MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan
8-16

Combining model-based classifiers for face localization

R.Belaroussi[1] L. Prevost[1] M. Milgram[1]
maurice.milgram {lionel.prevost} @upmc.fr
Univ. Paris VI [1]LISIF-PARC, BC252
4 place Jussieu 75252 Paris cedex 05 France

Abstract

We present a method to localize a face in a color image combining connexionist models (auto-associator networks), an ellipse model based on the orientation of the gradient and a skin color model. A linear combination of each model response is performed. Given an input image, we compute a kind of probability map on it with a sliding window. The face position is then determined as the location of the absolute maximum over this map. Improvement of localization rates of individual detectors is clearly shown and results are very encouraging.

1 Introduction

Face detection in an image without any hypothesis is a tough task because of the high variability of the pattern to be detected [1]. As in many detection issues, it is almost impossible to define the opposite class, the non-face patterns, which drives researchers to choose the models approach. Solutions implemented in a large number of face detection applications (biometric, presence detection, visiophony, indexation, car driver detection, virtual reality, lips reading) start with simplifying the problem by making assumptions : fixed camera and known background, use of motion information [2], strong hypothesis on the face location, special background for an easy extraction of the silhouette or special lighting conditions (use of infra-red, for example). Face localization (the face is in the image and we want to know where) is not simpler without additional assumption.
We find here the two approach common to the Pattern Recognition : structural and global. Structural approaches [3] try to detect primitives of the face (eyes, mouth, nose, head edge) then combine the results using geometrical and radio metrical models [4], or constellations analysis [5]. Global approaches process a sub-image of the input image into a feature vector (momentum, projection, gray level, wavelet...). These approaches use learning and test database to estimate parameters of the classifier. In the global approach, parameters can be weights (neural networks) [6] or terms of a covariance matrix (statistical classifier) [7]. A choice is then to be made between the model approach and the discriminant one. A model does not require counter examples, which may seems an advantage but actually decreases classifier efficiency : generalization in a high dimension space (221 for 13x17 sub-images) is tough without knowing where are the vectors that might be confused.

Our face localization approach combines an auto-associator network appearance based model, and an ellipse detector both based on the image gradient's direction, and a coarse skin color model in YCbCr color space.
Section 2 describes these three detectors, and their linear combination shall be found in Section 3. We present in Section 4 our experimental results and the benefits of the combination, before the conclusions and perspectives.

2 Basic detectors

A part of the information of a face image lies in orientation of its edges. A huge advantage of the edge's orientation is its relative invariance to the skin tone. Two of the three detectors presented in this section use this information : the neural network and the Hough transform.
Evaluation of the orientation of the gradient on the edges requires a low pass filtering of the image. Gradient field estimation uses Roberts masks, so that horizontal gradient is calculated by $I_x = I_{filtered} \otimes [1 \quad 1]$ and vertical gradient with $I_y = I_{filtered} \otimes \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Then the gradient magnitude is thresholded. For the generalized Hough transform, a global threshold is applied over the whole input image; this threshold was optimized over 168 images and is equal to 12. The thresholding for the diabolo is defined over each 13x17 sub-windows of the input image, so that 20% of the pixels are then regarded as edges. Orientations of these edge pixels are then quantized on N=36 values.

2.1 An appearance-based model : the auto-associative multi-layer perceptron (diabolo)

The next diagram presents the pre-processing of each example :

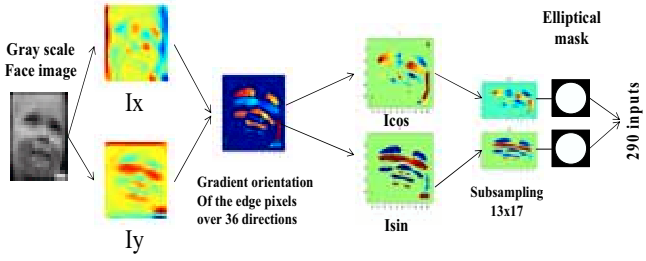


Figure 1. Training examples preprocessing

Each pixel is described by two features (I_{cos},I_{sin}) : $I_{\cos}(i,j)=\cos\left(\frac{2\pi}{N}\ orienx(i,j)\right)$ and $I_{\sin}(i,j)=\sin\left(\frac{2\pi}{N}\ orienx(i,j)\right)$ for the edges with $orienx = round\left(\frac{N}{2\pi}\arctan\frac{I_y}{I_x}\right) \bmod N$; (0,0) for the non edge pixels.

This step produces two arrays I_{cos} and I_{sin}, and we subsample them to size 13x17 using a bicubic interpolation. We crop these arrays with an elliptical mask -in order to remove border pixels values. This provide a 290 feature vector (note the dimension reduction from 17x13x2=442 due to the elliptical mask).

Processing of these examples is done with an aut-associator neural network, the so called "Diabolo" [8]. Such a network – which was successfully used for hand written characters recognition [9] - is trained to reconstruct an output identical to its input. It implements a specialized compression for its hidden layer has much less units than input or output does. A non-face image should be badly compressed and so the reconstruction error (square root of the mean square error between the input and the calculated output) would be higher than for a face image. The neural network is trained using 1602 face images as a learning set and a 178 face images set in cross-validation. Weight and bias values are updated according to gradient descent with adaptive learning rate. After an exhaustive search we found that for a 290 elements inputs (corresponding to a 17x13 retina) 18 hidden neurons are optimal. The gray level image is scanned at the resolution corresponding to the size of the face with a 13x17 retina, and at each position of the image a reconstruction error is calculated (figure 2) :

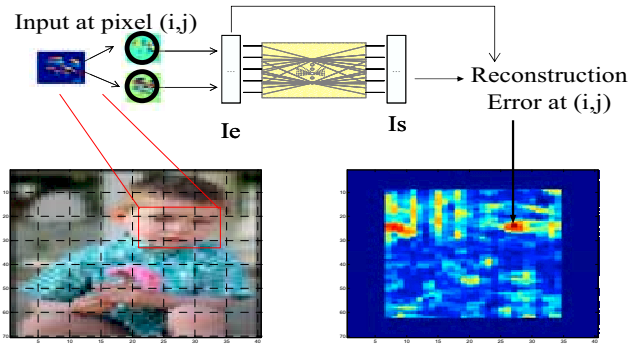


Figure 2. DiaboloMap : hot colours are small reconstruction error

so that an array of reconstruction errors is calculated –we will referre to it as DiaboloMap.

2.2 Ellipse detector based on Generalized Hough Transform

First the gray level dynamic of the input image is linearly adjusted between 0 and 255. This operation proved to be better than performing histogram equalization. Orientation of the gradient over the whole gray level image is then determined. Then a Generalized Hough Transform (GHT) is performed on the resulting orientation map : the HT constitutes a popular method for extracting geometrical properties. When the edge orientation is used and when it is applied to non parametric curves, the HT becomes the Generalized HT. Each edge pixel votes for all possible location of the shape (actually for the location of the barycentre). For ellipse detection, there is a simplified structure for the GHT based on the geometrical properties of ellipses. The method consists in casting votes for a line through each boundary pixel with an orientation determined by the edge one. We suppose that we know the orientation of the ellipse. So for each point M, a simple lookup table specifies the angle between the tangent Mt (to the boundary) and the radius MO (O is the centre of an ellipse passing through M). Faces are modeled as vertical ellipses with a specific eccentricity so we can build up our lookup table to cast votes from each edge pixel, knowing its gradient orientation. It provides a vote array which maximum correspond in the image to the position most likely to be the center of an ellipse with a horizontal minor axis a=8, and a vertical major axis b=10.

Due to the cluttered background this maxima does not correctly locate ellipse. To decrease effect of the background on votes, the vote map is first convolved with a 8x8 window then it is scanned using a 13x17 mexican hat which provides a new score map used to defined face location, we will referred to it as HoughMap (figure 4).

2.3 Skin color model

Independently of gradient's orientation information, a coarse skin detection has been implemented in the YCbCr color space [10]. Our coarse skin color filter is defined by : $Cb \in [105 \quad 130]$ and $Cr \in [135 \quad 160]$.

These thresholds were experimentally tuned using images with people. Here is the confusion matrix of this threshold operation over 145,736,580 skin pixels (set 3 of the ECU database [11]) and 667,359,498 non-skin pixels :

Tableau 1. Skin color detector confusion matrix

Classification / Class	Skin	Non-skin
Skin	76%	24%
Non-skin	19%	81%

This skin detector is coarse and in some case no skin at all is filtered but the combination of the detectors enables us to use a simple model. The advantage of YCbCr space is that it separates luminance (Y), chrominance (Cb-Cr). Skin being characterized by specific chrominance information, the filter can be applied to any ethnic skin color but our thresholding is not universal because the chrominance components is actually related to the luminance value Y [12]. In poor or bright illumination condition the filtered components are spurious. The noisy non-skin mask is then low pass filtered by convolving it with a rectangular 13x17 window. A morphological

operation is then applied in order to fill eventual holes, and the result is thresholded. At each point of the image, we compute proportion of skin-color pixels inside the 13x17 neighbors. The result map is called SkinMap.

3 Combining different detectors

We have implemented three detectors for a color image, which result in three maps : DiaboloMap, HoughMap, and SkinMap. We decided to combine linearly these three maps to improve localization performances [13].

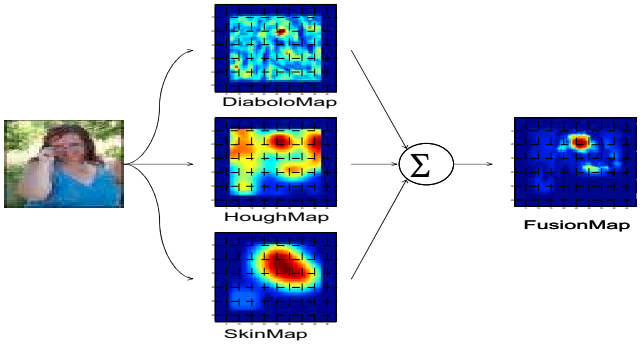


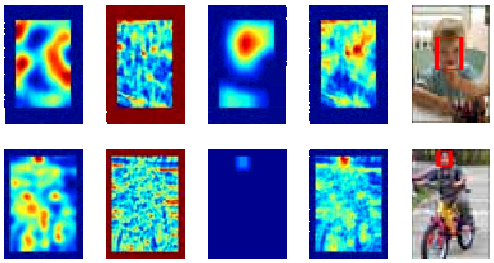
Figure 3. Overview of the face localization system

For that purpose, each detector map is linearly adjusted onto $[-1 \quad 1]$. We will denote these normalized maps as D, H, and S. Using the three detectors, a pixel (i,j) in the original image is then featured by $I_{i,j} = [H_{i,j} \quad D_{i,j} \quad S_{i,j}]$. The 100 images were used to learn linear combination parameters using a gradient descent stopped by cross validation :

$$FusionMap_{i,j} = a \cdot H_{i,j} - b \cdot D_{i,j} + c \cdot S_{i,j} + d$$

(a=0. 2280, b=0.2620 , c=0. 1229 and d= -0.7198)

We can notice that the weight of the skin detector is smaller than Hough and diabolo detectors ones.



(a) (b) (c) (d) (e)

Figure 4. (a) HoughMap (b) DiaboloMap (c) SkinMap (d) FusionMap and (e) the corresponding face location on the original image

4 Face localization performances evaluation

To determine face location, the input image is scanned (using the three calculated maps) and the location of the face is defined as the maximum of the resulting map FusionMap. In the first example of figure 4, HoughMap maximum failed to locate correctly face, while the combination system did. In the second example, the SkinMap maximum is very low (0.23), but the combination brings a correct face location.

We use the ECU face database [11] which is composed of color images (set 1), and two corresponding ground truth : one for the faces (set 2) and one for the skin (set 3). This database includes faces with various poses and skin tones.

A first test uses a set of 1353 images (non overlapping with the training and cross-validation corpus) containing only one face, to evaluate localization rate. For each image, face's size is supposed to be known, which enables us to apply a 13x17 window sliding strategy, this knowledge is equivalent to knowing the distance between the camera and the person to be localized. Each image of the test set is first resized so that the face (defined by the ground truth) reaches the size 13x17. This size was chosen to respects faces aspect ratio, and is a good compromise between face's feature visibility (by human vision) and computational efforts.

A face will be correctly localized if the detection bounding box covers at least 60% of the area of the bounding box defined by the ground truth. Figure 5 shows examples of detections at 60% within the ground truth (the face image in the middle) :

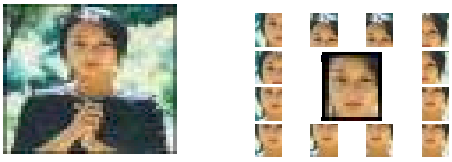


Figure 5. Examples of correctly localized faces

Using this performance measure, the image (figure 5) which is 49x53 contains 111 face sub-images, and 1258 non-face sub-images.

To evaluate improvement brought by the combination of the 3 detectors, the proportion of pixels of the ground truth inside the detection (which is the location of the maximum for HoughMap and FusionMap, and of the minimum for DiaboloMap) is calculated for each test image (figure 6). Using the DiaboloMap minimum position alone, 656 faces are correctly located (proportion of good pixels greater than 0.6) in our test set (48.5%). Using the maximum of HoughMap alone to define the location of the face, 903 faces are correctly detected (67%).

After combination, 1166 faces (over the 1353 test images) are correctly detected which increases the detection rate to 86 %. The combination of the three detectors decreases the error rate of 50%.

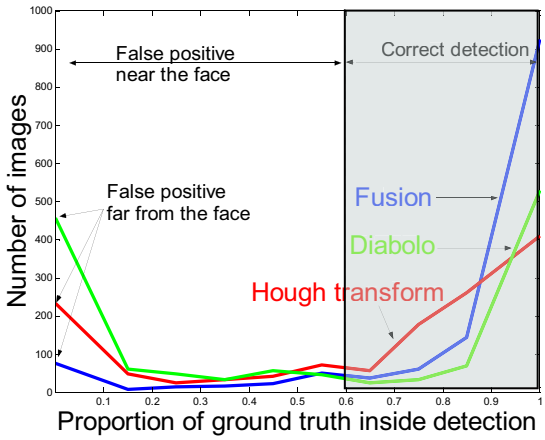


Figure 6. Histogram of the proportion of good pixels inside the best window. Fusion is better both in correct localization (proportion>0.6) and in false positive.

A second test was performed on 205 multiple faces images (non overlapping with the training and cross-validation corpus) containing a total of 482 faces. In single face images, face location is defined as the position of the maximum of FusionMap. In a N faces image (N is supposed to be known in a localization problem) the N highest maxima (with a sufficient distance to avoid overlapping detections) of FusionMap correspond to the location of the faces. 396 faces are correctly localized (82%), some examples are shown in figure 7.

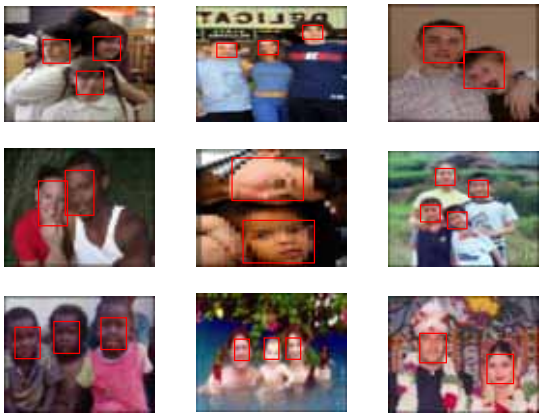


Figure 7. Multiple faces localization : the number of faces are supposed to be known

5. Conclusion and prospects

This communication aimed to present a significant contribution to the face localization task. We have presented three different detectors: skin color, auto-associative multi-layer perceptron and Ellipse Hough Transform. We have shown that a linear combination of the three feature maps drastically improves localization rate both on single and multiple face(s) images.

Several improvements are in progress: more sophisticated skin color models like ellipsoidal thresholding, Gaussian density functions [14] or mixture of Gaussians [15]; neural combination of our feature maps should bring better results than the linear one. Moreover we have proposed a robust eyes detection algorithm; it is envisaged to use it for face candidates verification.

References

[1] M.-H Yang, D. Kriegman, and N. Ahuja, Detecting Faces in Images: A Survey, *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 34-58, Jan. 2002.*

[2] R.Choudhury Verma,C.Schmid & K.Mikolajczyk, Face Detection and Tracking in a Video by Propagating Detection Probabilities, *IEEE Trans. Pattern Analysis and Machine Intelligence,vol.25, no 10, Oct. 2003.*

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

[4] A. Yuille, P. Hallinan & D. Cohen : Feature Extraction from Faces Using Deformable Templates, *Int J. Computer Vision, vol. 8, no. 2, pp. 99-111, 1992.*

[5] S.M. Bileschi & B.Heisele, Advances in Component Based Face Detection, *Proceedings of the IEEE international Workshop on Analysis and Modeling of Face and Gestures 2003.*

[6] C.Garcia & M.Delakis, Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, no. 11, Nov. 2004.*

[7] K.K. Sung , T. Poggio, Example-Based Learning for View-Based Human Face Detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence, v.20 n.1, p.39-51, January 1998*.

[8] R. Féraud, O. Bernier, J. Viallet, and M. Collobert, A Fast and Accurate Face Detector Based on Neural Networks, *IEEE Trans.Pattern Analysis and Machine Intelligence, vol. 23, no. 1, pp. 42-53,Jan. 2002.*

[9] H.Schwenk,M.Milgram : Transformation invariant auto-association with application to handwritten character recognition, *NIPS'7 (Neural Inf. Proc. Syst.),pp 991-998, 1995.*

[10] D.Chai and K.N.Nang Locating facial region of a head-and-shoulders color image. *Proceedings of the Third International conference on Automatic Face and Gesture Recognition, p.124-129, 1998.*

[11] S. L. Phung, A. Bouzerdoum, and D.Chai, Skin segmentation using color pixel classification: Analysis and comparison, *IEEE Trans. Pattern Analysis and Machine Intelligence, to be published in 2005.*

[12] M.Hu, S.Worrall, A.H. Sadka, A.M. Kondoz, Automatic scalable face model design for 2D model-based video coding, *Signal Processing: Image Communication, 19 (2004) 421-436*

[13] L.Prevost & M.Milgram, Automatic Allograph Selection and Multiple Expert Classification for Totally Unconstrained Handwritten Character Recognition, *ICPR'98, (1), pp 381-383, Brisbane, Australia, 1998.*

[14] M.-H Yang and N.Ahuja, Detecting human faces in color images, *Proceedings Of IEEE International Conference on Image Processing, vol. 1, p.127-130, 1998.*

[15] S.J.McKenna, S.Gong, Y.Raja, Modelling facial colour and identity with gaussian mixtures, *Pattern Recognition 31, No.12, pp.1883-1892, 1998.*