EFFICIENT FACE DETECTION FOR MULTIMEDIA APPLICATIONS

Nicolas Tsapatsoulis, Yannis Avrithis and Stefanos Kollias

Department of Electrical and Computer Engineering
National Technical University of Athens
Heroon Polytechniou 9, 157 73 Zographou, Greece
e-mail: {ntsap,iavr}@image.ntua.gr

ABSTRACT

Face detection is becoming an important tool in the framework of many multimedia applications. Several face detection algorithms based on skin color characteristics have recently appeared in the literature. Most of them face generalization problems due to the skin color model they use. In this work we present a study which attempts to minimize the generalization problem by combining the M-RSST color segmentation algorithm with a Gaussian model of the skin color distribution and global shape features. Moreover by associating the resultant segments with a face probability we can index and retrieve facial images from multimedia databases.

1. INTRODUCTION

In the past the term *face detection* was strongly related with the face recognition task; this fact had a deep impact on the developed algorithms. In order to achieve the required accuracy of detection, rigorous constraints had been posed in the digital image environment [9]. Moreover the great majority of the corresponding algorithms were based on gray-scale images utilizing template matching, image invariants or low level features for the detection of local facial features like eyes, nose and mouth [10][12][13].

Recently, the rapid development of multimedia applications added importance to the face detection task and decoupled it from face recognition. For applications like image and video indexing and retrieval, video scene classification and video / news summarization face detection has become a valuable tool. Modern applications, however, require fast implementations with sufficient accuracy rather than exhaustive procedures providing higher precision. As a result, well established algorithms which had been successfully utilized for face recognition are not appropriate or require re-development. Moreover, multimedia applications tend to employ color characteristics in contrast to traditional face detection.

The work presented in [11] inspired many researchers for implementing color based face detection algorithms. The basic idea of [11] is the modeling of skin color using the chrominance components of the *YCrCb* color model. Most of the studies based on this idea reveal that considerable effort is required in post processing steps to achieve remarkable results [5]. It is not clear however whether the post processing steps are sufficient for face detection based on skin color characteristics. Although the skin color subspace covers indeed a small area of the *Cr-Cb* chrominance plane, it cannot be modeled in such a general way to be efficient for all images that include faces. To improve the generalization ability the skin color model should be "relaxed" leading to an increased amount of *false alarms*. On the other hand a "rigorous" model increases the number of *dismissals.* Moreover the influence of the luminance channel *Y* is not totally negligible.

In this study we approach the face detection in a different way: We still use the skin color characteristics but in a second stage. In the first stage the M-RSST general purpose color segmentation algorithm [1] is applied to the image / frame under examination. Each segment is then associated with a skin color probability, while shape constraints are taken into account in a third stage in order to estimate an overall *face probability*.

This scheme improves the efficiency of face detection in three main ways: First, segments rather than macroblocks or pixels are associated with a skin color probability. By using pixel or macroblock probability some kind of threshold should be applied in order to merge pixels / macroblocks for creating candidate skin regions; in contrast, thresholding can be left out in the last stage of the proposed algorithm. Actually segment probabilities can serve as outputs of the system and utilized for facial indexing and retrieval purposes [2]. Moreover, since every segment retains its probability, we can extend the generalization performance of the algorithm by using adaptive skin color models [2]. Face segments with low probabilities due to the particular image environment can be reassessed after adaptation. Finally, segments are much less sensitive to lighting conditions than pixels or macroblocks. This is of particular interest since face shape is more efficiently captured and can be reliably used in the last stage of the algorithm to move from skin color probability to face probability.

2. COLOR SEGMENTATION

The Multiresolution Recursive Shortest Spanning Tree (M-RSST) algorithm, first introduced in [1], is our basis for color segmentation and briefly described in the sequel. The M-RSST is an efficient, multiresolution implementation of the conventional RSST [7] algorithm. It is approximately 400 times faster than the conventional RSST for typical image sizes and can be employed for direct segmentation of MPEG video streams with minimal decoding.

Initially a multiresolution decomposition of an input image I is performed with a lowest resolution level of L_0 so that a hierarchy of frames $I(0)=I$, $I(1),\ldots,I(L_0)$ is constructed, forming a truncated image pyramid, with each layer having a quarter of the pixels of the layer below. The RSST initialization takes place for the lowest resolution image $I(L_0)$ by creating a partition of regions (segments) of size 1 pixel each and generating links all 4-connected region pairs. Each link is assigned a weight equal to the distance between the two respective regions, which is in general defined as the Euclidean distance between the average color components of the two regions, using a bias for merging small regions. Using the *YCrCb* color space for example, a distance measure between two adjacent regions X and Y is defined as

$$d(X,Y) = \| \mathbf{c}_X - \mathbf{c}_Y \| \frac{a_X a_Y}{a_X + a_Y} \qquad (1)$$

where $\mathbf{c}_X = [\ Y_X,\ Cr_X,\ Cb_X\ ]^T$ contains the average color components of region X and a_X is its area, i.e. the number of pixels within the region.

Then an iteration begins, involving the following steps: *(i)* region pairs are recursively merged in ascending order of the corresponding link weights, using the conventional RSST iteration phase, *(ii)* each boundary pixel of all resulting regions is split into four new regions, whose color components are obtained from the image of the next higher resolution level, *(iii)* the new link weights are calculated and sorted. This "split-merge" procedure is repeated until the highest resolution image $I(0)$ and a minimum distance threshold are reached.

Due to the definition of the distance measure $d(X,Y)$ and to the low resolution of the initial image $I(L_0)$, small segments corresponding to facial details are in most cases eliminated and a single segment is retained for the whole facial area; for this reason, the M-RSST can be successfully employed for color based face detection [2]. As demonstrated in the experimental results, however, there still exist several cases – especially for large face segments – where even an optimal selection of the distance threshold cannot yield a single segment for the facial area without merging this segment with neighboring image areas. For this reason, a second step of segment merging is applied, based on skin-tone color distribution.

3. SKIN –TONE COLOR DISTRIBUTION MODEL

It is stated in some classic studies [6][8] that skin-tone colors are spread over a small area of the *Cr-Cb* chrominance plane of the *YCrCb* color model. This fact has been successfully used in some recent studies for face detection in color images and video sequences [5][11]. In our work, we approximated skin-tone color distribution using a two-dimensional Gaussian density function. Assuming that the mean vector $\boldsymbol{\mu}_0$ and the covariance matrix $\mathbf{C}$ are robustly estimated, the likelihood of an input pattern $\mathbf{x}$ is given by:

$$P(\mathbf{x} \mid \boldsymbol{\mu}_0, \mathbf{C}) = \frac{\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T \mathbf{C}^{-1} (\mathbf{x}-\boldsymbol{\mu}_0)\}}{(2\pi)^{\frac{k}{2}} \cdot |\mathbf{C}|^{\frac{1}{2}}} \qquad (2)$$

where $k = 2$ is the number of chrominance components. For the estimation of $\boldsymbol{\mu}_0$ and $\mathbf{C}$ we used as training data facial pixels of different races, obtained from regular TV clips, color images and personal video cameras.

Color segments represented by their average chrominance components can be assessed using the above model to give a probability of being skin segments. Since color probability can be reliably estimated, it can be further exploited for segmentation. In particular, a small distance threshold is chosen for termination of the segmentation procedure, so that each face may yield more than one segments – typically two to six segments – but none of them is merged with segments of other objects or the image background; see Figure 1(c) for instance. A second *skin-color merging* iteration is then applied, using the distance measure

$$d_C(X,Y) = [\max(1-p_X, 1-p_Y)]^2 \qquad (3)$$

for link weights instead of $d(X,Y)$ given in (1), where p_X , p_Y are the *skin-color probabilities* associated to segments X and Y, respectively. Thus, segment merging does not rely on color difference or size, but solely on the probability of belonging to a face area. Consequently all adjacent face segments are merged into a single segment while the remaining image partition map is not affected.

Although the proposed Gaussian model is rather efficient for the classification of segments as skin and non-skin ones, it is much more valuable if it is used in an adapted form. In an interactive content-based retrieval environment, for instance, face segments selected by the user can be exploited for the re-estimation of the model's parameters [2].

4. SHAPE FEATURES

The above approach leads to accurate extraction of an arbitrary number of candidate face segments from an input image. The segmentation map along with the associated probabilities given in (2) can thus be directly used for face detection. However, since objects unrelated to human faces but whose color chrominance components are still similar to those of the skin might be present in an image, object contour shape is also taken into account. Image segments whose chrominance components match the adapted probabilistic model are assessed against their shape to verify that they correspond to human faces.

Ideally, the similarity of object shape to an ellipsis (or a more accurate face shape template) can be calculated, since face shape is supposed to be elliptical. Shape matching under arbitrary deformations can be achieved using, for example, active contour models (snakes) or deformable templates [4]. Moreover, rigid motion or affine transformations such as translation, rotation, scaling and skew can be efficiently removed using affine invariants or curve normalization [3].

In most realistic cases, however, the object contours obtained through color segmentation are far from the ideal. Even if contour curves are simplified or smoothed, using splines for instance, shape matching usually gives poor results. For this reason, only global shape features are taken into account. In particular, the *compactness* of each shape is first obtained using the perimeter and the area of the corresponding segment:

$$g_X = \frac{1}{4\pi} \frac{r_X^2}{a_X} \qquad (4)$$

where r_X denotes the perimeter (number of contour points) and a_X the area (total number of pixels) of segment X. Note that the maximal compactness is achieved for a circular shape and is equal to one, hence g_X as defined above is always normalized in the interval [0,1]. The shape *elongation* is then obtained through its Hotelling, or discrete Karhunen-Loeve transformation. Let the $N\times1$ vectors $\mathbf{x}$ and $\mathbf{y}$ denote the coordinates of N contour points representing the closed shape boundary of segment X. The 2×2 covariance matrix of these points with respect to their center-mass point ($\mu_\mathbf{x}$, $\mu_\mathbf{y}$) is given by

$$\mathbf{C} = \frac{1}{N}[\mathbf{x}-\mu_x \quad \mathbf{y}-\mu_y]^T[\mathbf{x}-\mu_x \quad \mathbf{y}-\mu_y] \tag{5}$$

The two eigenvectors of this covariance matrix express the principal – minor and major – axes of the shape, while the ratio of the its eigenvalues defines the elongation of the shape of object X:

$$\ell_X = \sqrt{\lambda_2/\lambda_1} \tag{6}$$

where λ_1, λ_2 are maximum and minimum eigenvalues, respectively. The above global shape features are fairly robust to segmentation noise and normalized in the interval [0,1]. They are also invariant to translation, scaling and rotation.

Experimental results have shown that typical values corresponding to face segments range from 0.44 to 0.79 for shape compactness and from 0.59 to 0.91 for elongation. Consequently, shape matching is achieved by transforming compactness and elongation with appropriate non-linear functions taking values in the range [0,1], similarly to fuzzy membership functions. Finally, the transformed compactness g'_X and elongation ℓ'_X are combined with the skin-color probability p_X using a weighted geometric mean, and an overall *face probability* is obtained, denoted as f_X. Since p_X is usually more reliable for face detection – skin color is far more characteristic for a human face than its shape – it assigned a higher weight, except if g'_X and ℓ'_X take values close to zero. This means that shape features are in effect only used to discard face segments that possess extremely irregular shape although they match the skin-color probabilistic model.

5. EXPERIMENTAL RESULTS

The visual content used in our experiments included a database created in the framework of project PHYSTA of the Training Mobility and Research Program of the European Community. From this database, 200 images have been selected from various video-clips recorded from BBC's broadcasted program. In addition, another set of 200 images has been used, including images from several shots of TV news recordings and commercials. In total, 321 out of the 400 images contained at least one face.

Some preliminary results are presented in the sequel. First, color segmentation is considered in Figure 1. It can be seen that the two face segments of the original image are accurately extracted using the proposed M-RSST segmentation and skin-color merging procedures. This cannot be achieved through segmentation only, since the largest face segment gets over-segmented as shown in Figure 1(c) – in fact Figure 1(b) was obtained from Figure 1(c) through skin-color merging. An attempt to increase the distance threshold results in the right part of the speaker face being merged with her body, as in shown Figure 1(d).

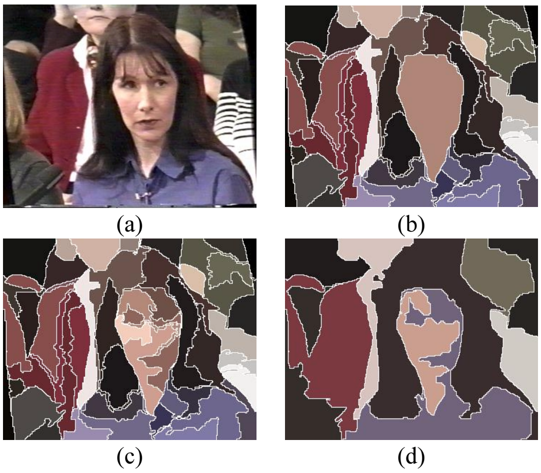

Figure 1: Color segmentation. (a) Original image, (b) segmentation with skin-color merging, (c) and (d) segmentation without skin-color merging & distance threshold equal to 2 and 22, respectively.

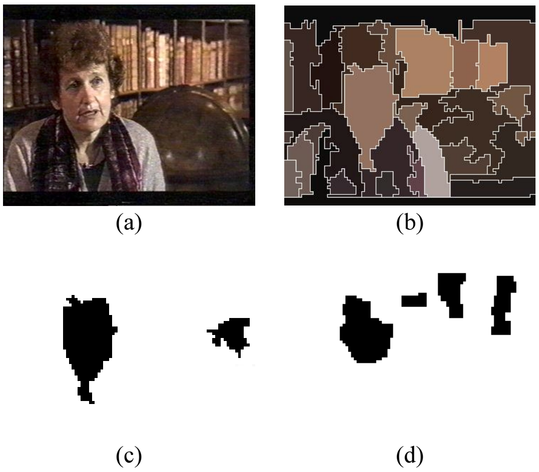

Figure 2: The proposed method vs. skin-color thresholding. (a) Original image, (b) segmentation at resolution level 2 (4×4 blocks), (c) thresholding on face probability f_X and (d) direct image block thresholding (without segmentation).

As discussed in the Introduction, face detection can be achieved by direct skin-color probability thresholding on image blocks (or macroblocks from encoded video streams), combined with suitable morphological operations on the resulting binary image masks. This approach is compared to the proposed one in Figure 2. Segmentation and skin-color merging are performed with a target resolution level equal to two, corresponding to 4×4 image blocks, as depicted in Figure 2(b). Shape features are calculated and the overall face probability f_X for all segments is obtained. Optimal thresholding of this face probability produces the image mask of Figure 2(c), while direct skin-color probability thresholding on 4×4 image blocks generates the mask of Figure 2(d). It is observed that although the face segment is distorted in the latter case, unrelated objects such as parts of the bookshelf cannot be avoided.

Finally, the resulting face probability map is demonstrated in Figure 3 for a variety of input image resolution, quality, contrast, number of face segments and lighting conditions. In all cases, the face probability f_X shown in column (c) outperforms the skin-color probability p_X shown in column (b) of Figure 3 as far as face detection is concerned. This is expected since face segments that possess extremely irregular shape are discarded even if they match the skin-color probabilistic model.

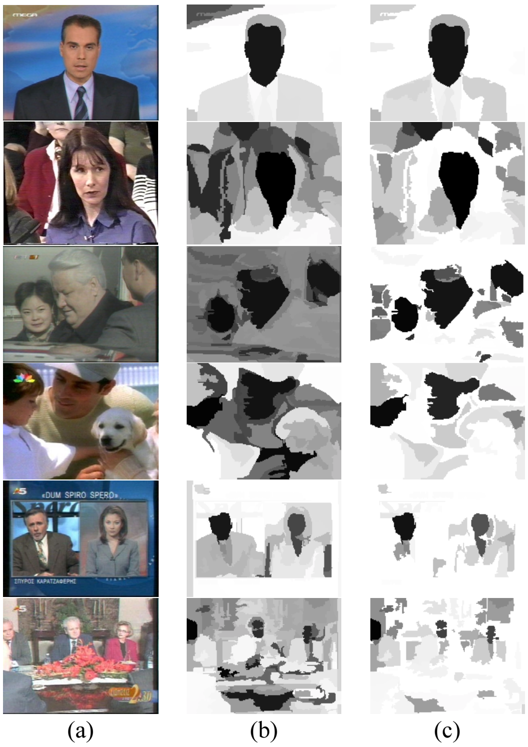

Figure 3: Face detection in a variety of situations. (a) Original images, (b) skin-color probability map, (c) final face probability map (including shape features).

Faces detected from the face probability map range from ideal to barely distinguishable, depending on input image quality and complexity. Although face shape may be distorted in some cases, all faces can be detected as distinct image segments with minimal false alarms. Moreover, probability maps contain more information than thresholded binary image masks and can be employed for constructing alternative image representations.

6. CONCLUSION

The proposed technique provides with a fast and efficient means for detecting human faces in image and video databases and can be used for a variety of multimedia applications, such video partitioning, browsing and summarization, indexing of TV news, interactive content-based retrieval and multimedia database management. It involves simple color and shape features and can be directly applied to compressed video sequences. Its performance is superior to direct skin-color probability thresholding in terms of face shape accuracy and average number of false alarms or dismissals. Moreover, the generated face probability maps contain more information than thresholded binary image masks and can be employed for constructing alternative image representations for enhanced visual content description, such as image partition graphs. Finally, integration of skin color features in the segmentation process results in enhanced image partitioning. An overall experimental evaluation suggests that proposed method successfully tackles the trade-off between speed and efficiency for face detection.

7. REFERENCES

[1] Y. Avrithis, A. Doulamis, N. Doulamis and S. Kollias, “A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases,” *Computer Vision and Image Understanding* **75** (1/2), pp. 3-24, July 1999.

[2] Y. Avrithis, N. Tsapatsoulis and S. Kollias, “Color-Based Retrieval of Facial Images,” *Proc. of EUSIPCO*, Tampere, Finland, Sept. 2000.

[3] Y. Avrithis, Y. Xirouhakis and S. Kollias, “Affine-Invariant Curve Normalization for Shape-Based Retrieval,” *Proc. of Int. Conf. on Pattern Recognition (ICPR)*, Barcelona, Spain, Sept. 2000.

[4] A.D. Bimbo and P. Pala, “Visual Image Retrieval by Elastic Matching of User Sketches,” *IEEE Trans. PAMI* **19** (2), pp. 121-132, 1997.

[5] C. Garcia and G. Tziritas, “Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis,” *IEEE Trans. on Multimedia* **1** (3), pp.264-277, Sept. 1999.

[6] L. A. Harwood, “A Chrominance Demodulator IC with Dynamic Flesh Correction,” *IEEE Trans. Consumer Electron.* **CE-22**, pp. 111-117, Feb. 1976.

[7] O. J. Morris, M. J. Lee and A. G. Constantinides, “Graph Theory for Image Analysis: an Approach based on the Shortest Spanning Tree,” *IEE Proceedings* **133**, pp.146-152, April 1986.

[8] T. Rzeszewski, “A Novel Automatic Hue Control System,” *IEEE Trans. Consumer Electron.* **CE-21**, pp. 155-162, May 1975.

[9] A. Samal and P.A. Iyengar, “Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey,” *Pattern Recognition*, **25** (1), pp. 65-77, 1992.

[10] N. Tsapatsoulis, N. Doulamis, A. Doulamis, and S. Kollias “Face Extraction from Non-uniform Background and Recognition in Compressed Domain,” *Proc. of ICASSP*, Seattle WA, May 1998.

[11] H. Wang and S.-F Chang, “A Highly Efficient System for Automatic Face Region Detection in MPEG Video,” *IEEE Trans. CSVT* **7** (4), August 1997.

[12] G. Yang and T.S. Huang, “Human Face Detection in Complex Background,” *Pattern Recognition* **27** (1), pp. 55-63, 1994.

[13] K.C. Yow and C. Cipolla, “Feature-based Human Face Detection in Complex Background,” *Image and Vision Computing Recognition* **15**, pp. 713-735, 1997.