A SIMPLE AND FAST COLOR-BASED HUMAN FACE DETECTION SCHEME FOR CONTENT-BASED INDEXING AND RETRIEVAL

Sangkeun Lee

Samsung DMS Lab.
Irvine, CA 92606, U.S.A
sangkny@yahoo.com

Monson H. Hayes

Georgia Institute of Technology
Atlanta, GA 30325, U.S.A
mhh3@ece.gatech.edu

ABSTRACT

The objective of this work is to provide a simple and efficient method to detect human faces in the compressed domain. We separate skin regions from non-skin regions using the Bayesian decision rule with a Gaussian mixture model (GMM). Next, we detect the location of a face using a deformable template that considers shape and orientation of the face. Then, face candidates are verified by using the second moments computed directly from the DCT coefficients in the face region. In particular, the coefficients are manipulated differently according to the orientation of the face. Good results have been obtained for a large variety of video sequences. This algorithm can be applied to JPEG images without any modification as well.

1. INTRODUCTION

The human face is an important subject in image and video databases, because it is a unique feature of human beings and is ubiquitous in photos, news, video, and documentaries. Faces can be used to index and search images and video, classify video scenes, and segment human objects from the background [1]. Face detection is performed to determine if there are any faces in an image and locate the position of each face [2].

Different approaches have been developed in recent years for the face detection problem. Some of the most representative works include shape-feature approaches [4]. Neural network approaches are used in [10], and template matching methods are proposed in [3]. These approaches have tended to focus still on gray-scale images. While they report good performance, they are often computationally expensive. This is especially true of neural network approaches, since they require processing for each possible position and scaling of the image. Another method for detecting faces is to use color information. Skin-color based face detection approaches have several advantages over other methods since under constant lighting conditions, color is almost invariant against changes in size, orientation, and partial occlusion of the face. Moreover, the processing of color information has proven to be much faster than the processing of other facial features which is an important point when dealing with video sequences [1].

We developed and implemented a color-based system for fast detection of human faces in images and video sequences where the faces appear in the compressed domain. First, skin regions are separated from non-skin regions using a *Bayesian decision rule* with a GMM. After that, the human faces are located within the skin regions through the template matching and the observation of the directional characteristics in the DCT coefficients of each region. We only analyze I-frames from MPEG streams in order to avoid costly decompression of the other frame types.

2. DETECTION OF SKIN REGIONS

The first stage of the proposed scheme checks each block of the video frame to see if there exist potential face areas using skin-tone color statistics. In order to generate skin-tone color statistics in the chrominance C_bC_r plane which is naturally related to the compressed format, we need a reliable skin color model that is adaptable to people of different skin colors and to different lighting conditions. In this paper, a total of 315 skin sample patches are used to determine the color distribution of human skin in chromatic color space using a Gaussian mixture model. We select M=2 and 3 for the mixture number of skin and non-skin tone samples, respectively. Similarly, non-skin patches are fitted into another GMM as well. In particular, another 300 patches for non-skin model were obtained to overcome a drawback of the color-based skin detection scheme from the regions which had a similar skin color but were not actually skin regions. Then, we apply *a Bayesian decision rule for minimum cost* [1] to MPEG video streams and classify each MPEG block as a candidate skin or non-skin one. Only the DCT DC-coefficients of the corresponding Y, C_b, and C_r blocks, which are equivalent to the average values of the blocks in the spatial (pixel) domain, are used. A block is classified as a skin block if its chrominance values fall within the small region, and its luminance value is within the interval $40 \leq \mathrm{Y} \leq 240$ [7]. After classifying the block, a binary map image is generated for each I-frame of each video, where a "one" indicates a skin region, and a "zero" indicates a non-skin one. The binary map image is post-processed by morphological operations (erosion and dilation executively) to eliminate noise, smooth the boundaries, and fill in holes in the image. The most important advantage of this simple algorithm is that the skin detection can be implemented with a look up table (LUT) which makes the algorithm extremely fast.

Most of the faces that researchers have considered [1, 10, 7] are oriented vertically. However, some of the faces have a considerable orientation. In order to have a higher performance, this feature should be considered for face detection. Once the system has determined that skin regions exit, it proceeds to analyze some characteristics in the particular regions. In this paper, the shape of a potential face region is approximated by an ellipse. To study each region in the binary map image, some features, which are a center, an area, a height, a width, and an angle (θ) of each region, are calculated [9]. The constraints of the region size and the ratio of the height to the width of the region are applied to remove

0-7803-8554-3/04/$20.00 ©2004 IEEE.

undesired segments which may correspond to skin-tone objects or background, and to save computing time.

3. SPATIAL DOMAIN FACE DETECTION

This section shows how to do the matching between the part of the image corresponding to the skin region and the template face. A human frontal face template is used to make a decision in determining if a skin region represents a face or not. This template was chosen by averaging 20 frontal-view faces of males and females wearing no glasses and having no facial hair. This template was sub-sampled by block (8×8 pixels) to be consistent with the region in a luminance DC-image [5], and filtered with low pass filter. Human faces (frontal faces and side faces) with orientation are detected using the template matching in the spatial domain and moments in the compressed domain. An example for the matching procedure is shown in Figure 1. In the figure, the template face is

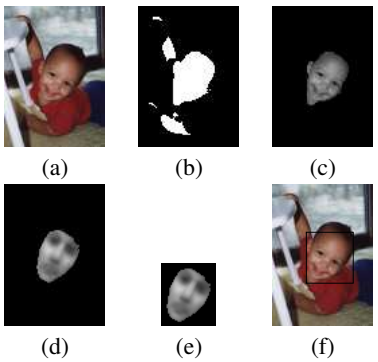

Fig. 1. Template matching procedure; (a) Original DC-image, (b) Segmented image of (a), (c) Gray-scale region of (b), (d) Resized and rotated template image, (e) Bounded template face, and (f) Matching result.

resized, rotated, and positioned to the same coordinates as the skin region. Specifically, the front face model is resized according to the height and width of the region computed, and rotated according to the target orientation. Now, the template face is aligned to the same direction as the skin region (d). Then, we select the same size boundaries (e) from both the target image(c) and the model image (d) to reduce the computation. Finally, the cross-correlation value is computed between the part of the image corresponding to the skin region and the template face, and the face region is indicated by the rectangular box (f). We empirically determined from our prototype set that a good threshold value of the cross-correlation for classifying a region as a frontal face is 0.8, and a value for ignoring a potential region is 0.2. These values have 97% accuracy in the prototype set which consists of 78 frontal faces with different sizes in 30 color images.

It is important to note that we need to determine two threshold values (lower and upper values) because the purpose of our system is to find all the faces whether they are front or side views, but only the frontal template is used for our template matching process, and this matching process gives corresponding outputs only to frontal faces. Therefore, the lower value is used to discard a potential face region and the upper value classifies the region as a face in the template matching process. However, the regions that score between the lower and the upper values go to the next stage to verify if they are face regions in the compressed domain.

4. COMPRESSED DOMAIN FACE DETECTION

The main purpose of this stage is to verify the human faces based on the previous template matching results. One of the most important characteristics of our method is that we use moments directly computed from the DCT coefficients of the Y-component to determine the face regions according to their orientation, and these moments are manipulated by the orientation of a potential face region. Direct computation of moments in the DCT domain was proposed in [8]. The first moment (mean) and the second moment (variance) are defined, respectively, in a luminance block with size 8×8 as

$$\mu_1 = \frac{1}{8} \cdot \mathrm{DCT}(0,0), \tag{1}$$

$$\mu_2 = \frac{1}{64} \cdot \sum_{u,v=0}^{7} \mathrm{DCT}(u,v)^2 \text{ for } u \neq 0 \text{ and } v \neq 0 \tag{2}$$

where, DCT(u,v) is the DCT coefficient at (u,v) of a block in MPEG. The concise expressions (1) and (2) show that the mean is directly derived from the DC coefficient, while the variance is just the average of squared AC coefficients. In order to use these moments as a measure for detecting a face region, we need to partition the DCT coefficients corresponding to the directional features [1, 11]. There are more discontinuities of intensity levels in the vertical direction compared to the horizontal direction of the image in face regions because of the existence of the eyes, the nose-mouth junction, and the lips in the face regions. These discontinuities are indicated by the DCT coefficients in the frequency band. The grouping of the DCT coefficients, where groups H, V, and D are sensitive to vertical, horizontal, and diagonal edges, respectively, is shown in Figure 2. In these masks, a "1" indicates the

$$\mathbf{H} = \begin{bmatrix} 0&0&1&1&1&1&1&1 \\ 0&0&1&1&1&1&1&1 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 0&0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0 \\ 0&1&1&1&0&0&0&0 \\ 0&0&1&1&1&0&0&0 \\ 0&0&0&1&1&1&0&0 \\ 0&0&0&0&1&1&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \end{bmatrix}$$

Fig. 2. The partitioning of the DCT coefficients.

value of the DCT coefficients at that position, and a "0" means the opposite. Given a potential $m \times n$ face region B and its area A in the binary image, the second moments based on the direction feature are defined at the corresponding DCT blocks in the luminance image as

$$\mu_{2H} = \frac{1}{64} \cdot \frac{1}{A} \cdot \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \sum_{u,v=2}^{7} \mathrm{B}(i,j) \cdot \mathrm{DCT}_{ij}(u,v)^2 \cdot \mathrm{H}(u,v),$$

$$\mu_{2V} = \frac{1}{64} \cdot \frac{1}{A} \cdot \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \sum_{u,v=2}^{7} \mathrm{B}(i,j) \cdot \mathrm{DCT}_{ij}(u,v)^2 \cdot \mathrm{V}(u,v)$$

where, $\mathrm{DCT}_{ij}(u,v)$ is the DCT coefficients of 8×8 block corresponding to the pixel (i,j) of the potential face region B. The ratio of the vertical moment to the horizontal moment can be used to

detect the region that has more discontinuities in the vertical direction. However, these masks can be used with the assumption that the potential face regions are vertically positioned, or only tilted a little. In order to select the AC coefficients directly in the DCT domain according to its orientation of a potential face, we separate the DCT coefficients into four regions as shown in Figure 3. In this figure, only horizontal edges (vertical discontinuities) are

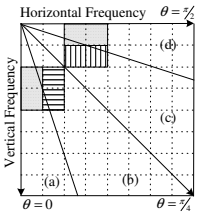

Fig. 3. Separated regions in the DCT domain.

considered to explain the procedure of selecting AC coefficients. As a horizontal edge ($\theta = 0$) rotates to become a vertical edge, the corresponding features are indicated by the regions (a) through (d). Similarly, this observation can be applied to vertical edges. We expect that a tilted human face can be compensated and dealt with as an upright face by selecting appropriate AC coefficients according to its orientation.

For verifying a human face in the DCT domain, the required number of DCT coefficients is equal to the original image size which is 64 times bigger than the DC-image size we have dealt with. However, to reduce the number of computations and save the decoding time for each coefficient, we use only eight AC coefficients in a block, which are illustrated as solid and striped boxes in Figure 3. In [6], edge information in the DCT domain is successfully extracted by using five AC coefficients of each block in zigzag order, and is used to identify shot change points. In particular, we use two sets of masks to compute the second moments based on the angle of a potential face as illustrated in Figure 4.

$$\underset{\mathbf{H_1}}{\begin{bmatrix}0&0&1&1\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}}\quad\underset{\mathbf{V_1}}{\begin{bmatrix}0&0&0&0\\0&0&0&0\\1&0&0&0\\1&0&0&0\end{bmatrix}}\quad\underset{\mathbf{H_2}}{\begin{bmatrix}0&0&1&1\\0&0&1&1\\0&0&0&0\\0&0&0&0\end{bmatrix}}\quad\underset{\mathbf{V_2}}{\begin{bmatrix}0&0&0&0\\0&0&0&0\\1&1&0&0\\1&1&0&0\end{bmatrix}}$$

Fig. 4. Two pairs of partitioning AC coefficients.

Each pair of masks is selected according to its orientation of a potential face by

$$[\mathrm{H}_\xi\,,\mathrm{V}_\xi]^T = \begin{cases} \xi = 1 \;\text{ if } 0 \le |\theta| < \theta_1\,, \\ \xi = 2 \;\; \theta_1 \le |\theta| < \pi/4\,. \end{cases} \tag{3}$$

When an angle $|\theta|$ is larger than $\pi/4$, the angle is updated by $\theta = \pi/2 - |\theta|$, and Eq. (3) is applied again. Then the horizontal mask H_ξ and the vertical mask V_ξ are exchanged based on the symmetry property. These operations can reduce the computation complexity by almost 3 or 6 times. Moreover, the selected masks can save partial decoding time since several numbers of the corresponding AC coefficients are necessary instead of all the AC coefficients. In order to get an AC coefficient in the DCT domain, partial decoding procedures including Huffman decoding and inverse quantization are required.

5. EXPERIMENTAL RESULTS

To evaluate the performance of the face detection system, we focus on the overall speed and accuracy of the results. The algorithm executes in real-time for locating the human faces. In particular, to show that our proposed algorithm performs competitively against existing algorithms, the face detection algorithms developed by Rowley*et al.* [10] at Carnegie Mellon University and Wang *et al.* [1] were selected for comparison in processing accuracy and speed for detecting faces. The proposed algorithm has been evaluated using the first test set in Table 1. This test data set

Table 1. Characteristics of the faces in the test data set.

Type	Number
Frontal	58
Semifrontal	29
Side	12
Little tilted (±15°)	87
Much tilted (±90°)	12
Total faces	99

contains 50 images that have been extracted as key frames from the I-frames of various MPEG videos. The size of each frame is 352×240 pixels. The sequences consist of a commercial film (CF), news, and movies. This set of 50 images contains 99 faces, which are again classified into two categories according to their orientations, and 5 images which do not contain faces. They cover most of the cases that the algorithm has to deal with. In Table 1, we give a detailed description of the content of this set according to facial characteristics. In order to evaluate the proposed human face detection algorithm using a template and moments in the compressed domain, we implemented the algorithm in C++ and performed the experiments using a Pentium II 400 MHz PC, using the Windows operating system.

In this experiment, the test set includes color images with multiple faces of different sizes, different colors, different positions, and a frame which does not contain any faces. Our first test data set of 50 images was tested using the interactive demonstration of the CMU face detector at
http://www.vasc.ri.cmu.edu/cgi-bin/demos/findface.cgi that allows users to submit an image for processing in batch mode and to retrieve the resulting image with bounding boxes overlaid on the detected faces. The results of the proposed algorithm compared to

Table 2. Face detection results.

Results	The proposed method	CMU face detector	Wang's face detector
Correct	94 (94.95%)	92 (92.93%)	87 (87.88%)
False alarms	7 (7.07%)	12 (12.12%)	5 (5.51%)
False dismissals	5 (5.05%)	7 (7.07%)	12 (12.12%)

the CMU face detector and Wang's face detector on our test data set are illustrated in Table 2.

It is important to note that the minimum sizes for the compared face detectors are 20×20, 48×48, and 24×24, respectively, for the CMU, Wang's, and our system. The CMU system and our scheme found several small faces, whereas Wang's face detector could not find them. A main difference between the compared systems is whether they use color information or not. The

CMU face detector does not use color information. Only the geometrical intensity features of a face are used in the pixel domain. Therefore, a set of candidate face areas cannot be built, and the neural network filters have to be applied at every pixel location in each image of the multiscale pyramid. The CMU scheme is still more time consuming than Wang's and our own. Moreover, the CMU face detector may find false alarms in complex backgrounds which have non-skin color areas. The proposed system does not have this drawback. On the other hand, Wang's scheme and our scheme fail in the case of extreme lighting conditions causing bad skin color segmentation.

As shown in Table 2, our algorithm detects 94 of the 99 faces which means a successful detection rate of 94.95%, whereas the CMU system and Wang's system detect 92 faces and 87 faces, respectively, leading to successful detection rates of 92.93% and 87.88%. We observed that the CMU face detector and Wang's face detector are more sensitive to the orientation of a face than the proposed method, especially for faces with a lot of tilt. The main reason may be that Wang's system did not take the orientation of the face into account, assuming faces that are upright or have little tilt, and the CMU face detector did not consider this situation during the neural network filters. Fewer false alarms are obtained using Wang's face detector; however, Wang's scheme provides many false dismissals (12.12%) mainly in small faces and faces with a lot of tilt. Seven false alarms and five false dismissals (5.09%) are obtained with our method, whereas 12 false alarms and seven false dismissals (7.07%) are obtained using the CMU face detector. After our chrominance segmentation step, a number of potential false alarms appear, but our shape constraints reduce them in a very efficient way. Especially, the selective moment calculations using intensity face texture directly in the compressed domain is effectively used when the result of template matching is in some range.

For an overall system speed comparison, a fast system by Wang *et al.* [1] is implemented as a baseline system. We used 100 I-frames of a news sequence, whose frame size was 352×240, as the second test set. The main differences between Wang's system and our system are the processing image size and the mask size of DCT coefficients. Wang's system is based on the macro-block level image size whereas our system is based on the block level image size. To verify a face in the compressed domain, Wang's detector needs to decode all the coefficients, but our scheme selects and decodes only several coefficients if necessary. Our method decodes some DCT coefficients selectively according to the results of the template matching. Even if Wang's system is much faster in the skin-tone segmentation stage, it requires a large number of computations to verify a face in the DCT domain. The elapsed time of the two systems depends on the output of the skin-tone detection stage. After segmentation, if only a few potential regions are shown in the binary image, then little template matching is involved, so less time is needed. On the other hand, if the video sequence contains many skin color regions and is complex, the template matching step will take longer.

The overall processing time comparison is shown in Table 3 for the test sequence. The average run time of Wang's system (31.3 msec.) is about 1.53 times faster than the proposed system (47.9 msec.). It is worth noting that Wang's face detector has many false dismissals, but our method detects them because of our minimum face size limitation. In this case, Wang's detector does not have to go to the next verifying stage. When the group of pictures (GOP) in MPEG format is taken into account, the GOP has one I-frame and consists of almost 15 frames. It is enough to provide real-time processing only if a system performs an algorithm during decoding two GOPs per second. Therefore, our system is quite fast and performs in real-time under our experimental environment.

Table 3. Overall speed comparison between the proposed system and Wang's system.

Method	The Wang's system	The proposed system
Average elapsed time	31.3 msec.	47.9 msec. (1.53)

6. CONCLUSIONS

This paper proposes a simple and fast processing algorithm for detecting human faces in the compressed domain.

Experimental results show a high detection rate (94.95%) regardless of the number, size, and orientation of the faces. The algorithm executes in real-time. We believe that the proposed system can be used as a significant tool for video analysis.

7. REFERENCES

[1] H. Wang and S. F. Chang, "A highly efficient system for automatic face region detection in MPEG video," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 7, no. 4, pp. 615-628, 1997.

[2] M. -H. Yang and D. J. Kriegman, "Detecting faces in images: A survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34-58, 2002.

[3] B. Moghaddam and A. Pentland, "Probabilistic visula learning for object recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696-710, 1997.

[4] K. C. Yow and R. Cipolla, "Feature-based human face detectin," *Image and Vision Computing*, vol. 15, no. 9, pp. 713-735, 1997.

[5] B. L. Yeo and B. Liu, "Rapid scene analysis on compressed video," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 5, No. 6, pp. 533-544, Dec. 1995.

[6] S. W. Lee, Y. M. Kim, and S. W. Choi, "Fast scene change detection using direct feature extraction from MPEG compressed videos," *IEEE Trans. on Multimedia*, vol. 2, no. 4, pp. 240-253, Dec. 2000.

[7] C. Garcia and G. Tziritas, "Face detection Using quantized skin color regions merging and wavelet packet analysis," *IEEE Trans. on Multimedia*, vol. 1, no. 3, pp. 264-277, 1999.

[8] G. Feng and J. Jiang, "JPEG image retrieval based on features from DCT domain," *Int'l Conf. Image and Video Retrieval*, LNCS 2383, pp. 120-128, 2002.

[9] R. M. Haralick and L. G. Shapiro, *Computer and robert vision*: Vol. I, Appendix A, Addison-Wesley, 1992.

[10] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," in *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23-28, 1998.

[11] Y. S. Ho and A. Gersho, "Classified transform coding of images using vector quantization," in *IEEE Int'l Conf. ASSP*, pp. 1890-1893, 1989.