Comparison of Different Combination Strategies for Face Localization

R. Belaroussi
Universite Paris VI
LISIF-PARC, BC252
4 place Jussieu 75252 Paris cedex 05 France

L. Prevost
Universite Paris VI
LISIF-PARC, BC252
4 place Jussieu 75252 Paris cedex 05 France

M. Milgram
Universite Paris VI
LISIF-PARC, BC252
4 place Jussieu 75252 Paris cedex 05 France

Abstract - *We present a comparison between different image fusion methods dedicated to the localization of faces in color images. The data to combine results from a connexionist model (auto-associative network), an ellipse model based on Generalized Hough Transform, and a skin color model. The combination methods compared have clearly different approaches. They include the Bayesian classifier with parametric or non-parametric technique, a fuzzy inference system, and a connexionist approach of the weighted average. 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 baseline detectors localization rates is clearly shown and prevalence of the weighted average is reported.*

1.0 Introduction

Face detection in an image has become a very important issue for many applications such as biometric, presence detection, video-conferencing, visiophony, indexation, car driver monitoring, virtual reality, lips reading, gaze tracking. Because of the high variability of the pattern to be detected, face detection without any hypothesis is a tough task [1]. 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 detection applications start with making assumptions. The face localization issue can be regarded as a face detection problem knowing the number of faces in the image. It is not simpler without additional assumption.

Structural and holistic approaches, common in Pattern Recognition, are applied. Structural approaches [3] try to detect facial landmarks (eyes, mouth, nose, head contour) then combine the results using geometrical and radiometrical models [4], or constellations analysis [5].

Holistic approaches of face detection process a sub-image of the input image into a feature vector (momentum, projection, gray level, wavelet...). These approaches estimate the classifier parameters on a training set. Parameters can be weights (neural networks) [6], [7] or terms of a covariance matrix (statistical classifier) [8]. As in many detection issues, it is almost impossible to define the opposite class, the non-face patterns, which drives researchers to choose the model-based approach. 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 patterns that might be confused. Another way is to design a combination of several detectors (classifiers). [9] did it to perform face detection and classifier combination has also been used in character [10] and face recognition [11].

In order to get the best of both worlds, our method makes co-operate holistic and structural approaches. Generalization capability of a single classifier is limited, especially in a high dimension space. A more reliable decision can be obtained by combining output of several experts [12] : the face localization issue is divided in sub-problems easier to deal with. Various information are extracted from the same image using different kind of detectors. Some try to model global features while the others concentrate on structural features. Each face cues are searched by a relevant experts : elliptical shape, global appearance and skin color. Cooperation between experts exploits their complementarities and can also handle conflicts between sources.

An auto-associator network appearance based model and an ellipse detector are based on the image gradient's direction. A coarse skin color model in YCbCr color space is also implemented. The combination of these three detectors is done via various methods that are compared: Bayesian classifier, fuzzy inference system and neural networks.

Section 2 describes the three detectors and the combiners. Comparison of the combinations is detailed in Section 3 along with our experimental results and the contribution of the combination. The last section is devoted to conclusions and prospects.

2.0 Combining model based classifier for face localization

2.1 Basic detectors

Face localization task consists in, knowing the number of person in an image, finding the correct location of their face. As this paper aims at comparing several combination strategies, the problem is simplified: for each image, face's size is supposed to be known. This enables us to apply a sliding window strategy at only one scale. This knowledge is equivalent to knowing the distance between the camera and the person to be localized.

At each position of the image three facial characteristics are computed : appearance of the interior of the face, its elliptical shape and the proportion of skin pixels. Each of these three models results in a single number for a given sub-image. The computations are done over a sliding window with a fixed size and each sub-image is featured by a 3D vector [H D S]. This process leads to three maps for a given image, and the combination of these maps to a final face probability map as shown in figure 4.

Edge orientations information is processed by an appearance-based model (so called Diabolo) and an ellipse detector (Generalized Hough Transform). The Diabolo [13] was successfully used for handwritten character recognition [14], face detection [13] and compression [15]. It is an auto-associator network : its number of output equals its number of input. It is trained to reconstruct an output identical to its input, and only face examples constitute the training database. It implements a specialized compression for its hidden layer has much less units than input or output does. So a non-face image should be badly compressed and 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, see figure 1. The detailed coding of the input image is explicated in [16].

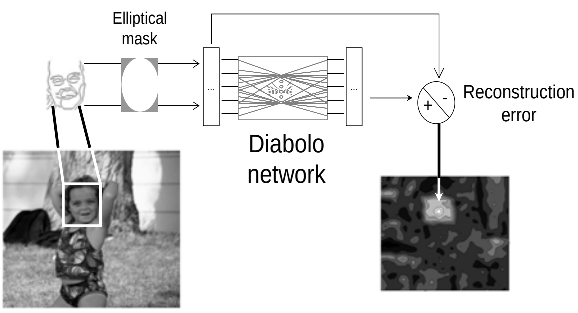

Figure 1 DiaboloMap: array of reconstruction errors calculated at all position of the image.

Interior part of the face are used to train the network using an elliptical mask to reduce border effect and in order not to model the elliptical shape of the face. The Diabolo is trained to model facial features: mouth and eyes, mainly. This approach is different from a face detector based on neural network which takes the face contour into account: this enhances the face detection rate. Our approach aims at compute face contour and facial feature separately. This way, redundant information between our two first detectors are reduced. The elliptical shape of the face is search using a Generalized Hough Transform [17]: faces are modeled as vertical ellipses with a specific eccentricity.

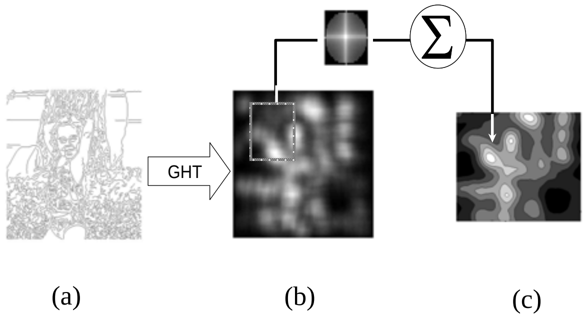

Figure 2 Edge detection (a), Generalized Hough Transform (GHT) computed over gradient orientation of the edge (b) and resulting HoughMap (c).

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 upright ellipse. The accumulator is scanned with a sliding window and at each position a weighted average of the number of vote is calculated as shown in figure 2.

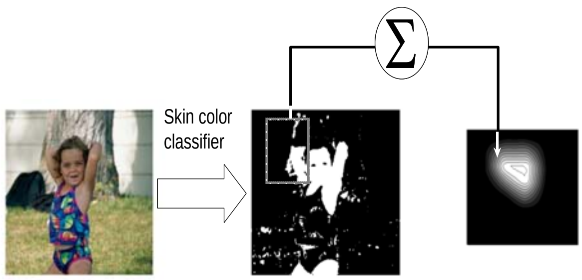

Figure 3 SkinMap : proportion of skin pixels array.

A coarse skin color model [16] of chrominance in Cb and Cr channels of YCbCr color space is implemented. A pixel is classified as a skin pixel if it falls in a fixed range of Cb and Cr. Proportion of skin-like pixels in the sliding window constitutes the third detector (see figure 3).

2.2 Combination algorithms

We have implemented three holistic detectors for a color image, which result in three maps : DiaboloMap, HoughMap, and SkinMap. When each detector alone failed to model facial features, the combination of the three sources can achieve this task very well. The combination can also handle conflicts between sources.
For that purpose, each detector map is linearly adjusted onto [-1 1]. Using the three detectors, a pixel (i,j) in the original image is then featured by $I_{i,j}$ = [H D S].
Several architectures exists for data fusion, we can divide them into three kinds : serial (or sequential), parallel and hybrid (mixing sequential and parallel, with feed-back or interaction…). Our face localization system has a parallel architecture (see figure 4).

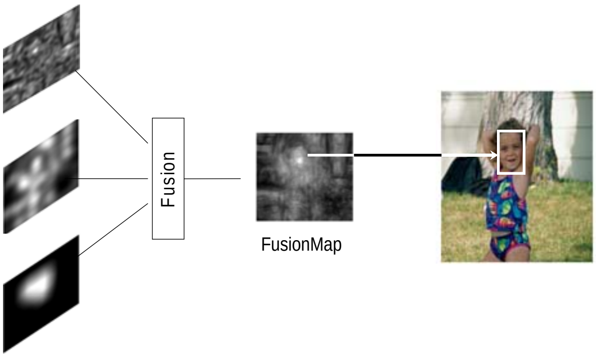

Figure 4 Overview of the face localization system.

Combination rules are various, depending on the application : mean, weighted sum, product or maximum of experts outputs, majority vote, fuzzy rules, neural networks, or neuro-fuzzy inference for example.
Several algorithms have been proposed for combining our three detectors : parametric and non-parametric combination strategies are described in this sub-section. The next section is dedicated to their comparison.
The training data set is made of 19.579 face examples and 482.783 non-face examples, except for the neural combination as explained hereafter.

2.2.1 Bayesian classifier : parametric and non-parametric approaches

The Bayesian decision rule is a classical technique in statistical pattern classification [17]. A sub-image featured by a 3D vector $\vec{X} = \left[H \ D \ S\right]$ is a face if

$$\frac{P\left(\vec{X} \mid face\right)}{P\left(\vec{X} \mid non\text{-}face\right)} \geq \tau \qquad (1)$$

where $P\left(\vec{X} \mid face\right)$ and $P\left(\vec{X} \mid non\text{-}face\right)$ are respectively the conditional probability density function (pdf) of the class face and non-face. τ is a threshold usually estimated over a training set. As the application presented in this paper is face localization and not face detection, no estimation of τ was done. The face location shall be the one that maximize the value of the likelihood ratio (left hand side of equation 1) :

$$Face\ location \leftrightarrow \max\left(\frac{P\left(\vec{X} \mid face\right)}{P\left(\vec{X} \mid non\text{-}face\right)}\right) \qquad (2)$$

Parametric and non-parametric estimations of the class-conditional pdf are implemented.
The histogram technique is a non-parametric method. For each class a 3D histogram is computed using the training examples. Due to the small amount of face examples, it has only five bins per dimension : 5^3 being equal to 125, a mean of about 160 examples per bin is available. We combine the two histograms obtained into one histogram which bins values are the ratio of the bins frequency of the two preceding histograms (face/non-face). Resulting histogram values are then scaled into $\left[0 \ 255\right]$. When a test image is processed three planes are calculated corresponding to our three face models over a sub-window sliding at each position of the image (see figure 4). For each position of the test image a 3D vector is computed and a back-projection of the histogram is done by a look-up table operation. This back-project is the FusionMap illustrated in figure 4, face location should correspond to the position of its maximum value.
A parametric approach is to model both skin and non skin class-conditional pdf by a unimodal Gaussians. The face location is then defined as the position of the maximum of the logarithm of the likelihood ratio : $\left(\vec{X}-\vec{M}_{face}\right)^T \Sigma_{face}^{-1}\left(\vec{X}-\vec{M}_{face}\right)-\left(\vec{X}-\vec{M}_{non\text{-}face}\right)^T \Sigma_{non\text{-}face}^{-1}\left(\vec{X}-\vec{M}_{non\text{-}face}\right)$
where the parameters of the Gaussian $\left(\Sigma, M\right)$ are the mean and covariance matrix of each class computed over the training set. If $\vec{X}^{\,i} = \left(H^i \ D^i \ S^i\right)^T$ is the i^th example out of N_{faces} of the face training set:
$\vec{M}_{face} = \frac{1}{N_{faces}} \sum_{i=1}^{Nfaces} \vec{X}^{\,i}$ is the mean faces vector and
$\Sigma_{face} = \frac{1}{N_{faces}} \sum_{i=1}^{N_{faces}} \left(\vec{X}^{\,i} - \vec{M}_{face}\right) \cdot \left(\vec{X}^{\,i} - \vec{M}_{face}\right)^T$ is the covariance matrix of the face class.
Other parametric functional forms of the pdf were investigated. The simplest is a unimodal Gaussian of the face class : this assume that the non-face class is uniformly distributed. In this case the face location is defined as the maximum of the square Mahalanobis

distance to the mean center of face training examples. Mixture of Gaussians were also tested but led to very poor results. Due to the small amount of training data available, this method is out of scope in this paper.

2.2.2 Fuzzy inference system

A face sub-image should be featured by a small Diabolo reconstruction error D, a high number of GHT votes H and a high proportion of skin pixels S. A classical set approach would define a threshold on each face model values. H_{high} = { H | H > thresh } the set of high H (for instance) values and H_{small} = { H | H < thresh } the set of small H values would be separated by this sharp boundary: a H value slightly under that threshold is then considered as small which make few sense. The fuzzy logic approach is more flexible by admitting partial membership to a class [18]. It is also coherent with natural language by introducing the degree of membership of H value in the class “high” and “small” : H_{high} = { H, μ_{high}(H)} and H_{small}= { H, μ_{small}(H)}.

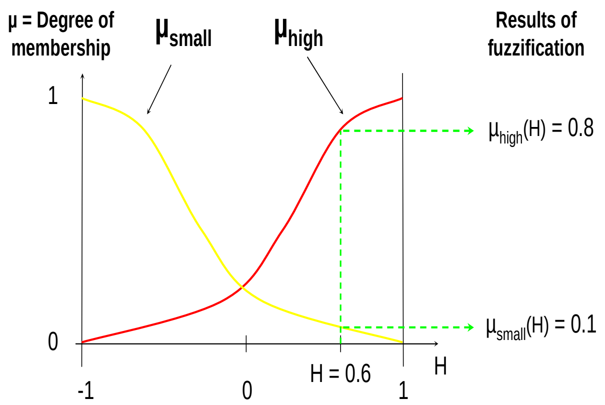

Figure 5 Membership functions of the class "H high" and "H small".

In figure 5 a value of H=0.6 belongs the “high” class at 80% and the “small” class at 10%.
S value is the normalized proportion of skin in the sub-image : as H, high values of S correspond to high probability of the sub-image to contain a face. D value is the normalized Diabolo reconstruction error : the smaller it is, the higher is the probability of the sub-image to be a face one. For these three sources two class are defined with respect to their value : high and small. As shown in figure 5, the membership functions for these classes are Gaussian functions centered respectively in +1 and -1.
To combine our three sources a fuzzy inference system of Mamdani type [19] was built. A fuzzy inference system requires to fuzzify inputs, to formulate a set of linguistic rules and logical operators, and to aggregate results of the fuzzy rules. Three output class are defined as fuzzy sets : non-face, unknown, and face patterns.

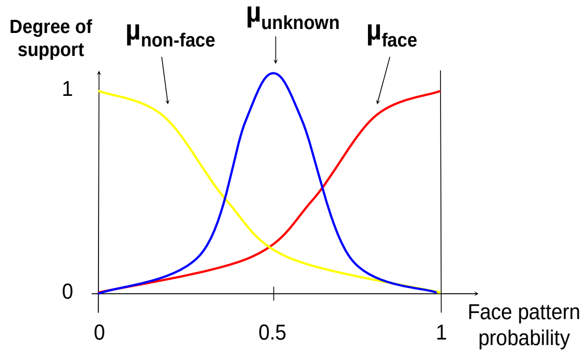

Figure 6 Output fuzzy sets membership functions.

Each output set is defined by a Gaussian membership function centered in 0 (non-face), +0.5 (unknown) or +1 (face), as shown in figure 6.
Considering only the ellipse model (H value), a simple statement can be formulated : if H is high then sub-image is a face. Consequent of this fuzzy rule assigns a fuzzy set to the output which membership function is a truncation of the “face” set depending on the degree of support and according to the implication method (ie the mathematical definition of “then”). Degree of support in this particular statement only involving H value is the degree of membership in the “H high” class. The “then” operator results in a membership function equal to the minimum between the degree of support and the output fuzzy set membership function (the green area in figure 7 showing the case of H=0.6).

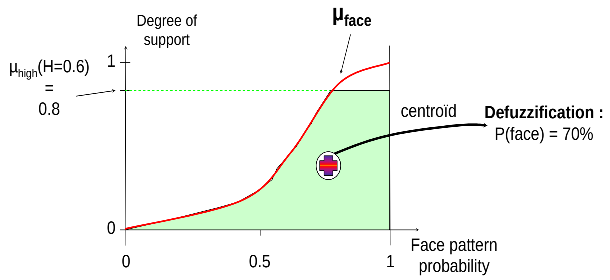

Figure 7 Implication method : “then” operation.

Finally a decision can be made out of the resulting function by resolving a single value representing the probability of the sub-image to be a face pattern. A typical defuzzification method is the calculation of the center of the area under the curve (centroïd).
Now consider a statement with multi-part antecedent : if H is small or D is high or S is small then sub-image is unknown. The “or” fuzzy operation is mathematically defined as maximum of the three calculated degree of membership: this minimum is the degree of support for the output “unknown” set. In figure 8 a sub-image is featured by [H D S] = [0.6 0.8 -0.2] : for each source, a degree of membership is calculated. The “or” operation resolve them to a single number : the higher value is kept

as degree of support for the rule shaping the "unknown" fuzzy set.

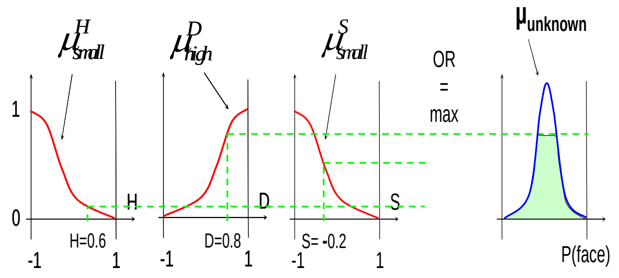

Figure 8 Application of fuzzy operator "or".

One rule by itself is show very poor localization rate. We found experimentally that the three following fuzzy rules are optimal for face localization purpose:
-if H is high and D is small and S is high then sub-image is a face,
-if H is small or D is high or S is small then sub-image is unknown,
-if H is small and D is high then sub-image is a non-face.
The "and" operator is defined as the minimum of the degrees of membership. The rules are given the same weight, and order of the rules is unimportant as they are evaluated in parallel as shown in figure 9.
One can notice that the skin detector is not taken into account in the last rule: our skin color model is not elaborated enough and this is also noted with a weighted average combination (see next section).

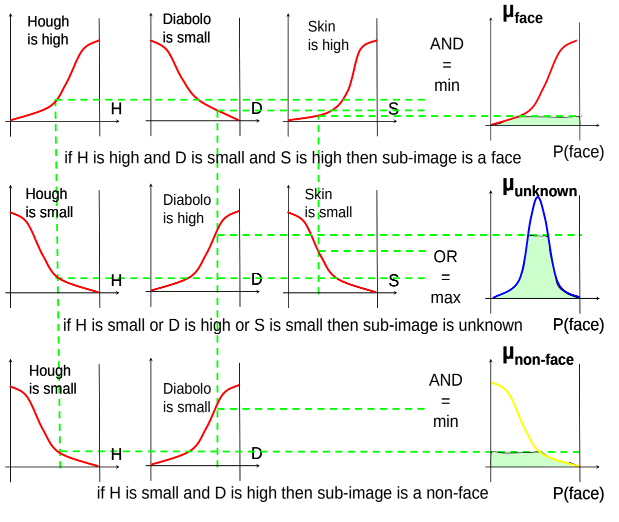

Figure 9 Fuzzy inference diagram representing the rules.

Aggregation of the output fuzzy sets consists in calculating a membership function as the maximum of the three consequent membership functions calculated before (see figure 10).
This membership function is finally defuzzified by calculating the centroid of it, which provide a single number: the probability that the input sub-image is a face image.

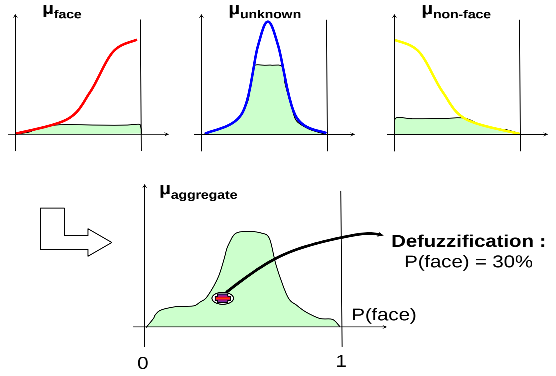

Figure 10 The aggregate output fuzzy set.

This process is applied at all position of the original image to construct the "fuzzy" FusionMap used to define face location.

2.2.3 Weighted average and multilayer perceptron

We investigate neural combination of the three face models: the three sources are the inputs of the multi-layer perceptron (MLP). The MLP hidden neurons have sigmoid activation function, and the output neuron has a linear activation function as described in figure 11.
The training database is divided in two set : one for neural networks training, the other for assessment of the best architecture. 12713 face examples and 341316 non-face examples are used as training examples to learn parameters (weights) of the MLPs. Training is done using a gradient descent with adaptative learning rate stopped by cross-validation**.** Gradient descent algorithm is a standard backpropagation in which the network weights are moved along the negative of the gradient of the performance function. The performance function implemented here is the sum over training examples of the square difference between target and simulation (output calculated by the MLP).

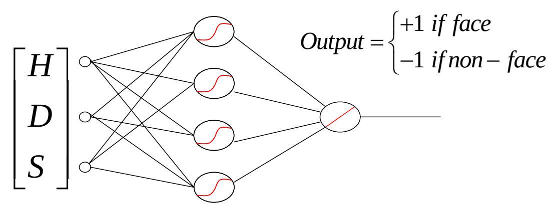

Figure 11 Combining the three detectors with a Multi-Layer Perceptron (MLP)

The network is trained for pattern classification : target is +1 when the input [H D S] correspond to a face and -1 else. Before training the MLP weights must be initialized: a different initialization leads to different weights, therefore to different networks. For a given

neural net architecture, several initialization must be tested in order to avoid the network to fall in a local minimum of the performance function different to its global minima.
During the test phase the MLP output is a value of the interval [-1 +1]. Output of the network is calculated at all location of the image, which produces the "neural" FusionMap : face location is the position of the maximum of this map. The optimal MLP architectures is searched over 50 images (not used during training) containing only one face. MLPs with different number of hidden neurons, and different initialization of the weights are trained then assessed over this set. This exhaustive search leads us to the conclusion that the best architecture correspond to zero hidden neurons. Actually a growing number of hidden cells do not dramatically decrease the localization rate: for numbers of hidden neurons under 3 the rates are quiet the same order. The natural approach is to choose the simplest architecture for the MLP: that is to say the best neural combination is a weighted average of the inputs :

$$FusionMap_{i,j} = a.H_{i,j} + b.D_{i,j} + c.S_{i,j}$$

where a=0. 2280, b= –0.2620 , and c=0.1229 after training. One can notice the weight of the S input: as in the preceding section it is half less important than the Hough or Diabolo response. This is due to the fact that the skin color model is pretty coarse.
This weighted average is compared to a simple average (same weight for the inputs : a=b=c=1) in the next section.

3 Experimental results

In order to compare the combination strategies we used the ECU face database [20] : we compare the face localization rate of the algorithms on a test set of color images not used during training. Each of these image contains only one person, and the rectangle bounding the face is the same size over the whole set. A face is considered as correctly localized or not by a human operator. A correct localization of the face contains the eyes, the mouth, and is well-centered on the face.

3.1 ECU face detection database

The ECU face and skin detection database was created in Edith Cowan University [20]. It has three sets of images each that was particularly useful in our study. The first set is made of original color images. The second set is the corresponding ground-truth location of the faces. The third set is the ground-truth of skin pixels.
Almost all the image are taken from the Web, and were selected to have wide variety of illumination conditions, background (mostly complex), face poses (upright, pan, tilted) and skin tones. The database is widely depicted in [20].
Our test uses a set of 1353 images non overlapping with the training and cross-validation corpus.

3.2 Comparison of the combining classifiers

In the preceding section, different combination algorithms have been proposed. They include Bayesian classifier with parametric (unimodal Gaussian of face and non-face) and non parametric techniques (histogram), fuzzy inference system, neural combination and weighted average.
It is important to outline the contribution of combination, and a reference for face localization rates.
If we only consider the appearance-based model alone, the face location is defined as the minimum of Diabolo reconstruction error over the whole image, as a face image should be better reconstructed than a non-face image. Under this consideration, 656 faces out of 1353 are correctly localized: the localization rate is **48.5%**.
Using the ellipse model alone, the face location is defined as the maximum number of vote given by the Generalized Hough Transform: 903 faces are correctly localized. The face localization rate is higher: **67%**.
The Bayesian classifier with the histogram technique reaches a rate of only **22%**. A unimodal 3D-Gaussian of the face class gives a poor **5%** of success. The fuzzy approach is more efficient with a face localization rate of **72%**. These methods bring no or poor improvement in the face localization rate.
An unimodal 3D-Gaussian of face and non-face class achieves **84%**: modeling the non-face pattern dramatically increases the face localization rate.
Amongst all these methods, the weighted average performs the best. With a localization rate of **86%** it outperforms all the other approaches. In order to measure the effect of the weights on the detection result, a simple average (ie all weights equal 1) is performed. With a rate of **80%** it perform well too, but less than the weighted average with the weights learned by gradient descent.
Amongst the multiple classifier systems, linear combiners are the most frequently used: a recent study can be found in [21] with a theoretical analysis based on the framework of [22],[23]: the analysis of linear combiners is still a promising path of research.

4. Conclusion and prospects

This communication aimed to present a significant contribution to the image fusion task with application to face localization. We have presented three different detectors: skin color, auto-associative multi-layer perceptron, and ellipse Hough Transform. We proposed three various combination scheme and compare them: Bayesian classifier, fuzzy logic and connexionist. An awesome improvement of localization rate is brought by the two last methods.

For the face detection/localization issue, several improvements are in progress: more sophisticated skin color models like ellipsoidal thresholding, Gaussian density functions [24] or mixture of Gaussians [25]. A more efficient appearance-based model is also elaborated, based on the Viola&Jones face detector [26] For the combination part, it is not clear when and why a combination method outperforms the others: quantitative and qualitative investigations of classifiers output correlation effect on combiners performance is under study.

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 and 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 and T. S. Huang. "Human face detection in complex background." Pattern Recognition, vol. 27, no. 1, pp. 53-63, 1994.

[4] A. Yuille, P. Hallinan and D. Cohen. "Feature extraction from faces using deformable templates." International Journal of Computer Vision, vol. 8, no. 2, pp. 99-111, 1992.

[5] S.M. Bileschi and B.Heisele. "Advances in component based face detection." IEEE international Workshop on Analysis and Modeling of Face and Gestures 2003.

[6] H. Rowley, S. Baluja and T. Kanade. "Neural network-based face detection." IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23-38, Jan. 1998.

[7] C. Garcia and 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.

[8] K.K. Sung and 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.

[9] B. Froba and W. Zink. "On the combination of different template strategies for fast face detection." Multiple Classifier Systems, pp. 418-428, 2001.

[10] A.F. Rahman and M.C. Fairhust. "Multiple classifier decision combination strategies for character recognition: a review." International Journal on Document Analysis and Recognition, (5) pp. 166-194, 2003.

[11] J. Czyz, J. Kittler and L. Vandendorpe. "Combining face verification experts." International Conference on Pattern Recognition, vol 2, pp. 28-31, 2002

[12] L. Prevost and M. Milgram. "Automatic allograph selection and multiple expert classification for totally unconstrained handwritten character recognition." International Conference on Pattern Recognition, vol 1, pp 381-383, 1998.

[13] 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, 2002.

[14] H. Schwenk and M. Milgram. "Transformation invariant auto-association with application to handwritten character recognition." Neural Information Processing Systems 7, pp. 991-998, 1995.

[15] D. DeMers and G. Cottrell. "Non-linear dimensionality reduction." Neural Information Processing Systems 5, pp 580–587. 1993.

[16] R. Belaroussi, L . Prevost and M. Milgram. "Combining model-based classifiers for face localization." IAPR conference on Machine Vision Application 2005, p.290-293.

[17] R.O Duda, P.E. Hart, and D.G.Stork. "Pattern Classification." John Wiley and Sons, 2001.

[18] L.A. Zadeh. "Outline of a new approach to the analysis of complex systems and decision processes." IEEE Transactions on Systems, Man, and Cybernetics, Vol. 3, No. 1, pp. 28-44, Jan. 1973.

[19] E.H Mamdani. and S. Assilian. "An experiment in linguistic synthesis with a fuzzy logic controller." International Journal of Man-Machine Studies, Vol. 7, No. 1, pp. 1-13, 1975.

[20] S.L. Phung, A. Bouzerdoum and D. Chai. "Skin Segmentation Using Color Pixel Classification: Analysis and Comparison." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 1, pp. 148-154, Jan., 2005.

[21] G. Fumera and F. Roli. "A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems." IEEE TPAMI. 27(6): 942-956 (2005)

[22] K. Tumer and J. Ghosh. "Analysis of decision boundaries in linearly combined neural classifiers." Pattern Recognition 29(2): 341-348 (1996)

[23] K. Tumer and J. Ghosh. "Linear and Order Statistics Combiners for Pattern Classification." CoRR cs.NE/9905012: (1999)

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

[25] S.J. McKenna, S. Gong, and Y. Raja. "Modeling facial color and identity with gaussian mixtures." Pattern Recognition, vol 31, no.12, pp.1883-1892, 1998.

[26] P.A. Viola, M.J. Jones. "Rapid Object Detection using a Boosted Cascade of Simple Features." CVPR (1) 2001: 511-518