. These possible face regions should be connected regions without any holes inside. The holes inside these regions may be caused by noises, highlights, and the face features that have colors other than the face skin. The processing of this stage includes the following steps:

1. In the first step, median filtering is applied to the whole image. The color of each non-skin pixel is replaced by skin color if the number of the skin pixels in its 3×3 neighbor is larger than four. This median filtering removes noises and smoothes the image.
2. In the second step, region growing technique [9] is applied to segment the connected skin regions.
3. In the third step, the segmented regions with their size smaller than 5×5 are deleted. This size is determined empirically.

4. The objective of the fourth step is to include the holes inside the skin regions into the skin regions. In this step, a smallest rectangular area that can cover each of the segmented skin regions is determined first. Then, a test is applied to each non-skin pixel inside the rectangular area. If at least three of the four lines emanating from the pixel in four different directions (north, south, east, and west) intersect the skin pixels within the boundary of the corresponding covering rectangular, the pixel is considered as an inside pixel within the possible face region.

After the four steps of processing, the segmented skin regions are referred to as candidate face regions. These regions are connected without holes inside them and are represented in separate data structures.

III. Face feature extraction --- energy thresholding method

The features we are most interested are the lips since they are the most reliable features to be detected. The active contour model has been successfully applied to face feature extraction [10]. Other methods such as deformable templates have also been proposed to extract face features [6]. These methods use the gray-scale variation of images and can accurately extract the shape of the features. However, these methods are usually very time consuming. For the purpose of human face detection, only a rough shape of the features need to be extracted. The requirement of the extraction task is high speed.

The characteristics of the lips in the image, including their shape, edge, and color, are formulated into three different energy terms in our proposed method. A threshold value for the sum of the three energy terms is empirically determined. If the energy of a pixel in the candidate face region is smaller than this threshold, the pixel is classified to be a pixel within the lip. This method, called "energy thresholding method", can efficiently extract a rough contour of the lip that can be used for face detection. In the following, the procedure for extracting the face features is described, then the definitions of the three energy terms are explained.

III.A Feature extraction procedure

Our feature extraction method first looks for a small region that covers the possible feature area in the candidate face region. Since only a small region of the image is processed, the computation can be performed efficiently. We first classify the candidate face region into three areas: the skin area, the lip area, and the brows area, according to their color characteristics. Fig. 3 shows the result after the application of this classification method to a face region.

After color classification of the candidate face region, region growing technique is then applied to segment the lip region. The lip region is defined to be the largest connected region among the regions whose colors are classified as lip color. Once this lip region is found, an $l \times w$ rectangular that covers this region is defined. The center of the rectangular is defined to be the center of the segmented possible lip region. For convenience, this rectangular is referred to as "candidate lip rectangular".

Starting from the center of the candidate lip rectangular, a region growing technique using pixel aggregation [11] is applied to segment a more accurate lip region. If the value of the energy of a pixel is smaller than a predefined threshold W_T, the pixel is included. The energy W of a pixel within the candidate lip

rectangular consists of three different energy terms which are named as shape energy W_{shape}, color energy W_{color}, and edge energy W_{edge}. Fig. 4 shows the result after the application of this energy thresholding method to the candidate lip rectangular in Fig. 3.

Figure 3: A candidate face region after feature color classification.

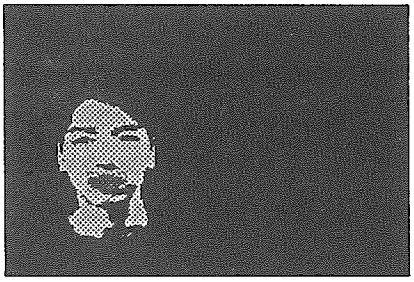


Figure 4: Lip feature segmented from the candidate lip rectangular in Fig. 3 using energy thresholding method.

III.B Shape energy term

The shape of lips is approximated by an ellipse with its major axis equal to the length of the lip and its minor axis equal to the width of the lip. The center of the ellipse is estimated to be located on the geometric center $(x_0, y_0)^T$ of the candidate lip rectangular. Since the estimated length and width of the lip are proportional to the length and width, l and w, of the candidate lip rectangular, respectively, the shape energy term is defined as

$$W_{shape} = w_s\left[\frac{(x-x_0)^2}{l^2} + \frac{(y-y_0)^2}{w^2}\right]$$

where $(x,y)^T$ is the coordinate of the pixel being tested and w_s is the weight of this energy term and its value is determined empirically.

The shape energy W_{shape} allows the prior knowledge of the shape of the lip be considered into the extraction process. As seen in Fig. 3, there usually exists a dark strip between the upper lip and the lower lip. The shape energy can adequately correct this kind of mistakes by taking the shape of the feature into consideration.

III.C Color energy term

Different faces have different colors in their lips. Therefore, a more accurate method for classifying each individual candidate lip rectangular is needed. Color classification of the candidate lip rectangular can be considered as a vector quantization problem. The code book of the quantizer has two code vectors: the lip color and the skin color. L*a*b* color coordinate system is used in this vector quantization problem since the L*a*b* coordinate system has the advantage of uniformity in measuring color differences [12]. Simple LBG algorithm [13] for vector quantizer design is used. The average skin color and the average lip color of the pixels in the candidate lip rectangular obtained from the color classification method described in Section III-A are considered as the initial code vectors for the LBG algorithm. Since there are only a small number of pixels within the candidate lip rectangular and the initial code vectors are very close to the final code vectors, the LBG algorithm can be performed very fast.

The color energy term W_{color} is defined to be

$$W_{color} = w_c\left[(L-L_0)^2 + (a-a_0)^2 + (b-b_0)^2\right]$$

where $(L, a, b)^T$ is the color of the pixel under test, $(L_0, a_0, b_0)^T$ is the lip color obtained from the LBG

algorithm, and w_c is the weight of this term whose value is determined empirically.

III.D Edge energy term

The edge energy W_{edge} is defined as

$$W_{edge} = w_e\left[(S_x - S_{0x})^2 + (S_y - S_{0y})^2\right],$$

where $(S_x, S_y)^T$ and $(S_{ox}, S_{oy})^T$ are the gradients of the pixel under test and the gradients of the center of the candidate lip rectangular, respectively, and w_e is the weight of this energy term whose value is determined empirically. The gradients of the pixel is computed by using Sobel edge detectors [12]. In order to obtain a robust value of the weight w_e that can apply to different images, the gradients used in the energy term are normalized by the maximum gradient values in the candidate lip rectangular.

IV. Detection of face regions

After the lip feature has been extracted, the shape of this feature is used to verify the existence of the face in the candidate face region. If the extracted feature area is recognized as a lip feature, the location of the face can be determined to be the smallest rectangular that covers the whole candidate face region. Three different shape descriptors are used for the recognition of the lips feature: the length/width ratio, the compactness, and the curvature.

The length/width ratio descriptor is defined to be the ratio of the length and width of the smallest rectangular that covers the whole extracted lip area. The compactness descriptor of the lips is defined to be the ratio of the area of the lip feature and the area of the smallest rectangular that covers the whole extracted feature. To compute the curvature descriptor of the lips, we first take sample points evenly along the boundary of the feature area. Then the angle at each sample point is computed as the angle between the point and its two nearest neighbors. The average value of all the angles at the sample points is defined to be the curvature of the lip. The verification values of these three shape descriptor are defined empirically to be, 0.25 to 0.9, from 0.5 to 0.8, and from 130 to 180 degrees, respectively.

V. Experimental results

450 color images are used to test the effectiveness of the proposed three-stage face detection method. These 450 color images are divided into three groups according to their sources. Each group has 150 images. Among each group, fifty of them contains no faces, while the other 100 images have one or more faces in the images. The source of the first group is a KODAK digital camera. The images of the first group are taken by this camera under different lighting conditions. The source of the second group is a color scanner. The images of the second group are obtained by scanning the pictures in magazines, newspapers, etc. The images of the third group are obtained from video signals. The video signals come either from a SONY V-8 camera or from the sites in the internet network.

The images in the three groups are tested separately. Fifty images of the one hundred images that contain faces in each group are used as the training images to obtained the color characteristics of the face skin and the lips. Then, the other one hundred images are used as the test images. Table I shows the percentages of the successful face detection from the fifty testing images that contain faces. The percentages of the incorrect face detection from the fifty testing images that contain no faces are all zero for the three groups of images. The size

of the test images is 378 × 252. The average and longest detection time are 3.8 and4.23 seconds, respectively on a Sun Spark-20 workstation.

Table I: Percentages of successful face detection in the images containing faces and the percentages of the mistaking face detection in the images containing no faces in the three groups of images.

Sources / Type of images	camera	scanner	video
with faces	96%	90%	76%
without faces	0%	0%	0%

VI. Conclusions

In this research, color characteristics are proposed to be utilized to detect human faces in color images. The proposed three stage method for face detection has been tested by experiments. The experimental results show that this method can detect faces in different images from different sources in an accurate and efficient manner. From the experimental results, it can also be concluded that this detection method is most useful when the images come from the same source or their sources are known beforehand. Since human faces are very common elements in the image and video signals of a multimedia system, the proposed face detection method should be useful toward the goal of content-based indexing and retrieval of video and image signals in multimedia systems.

VII. References

[1] S. Smoliar and HongLiang Zhang, "Content-based Video Indexing and Retrieval," *IEEE Multimedia Magazine*, pp. 62-72, 1994.

[2] A. Samal and P. A. Iyengar, "Automatic Recognition and Analysis of Human Faces and Facial Expressions: A survey," *Pattern Recognition*, vol. 25, no. 1, pp. 65-77, 1992.

[3] Dominique Valentin, Herve Abdi, Alice J. Otoole, and Garrison W. Cottrell, "Connectionist models of face processing: A survey," *Pattern Recognition*, vol. 27, no. 9, pp. 1209-1230, 1994.

[4] V. Govindaraju and S. N. Srihari, "A Computational Model for Face Location," in *Proc. 3rd Int. Conf. Comput. Vision*, pp. 718-721, 1990.

[5] I. Graw, H. Ellis, and J. R. Lishman, "Automatic Extraction of Face Features," *Pattern Recognition Letters*, no. 87, pp. 183-187, 1987.

[6] A. L. Yuile, D. S. Cohen, and P. W. Hallinan, "Feature Extraction from Faces Using Deformable Templates," in *Proc. CVPR*, pp. 104-109, 1989.

[7] G. Yang and T. S. Huang, "Human Face Detection in Complex Background," *Pattern Recognition*, vol. 27, no. 1, pp. 53-63, 1994.

[8] J. S. Judd, "Learning in networks is hard," in *Proc. 1st IEEE Int. Conf. Neural Networks*, San Diego, CA, June 1987, pp. 685-692.

[9] Ramesh Jain, Rangachar Kasturi, and Brian G. Schmunck, *Machine Vision*. Newyork: McGraw-Hill, 1995.

[10] Chung-Lin Huang and Ching-Wen Chen, "Human face feature extraction for face interpretation and recognition," Pattern Recognition, vol. 25, no. 12, pp. 1435-1444, 1992.

[11] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley, Reading, Mass., 1992.

[12] A. K. Jain, *Fundamentals of Digital Image processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[13] R. P. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine*, vol. 4. pp. 4-22, 1987.