Face localization in color images with complex background

Paola Campadelli, Raffaella Lanzarotti, Giuseppe Lipori

Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
Via Comelico, 39/41 20135 Milano, Italy
{campadelli, lanzarotti, lipori}@dsi.unimi.it

***Abstract*— In this paper we describe a two step algorithm which localizes faces in 2D color images depicting a single face on a complex background. Given a single image, the algorithm roughly determines the skin regions and then searches for eyes within them. A face is localized if at least one eye is present in a skin region. The system is based on a Support Vector Machine trained to separate sub-images representing eyes from others. The algorithm is robust to scale, illumination, pose variations and deals with partial occlusions. Results on several public databases are presented.**

Keywords: Face localization, skin color model, Support Vector Machine (SVM).

I. Introduction

In recent years a great deal of research work has been devoted to *face image processing* due to the interest of the pattern recognition problems it involves, and to the richness of potential applications such as face recognition and authentication, facial expression recognition, face tracking, model-based coding of video sequences, etc. Albeit these application fields have different aims, they all have a common denominator: to detect in images those locations where faces are present. This task is called *face localization* in case of input images depicting only one subject in the foreground[1].

Such task is challenging because of the face manifold owing to the high inter-personal variability (e.g. gender and race), the intra-personal changes (e.g. pose, expression, presence/absence of glasses, beard, mustaches), and the acquisition conditions (e.g. illumination and image resolution).

To our knowledge the best performance in face localization has been presented by Smeraldi and Bigun in [14], but the authors reduced the problem complexity by fixing the scale.

In this paper we propose a two-module system: at first it searches for skin regions within the image allowing to restrict the search area for the subsequent module (section II); then an SVM is applied only in correspondence to the skin regions with the objective of discriminating between faces and non faces (section III). In particular we trained the SVM to recognize eyes, meaning that an eye within a skin region validates it as a face. This solution tries to exploit the advantages of both feature invariant approaches [15], [17] and appearance based methods [7], [11].

Work partially supported by project "Acquisizione e compressione di Range Data e tecniche di modellazione 3D di volti da immagini", COFIN 2003

[1]On the contrary we define *face detection* the case when no assumption is made regarding the number of faces in the images.

II. Skin detection

The objective of this step is to greatly reduce the search area in input to the classifier by excluding the regions that do not correspond to skin on the basis of their color properties. Subsequently we exploit further information to better characterize faces among skin regions.

In the following we present the skin color model construction (II-A), the algorithm which selects the skin regions (II-B) and the experimental result of the method (II-C).

A. Skin color model

The skin color model consists of a mixture of two bidimensional Gaussians, each one parameterized by $(\mu_i, \sigma_i^2 I)$, since it is simple and statistically well justified [17]. The parameters are estimated using the EM algorithm [10], [13] on the basis of 4-million sample of skin colors, and referring to the chrominance components in the *YPbPr* color space.

In order to obtain a color sample which is representative of different illumination conditions and human races, we have gathered the skin colors from two databases, the BANCA-Controlled and the DBLAIV.

- **BANCA** [4] **- Controlled:** from this database we collected two million pixel colors, taken from the first and the last image of each section for all the 52 recorded people and for all the four sections.

- **DBLAIV[2]:** from this database we collected two million pixel samples, referring to a subset of 250 images chosen randomly among the 1130 images.

The obtained skin color model M spreads on a relative small portion of the color plane, showing that the skin colors, even

[2]it consists of 1130 color images collected in our **LAIV** laboratory [1], and representing an arbitrary number of people (even none) standing on an arbitrary background. The images are completely uncontrolled both regarding the acquisition conditions (very different illuminations, 5 cameras, etc.) and the face characteristics (pose, scale, expression, wearing or not glasses, partial occlusions, rotations, etc).

if corresponding to a wide range of illumination conditions and human races, are quite clustered together, making such characterization useful.

B. Skin Map determination

Given the skin color model M, we have to define a criterion to select in each image the pixels that with high probability do not correspond to skin. This allows to construct the *Skin-Map*, that is a binary image where the pixels corresponding to the skin are set to 1 and the others are set to 0. To this end, the straightforward method that consists in building the probability image and then thresholding it is not robust enough.

The method we propose does not assign the probability to single pixels but to homogeneous regions searched with the watershed technique. By doing so we segment the image into *catchment basins* to be considered as elementary units in the further steps. We adopted the algorithm proposed in [16] since it has linear computational time; it is based on the immersion simulations, and it proceeds labelling each basin with a progressive label, starting from the pixels with the darkest gray level. For this application, we applied the algorithm to the gradient of the low-pass filtered image, since we have experimentally observed that this allows to extract basins which well describe the objects in the scene. To this end, we adopted a Gaussian filter with $\sigma = 1$ and the Sobel filter.

We then characterize the catchment basins with their mean colors and use some further information, such as the basin's label and dimension, to better discriminate the basins corresponding to faces from the others. In particular, in order to find discriminant criteria, we analyzed 100 images representing generic scenes observing that:

1) face basins have generally *labels with high values*; this is due to the fact that in correspondence of faces there is frequently a high concentration of borders, that is high values in the edge images. Remembering that the watershed algorithm starts labelling the darkest gray levels, the basins corresponding to faces will have high values.
2) face basins are *small*, since in correspondence to the face there are many features which fragment the face region in many different basins.

Thus, combining these pieces of information we are able to define a robust *Skin-Map* following these steps:

- we eliminate the basins whose label values are lower than the 1% of the maximum label used in the image;
- we associate to the remaining basins the corresponding mean color in the *YPbPr* color space; we build the probability image P such that for every pixel p, $P(p)$ is its probability according to the skin color model. We threshold P putting to 1 the 5% of the highest values and thus determining a first *Skin-Map* approximation;
- finally, the area of the basins is analyzed, the idea being to eliminate the big ones. For each image, we estimate the mean μ and the standard deviation σ of the areas of the basins corresponding to the *Skin-Map*; we eliminate those basins whose area is greater than $(\mu + 5\sigma)$, under the condition that σ is high with respect to both μ and the image dimension $|Img|$, that is if the condition $[\sigma > (0.2 \cdot \mu)] \wedge [\sigma > (0.01 \cdot |Img|)]$ is verified.

These simple rules allow to reduce significantly the *Skin-Map*, simplifying the task of the SVM.

C. Experimental results

The skin detection module has been tested on 4 databases, each one with different characteristics and difficulties:

- **AR** [9]: it consists of 4000 color images corresponding to 126 people's faces (70 men and 56 women). Images feature frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). No restrictions on wear (clothes, glasses, etc.), make-up, hair style, etc. were imposed to participants. Each person participated in two sessions, separated by two weeks time.
- **XM2VTS** [6]: it consists of 1180 high quality images of single faces acquired in frontal position with homogeneous background.
- **BANCA** [4] **- Adverse:** like the **BANCA-Controlled** it consists of 2080 images, each one representing one person placed frontally to the camera and looking down, while in this database the background is non uniform. The image quality and illumination are quite poor.
- **DBLAIV:** we have already presented this database in section II-A. For this test we selected 150 images not used for the model construction and representing single faces which differ in pose, expression and scale.

In order to evaluate the *Skin-Map*, we distinguish its regions in *face regions*, and *non-face regions*, meaning that the former overlap (at least partially) the portions of the image representing faces, while the latter might correspond either to any body part other than the face, or to any portion of the image that has a mean color similar to skin. Moreover we define the logical notion of *Face-Map* as the portion of an image that corresponds to a face, that is the ground truth of the image. At this stage we do not consider the inclusion of *non-face regions* in the *Skin-Map* as an error, considering it will be the goal of the validator step to eventually discard them.

Thus, to evaluate the *Skin-Map* quality, we focus on the *face region* only. We state that an error (**false negative**) occurs if the intersection between the *Face-Map* and the *face region* does not contain at least one eye. Thus, if the *Skin-Map* has no false negative we conclude it is **correct**, no matter regarding eventual *non-face regions*. Of course it makes a big difference, with respect to the classifier task, if the *face region* corresponds exactly to the *Face-Map* or if it is either over or under estimated; for this reason we give a qualitative classification of the *face regions* according to the following notions: we consider **excellent** each *face region* which covers tightly the *Face-Map*; we say a *face region* is **good** if it covers at least either the triangle including the two eyes and the mouth or a vertical half of the face, and if its area does not exceed the double of the *Face-Map* area; in all the other correct cases the *face region* is considered **poor**. In table I the obtained results

are reported and figure 1 shows an example of a poor, a good and an excellent result.

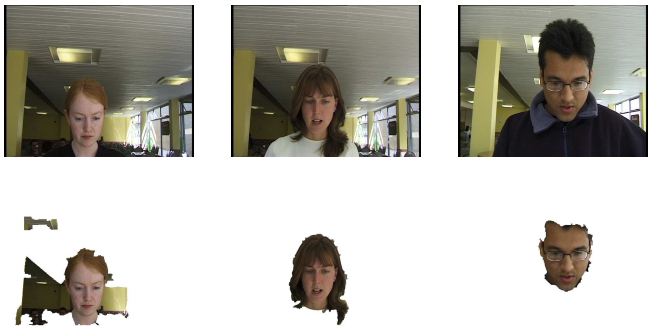

Fig. 1. Example of *Skin-Map* results (Poor, Good and Excellent respectively).

Database	False Negatives	Correct			
		Poor	Good	Excellent	Total
XM2VTS	0%	-	13.1%	86.9%	**100%**
BANCA Adverse	0.5%	0.5%	39.6%	59.4%	**99.5%**
AR	0.9%	7.5%	63.5%	28.1%	**99.1%**
DBLAIV	0.6%	3.3%	62.3%	33.3%	**99.4%**

TABLE I
SKIN DETECTION RESULTS.

As it can be expected, the best results occur with simple, high quality images (XM2VTS database) but we have very high detection rates even on images with complex background such as the BANCA-Adverse and the DBLAIV databases.

The main causes of errors are due to both the skin color model (which does not match with all the possible skin colors) and the elimination of big regions from the *Skin-Map* (which in some cases brings to deleting portions of face regions), but we have to observe that errors occur very seldom (in less than the 1% of faces both eyes are lost), bringing to a slight undermining of the successive classification step.

III. LOCALIZATION OF FACES WITH SVM

Given the *Skin-Map* of an image, we conceive a validation criterion to determine for each connected component R of the *Skin-Map* whether it constitutes a face or not. To do so, we search within it for (at least) one eye; in case of positive response we say we have localized the face, otherwise we discard the region R as non-face.

We base the validation step mainly upon the output of a statistical classifier, without taking into account any strong geometric knowledge of what constitutes a face. We mean to treat a wide set of situations that can arise in generic scenes: visibility of a single eye; faces tilted up to $|60°|$; head significantly bent down; closed eyes, subject looking away from camera objective; subject wearing transparent spectacles; subject manifesting a non-neutral expression. Besides, we have to consider that the *Skin-Map* can be over or underestimated, so weakening all the information we can gather from it.

We present in section III-A the construction of the SVM classifier, in section III-B we give a brief description of the validation technique and finally in section III-C we summarize the results obtained on several data sets.

A. Training the classifier

From now on we discard the color information restricting our attention to the spatial one. To build a SVM able to recognize eyes in a wide range of situations, we considered three different image databases to construct the training and test sets:

1) **FERET** [5] (1330 images): due to its variety and good quality it is suitable to model the general eye pattern (eyes belonging to vertical faces, both frontal and rotated up to $\pm 60°$, eyes eventually framed by transparent glasses);
2) **BANCA-Adverse** (210 images): useful to model the class of closed eyes;
3) **DBLAIV** (480 images): it includes in our classifier some knowledge about real world pictures. For instance it allows to model eyes taken from tilted faces and it enriches the class of negative examples due to the high complexity and variety of the backgrounds.

For all these images we have the coordinates of the eyes' centers, the nose tip and the mouth center as ground truths. Consequently we automatically extracted the two eyes (labelled as positive examples), six non-eye components chosen randomly out of eleven (see Fig. 2) plus four random examples from the background (or generally speaking from the complement of the eyes' bounding box). These latter ten examples are labelled as negative. The dimensions of the window used for extraction are related to the inter-eye distance d and to an estimate of the pose angle of the subject. The sample was the split into training and test set with proportion two third and one third respectively, giving rise to sets of cardinality 14842 and 7430.

In order to understand the difficulty of the classification task, we carried out two training experiments, each depending on a different representation of the same sample data:

- **Direct space representation** the sub-images have been contrast stretched and pyramidally down-sampled to the size of 16×16 pixels which is a trade off between the necessity to maintain low the computational cost and to have sufficient details to learn. We observe that this choice limits the size of regions that are candidate to be detected as faces (they must be greater than 40×40 pixels); this is a typical drawback of component based methods.
- **Wavelet space representation** we filtered the sample via an over-complete fast wavelet transform (FWT). Then we selected a small subset of significant wavelet coefficients suitable to represent the data for training purposes.

The introduction of the wavelet transform is motivated by the idea that it permits a better representation of the pattern, thus simplifying the classification task. The method we

adopted is based on the work by Papageorgiou et al. presented in [12].

The first step consists of transforming each example by a double-density one-dimensional FWT along each dimension of the image[3]:

$$c_{j-1,k} = \frac{1}{\sqrt{2}}(c_{j,2k} + c_{j,2k+1}) \quad k = 0,...,2^{j-1}-1, \quad j = 4,3,2$$

$$d_{j-1,k} = \frac{1}{\sqrt{2}}(c_{j,k} - c_{j,k+1}) \quad k = 0,...,2^{j}-2, \quad j = 4,3,2$$

which equals to perform the standard FWT and skip the downsampling step on the wavelet coefficients $d_{j,k}$. This means that instead of producing 256 coefficients per image, we generated about four times as many[4] in order to increase the variety of features to select among.

Let us call c_{j,k_1,k_2} and d_{j,k_1,k_2} the coefficients of the two-dimensional decomposition, where j is the detail level and (k_1,k_2) the two-dimensional shift. However we did not consider all the c_{j,k_1,k_2} and d_{j,k_1,k_2}; firstly we discarded the scaling coefficients c_{1,k_1,k_2} since they describe the mean illumination of the examples; secondly we discarded the detail level $j = 4$ because eye patterns are characterized by relatively small frequencies; thirdly we selected a subset of 94 coefficients which condense most of the characteristics of the pattern. The relevant features correspond to wavelets that represent vertical variations as the eye is rich in horizontal edges.

As the sample representations involved in the two experiments are quite different, we had to choose an appropriate classifier for each data set. In both cases we employed a C-SVM with Gaussian kernel, but we selected two different couples of parameters for the different classification tasks: $\gamma = 1e-3$ and $C = 10$ for the direct space representation, $\gamma = 4e-4$ and $C = 6$ for the wavelet one. Such parameters have been chosen after several trials as trade off between error reduction and generalization ability. By doing so, we obtained respectively 2060 and 1674 support vectors[5], being 3.1% and 2% the errors on the test sets. These results suggest that in the second experiment the examples are better separated, which makes the wavelet representation more suitable for the robustness of the detection technique.

B. Localization technique

The validation of the *Skin-Map* poses two major problems. First, it is necessary to reduce the number of points to consider for classification, while not excluding eye centers. Second, the absence of any assumption on the scale of faces forces the research of eyes on a range of possible dimensions. The two questions have implications both on the computational cost and on the accuracy of the technique.

[3]The non standard two-dimensional Haar basis corresponds to transforming the rows and the columns alternatively at each level j, while the standard one derives from transforming first all rows and then all columns. We prefer the non standard Haar basis because all its elements have a square support.

[4]The over-complete FWT is two times denser on each linear dimension of the image.

[5]In general a smaller number of support vector indicates an easier classification task.

Fig. 2. Facial components: 2 positive and 11 negative examples

For what concerns the first issue we restrict the scan to edges within the *Skin-Map*, according to the consideration that eyes usually lay on strong edges. In particular we limit the maximum number of points to visit in each region R by thresholding adaptively the edge image E_R.

The detail with which a region is scanned depends on its size, through the definition of the "radius" $r = \sqrt{A/\pi}$, where A is the area of R in pixels.

In fact the candidates p are exactly the points in the set $E_R \wedge G$, where G is the set of the interceptions of a grid of lines, spaced according to the scan step $s = r/k$ (being k a constant that regulates such detail). In this way we expect to consider approximately the same number of candidates independently of the region area.

Regarding the second issue, in order to evaluate a candidate point we need to extract an example at the correct scale. In other words, we expect the SVM response to be correct on eyes only if we feed in an example in which the eye dimension fits the model defined by the class of positive examples. We know by construction that the training examples have been extracted with a size based on the inter-eye distance d. On the other hand, during validation the only information that we have is the shape and size of R. Experimentally we see that if the *Skin-Map* of a face is *excellent*, then the ratio of the distance between the eyes over the radius r is about 0.6. So the natural choice for candidates extraction is the size $d = 0.6 \times r$.

However, complex pictures seldom have *Skin-Maps* very well defined. We must account for possible errors of over or underestimation of skin regions, which means to consider different possible dimensions for eyes (hypothetically) present in the region. For simplicity we extract only two additional examples of sizes $0.8 \times d$ and $1.2 \times d$, besides the optimal size d.

Let us call $\mathbf{x}_p$, $\mathbf{x}_p^-$ and $\mathbf{x}_p^+$ the examples corresponding to the same candidate p at the three different scales; we evaluate the *strength* of p by summing the margins of all three examples. Since the margin is proportional to the Euclidean distance of the example from the decision hyperplane, we treat it as a "measure" of the confidence with which the SVM classifies the example. Thus we define the function

$$f(p) = SVM(\mathbf{x}_p) + SVM(\mathbf{x}_p^-) + SVM(\mathbf{x}_p^+)$$

where $SVM(\mathbf{x}) = 0$ is the equation which defines the optimal separating hyperplane. Being the three scales quite close, we

usually observe a good correlation among the margins on positive examples, and the definition of f is useful to prevent the exclusion of a good candidate due to a wrong *Skin-Map* estimate and simultaneously to weaken the strength of a pattern that looks similar to an eye only at a certain scale.

Unfortunately the classification skills of the SVM are not sufficient to take for eyes all candidates p with $f(p)$ positive and discard all the others. What's more, we cannot foresee neither how many candidates will be placed near the eyes' centers, nor what will be the response of the SVM on them. Consequently it is necessary to determine a threshold to separate positives from false positives and a heuristic to group together all positives relative to the same eye. Our answer to the problem requires the introduction of two absolute threshold values, of opposite sign. First, we make a quick scan of the candidates only at the optimal scale d and discard all $\tilde{p}$ s.t.

$$SVM(\mathbf{x}_{\tilde{p}}) < \theta_1 < 0.$$

Second, we operate another scan in which we select all $p \neq \tilde{p}$ s.t. $f(p) \geq \theta_2 > 0$, then we aggregate them according to their mutual distance. The idea is to cluster all strong candidates closer to each other than a certain multiple m of s, and for each cluster cl we calculate its centroid c. Finally we strengthen the correctness of the position of c by adding the contribute of those candidates p s.t. $\theta_2/2 \leq f(p) < \theta_2$ (candidates which are less strong but still significant)and we say that each point c represents the center of an eye.

If we name C the set of all clusters detected by the previous technique, we can select the best two (if more than two are found) and state that they constitute the only eyes present in the image. In order to choose those clusters we coupled the elements of C in all possible ways and attributed to each couple a vote v_{ij}; if we define $v_i = \sum_{p \in cl_i} f(p)$ and call $(c_i)_y$ the y-coordinate of the centroid of the *i-th* cluster cl_i:

$$v_{ij} =_{def} \frac{v_i \cdot v_j}{1 + |(c_i)_y - (c_j)_y|}, \text{ where } i > j,\ cl_i, cl_j \in C.$$

To select the couple with maximum vote in order to localize the two eyes:

$$\text{if } v_{nm} = \max_{i,j \text{ s.t. } i>j} \{v_{ij} | cl_i, cl_j \in C\},$$
$$\text{then } c_n, c_m \text{ are the two eyes' centers.}$$

This definition means to favor cluster couples whose elements are strong (in terms of the classifier margin of their constituents) and horizontally aligned. This is reasonable because it is highly unlikely that the cluster set will contain false positives both stronger and better aligned than the two eyes.

The technique so far presented depends critically on many parameters: d, k, m, θ_1 and θ_2. Therefore we thought to use a special database to make an optimal choice for them. We selected 175 images from the BANCA-Controlled (it consists of good quality frontal portraits over a uniform background), in a way that the *Skin-Map* estimation for each image is at worst *good*. Then we exploited this ideal condition to make different experiments after which we set $k = 25$, $m = 3$, $\theta_1 = -1.5$, $\theta_2 = 1$ and $d = 0.6 \times r$.

C. Experimental results

We list here the results of face localization. Since it avoids any *ad hoc* consideration with respect to any particular database, it follows that we need to evaluate its performance over different kinds of image collections; to this end we refer to the database already presented, selecting subset of images not used to train the classifier:

- **AR** (396 images): difficult illumination conditions;
- **BANCA-Adverse** (183 images): generic background;
- **DBLAIV** (143 images) generic background, different poses and scales;
- **XM2VTS** (779 images);

.

The evaluation has been done on images whose *Skin-Map* is at least *good* (according to the definition given in II-C) to separate the errors of the validation module from the skin detection one. Tables II summarizes the performance of face localization. A point detected is 'positive' if it is closer to an eye center than $m \times s$ pixels (as defined in section III-B).

IV. Conclusions

In this paper we have presented a two step-method which automatically detects faces in still color images of generic scenes representing one subject. At first a skin detection module discards the regions whose color has very low probability to correspond to skin; this step allows to avoid an exhaustive search of the image, reducing the computational cost[6] and the number of possible false positives. Afterwards, a validation step based on a SVM classifier discriminates between faces and non faces among all the regions maintained by the previous step; the classifier is trained to recognize eyes, meaning that detecting at least one eye within a skin region determines the presence of a face. We have chosen to search for eyes instead of entire faces since it has been shown that component based methods are more robust to pose, expression and illumination variations.

We make only the assumptions that skin color is not strongly altered by artificial lights and that at least one eye per face is visible. Consequently we propose a very general method, robust to partial occlusions and to changes in pose, expression and scale. Differently from many scale-independent methods, which scan the image several times in a computationally expensive fashion, we limit our search only to three different scales and on a small subset of points, exploiting the information given by the *Skin-Map*. Regarding execution times, it takes approximately 8 seconds to compute the *Skin-Map* of a 800×600 image. After that, it takes roughly 6 seconds to detect or localize faces within the same image[7].

Face localization is a crucial first step for many applications such as face recognition, face expression analysis, face tracking, which all require the previous identification of the main

[6]In our experiments the *Skin-Map* area measures on average $1/4$ of the image area.

[7]We visited on average only 1% of the *Skin-Map* area, that is the 0.25% of the total image size. Each point candidate has a computational cost of about 5ms on a Pentium 4 with clock 3.2 GHz. Our Java code was interpreted by an optimized JRockit Java Virtual Machine.

TABLE II
FACE LOCALIZATION RESULTS

Localization ($\theta_2 = 1$)	**Eyes**			**Faces**			
Database	positives present	positives detected	false positives detected	positives present	positives detected	false positives detected	detection rate
XM2VTS	1558	1522	31	779	773	6	**99.2**%
AR	786	715	70	396	370	28	**93.4**%
BANCA Adverse	365	329	35	183	172	16	**94.0**%
DBLAIV	378	303	67	189	166	37	**87.8**%

facial features. For example Zhang and Martinez present in [18] a face recognition system in which the face localization and normalization is manually done on the AR database [9]. Many face and feature locators have been presented [8], [2], [19], each of them making some restriction on the input images. To our knowledge the most significant work has been presented by Smeraldi and Bigun [14]: they tested their application on the XM2VTS image collection [6], obtaining in 97.4% of cases the precise localization of the three main facial features (eyes and mouth), and the localization of at least two features in the 99.5% of cases. The main drawback of this method is that it is scale and pose dependent, limiting its real usability. In this paper we present a new face and eye locator that achieves very high performance on several data sets which differ in illumination, scale, pose, and quality. The results prove the system robustness and generality. Moreover in [3] we showed that, given the coordinates of at least one eye, it is possible to localize precisely the positions of both eyes, nose, mouth and chin in the 97.5% of images taken from several databases.

In order to improve the system performance, we can intervene on the two main steps separately. For what concerns the skin detector, we do not believe that including further skin samples could help: the model would cover an increasingly big portion of the color plane, resulting in a non-discriminative model. On the contrary we should deal with the limits of the method (over- and under-estimation of the skin regions) by weakening the dependence of the validator on the *Skin-Map* estimate, for example extending the range of scales to take into account or by integrating the skin information with some other invariant characteristic.

Concerning the validator, we think that the use of a local feature validator is correct, but we intend to strengthen the method by building several detectors, for instance one for eyes and one for mouths, in order to combine the judgements of the different SVMs and achieve a lower acceptance of false positives.

Experiments to evaluate the method also for the detection problem are ongoing.

REFERENCES

[1] Laboratory of Analysis of Images and Vision. *Web address: http://homes.dsi.unimi.it/ campadel/LAIV/.*

[2] J.D. Brand and J.S.D. Mason. A skin probability map and its use in face detection. *Proceedings of International Conference on Image Processing*, 2001.

[3] P. Campadelli and R. Lanzarotti. Fiducial point localization in color images of face foregrounds. *Image and Vision Computing Journal*, 22(11):863–872, 2004.

[4] The BANCA database. Web address: http://www.ee.surrey.ac.uk/Research/VSSP/banca/.

[5] The FERET Database. Web address: http://www.itl.nist.gov/iad/humanid/feret/. 2001.

[6] The XM2VTS Database. Web address: http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/. 2001.

[7] B. Heisele, P. Ho, J. Wu, and T. Poggio. Face recognition: component-based vesus global approaches. *Computer Vision and Image Understanding*, 91:6–21, 2003.

[8] O. Jesorsky, KJ Kirchberg, and RW Frischholz. Robust face detection using hausdorff distance. *Lecture Notes in Computer Science*, 2091:212 – 227, 2001.

[9] A.M. Martinez and R. Benavente. The ar face database. CVC 24, June 1998.

[10] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. John Wiley & Son, 1996.

[11] E. Osuna, R. Freund, and F. Girosi. Training support vector machines:an application to face detection. *Proceedings of International Conference on Computer Vision and Pattern Recognition, CVPR'97*, 1997.

[12] C.P. Papageorgiou and T. Poggio. Trainable pedestrian detection. In *ICIP99*, pages 35–39, 1999.

[13] R.A. Redner and H.F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26:195–239, 1984.

[14] F. Smeraldi and J. Bigun. Retinal vision applied to facial features detection and face authentication. *Pattern recognition letters*, 23:463–475, 2002.

[15] J.C. Terrillon, M.N. Shirazi, H. Fukamachi, and S. Akamatsu. Comparative performance of different skin chrominance models and chrominance for the automatic detection of human faces in color images. *Proceedings of the IEEE International conference of Face and Gesture Recognition*, pages 54–61, 2000.

[16] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991.

[17] M. Yang and Narendra Ahuja. Gaussian mixture model for human skin color and its applications in image and video databases. *SPIE Proceedings Storage and Retrieval for Image and Video Databases VII, 01/23 - 01/29/1999, San Jose, CA, USA*, pages 458–466, 1999.

[18] Y. Zhang and A.M. Martinez. Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class. *Proceedings of International Conference on Pattern Recognition (ICPR), 2004*, 2004.

[19] J. Zhou, X. Lu, D. Zhang, and C.Wu. Orientation analysis for rotated human face detection. *Image and Vision Computing*, 20:239–246, 2002.