RAIRO-Inf. Theor. Appl. 40 (2006) 123-139

DOI: 10.1051/ita:2006006

EYE LOCALIZATION FOR FACE RECOGNITION*

PAOLA CAMPADELLI¹, RAFFAELLA LANZAROTTI¹ AND GIUSEPPE LIPORI¹

Abstract. We present a novel eye localization method which can be used in face recognition applications. It is based on two SVM classifiers which localize the eyes at different resolution levels exploiting the Haar wavelet representation of the images. We present an extensive analysis of its performance on images of very different public databases, showing very good results.

Mathematics Subject Classification. 68T10, 68T45.

1. Introduction

Researches on automatic face recognition (FR) started in the seventies, but had a great impulse in the last decade thanks to the advance in technologies. The interests for this topic are both scientific and applicative: the former because FR is a very challenging pattern recognition problem; the latter for its numerous real-world applications such as human/computer interface, surveillance, and secure access. Indeed, the usefulness of biometric systems is evident and FR has the advantage of being non-intrusive, unlike other biometric systems, such as fingerprint analysis or retinal and iris scan.

A general statement of the face recognition problem can be formulated as follows: given still or video images of a scene, identify one or more persons using a stored database of faces.

In the last two decades hundreds of papers dealing with FR have been presented; Zhao and others reported in [44] both a fundamental survey of the main algorithms,

© EDP Sciences 2006

Keywords and phrases. Eye localization, face recognition, Haar wavelets, support vector machines.

 $^{^{}st}$ Work partially supported by the PASCAL Network of Excellence under EC Grant no. 506778. This publication only reflects the authors view.

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

and a performance analysis of the most efficient methods that work on still images. The evaluation has been done adopting a common protocol [30] and referring to a standard database[14].

This comparison gave two main contributions: first it identified a set of methods which outperformed the others; second, it clearly identified the conditions in which the existing algorithms achieve a certain level of success, thus indicating the areas which require further research.

The best performing algorithms were: the Principal Component Analysis (PCA) [35], a combined Principal Component Analysis and Linear Discriminant Analysis (LDA) [43], a Intrapersonal/extrapersonal Image Difference Classifier (IIDC) [26], and the Elastic Bunch Graph Matching (EBGM) [27]. Since that comparison, these methods were referred to as standard baselines¹.

Regarding the research open problems, the authors observed that even the best algorithms behave well only on images representing frontal faces, with neutral expression, acquired under tightly controlled conditions of illumination, and uniform background. The performance decreases drastically when even one of these conditions is not satisfied. Moreover all the methods need an accurate facial features (eyes, mouth, and nose) localization to normalize the faces to a common scale, and to align them precisely. Without accurate localization of the facial features, robust face recognition cannot be achieved.

Even recent papers [1,6–8,16,24,31,40–42], which tackle some of the critical aspects cited above, do not deal with the feature selection problem. Either implicitly or explicitly, they refer for their initialization to manual annotations of the facial features, leaving open the issue of their individuation or taking off-the-shelf solutions. This problem is still indicated as a crucial one [7,33].

Recently some research works [4, 10, 17, 19, 21, 34, 38] have devoted particular attention to the problem of eye detection and localization since

- eye appearance is less variant with respect to the other facial features, in case of pose and expression changes. This allows to define a more robust pattern to look for;
- eye detection validates the output of a face detector. In fact, even a good
 face detector is subject to a certain false acceptance rate and the presence
 or absence of eyes is a good criterion to confirm or reject the face detection
 output;
- if the detected eyes are precisely localized, the position of the other features can be easily determined.

Besides, eye localization methods also concern other research fields such as gaze tracking for man/machine interaction tasks [22], augmenting the importance of developing a robust, precise and efficient technique.

The method we propose is hierarchical and consists of three subsequent stages: at first a face detector approximately localizes the position and extension of the

¹An efficient implementation and a software environment have been developed [3] to allow researchers to rigorously compare their new algorithms with the standard ones.

faces represented in the image; then an eye detector processes the detected subimages to roughly individuate the eye positions; finally an eye localizer is applied to the found positions to further refine the localization precision. Here we focus on the latter two steps, which can be applied in cascade to any face detector [28,32,37] that returns a sort of bounding box delimiting each face candidate in the image.

The construction of the eye detector and the eye localizer is similar: both of them are built as a statistical classifier trained on examples represented via an overcomplete wavelet decomposition of their respective target patterns. For this reason we describe the eye detector in detail and then we shortly account for the specificity of the eye localizer.

The article is organized as follows: in Section 2 we face the problem of selecting the right features for the eye pattern representation in order to make it more suitable for classification. Section 3 presents the definition and construction of the Support Vector Machine (SVM) used for binary classification. Sections 4 and 5 describe respectively the eye detection and the eye localization modules. Finally, in Section 6 we report the performance of our technique on standard databases (XM2VTS, Banca, FRGC, and FERET) and we compare it with the best performing algorithms. This paper includes and significantly extends the work we published in [5].

2. Feature selection

The difficulty intrinsic to the task of pattern recognition requires an accurate choice of a suitable representation of the eye pattern. In [5] we have observed that the wavelet representation is more favorable than the direct representation as it leads to a smaller generalization error. Indeed, Haar wavelets permit to abstract the low level features (the information represented by the gray level values of the pixel) and allow to describe the pattern in terms of luminance changes at different frequencies, at different positions and along different orientations. Moreover, Haar coefficients are much faster to compute than other wavelet coefficients.

Ideally the set of coefficients resulting from the analysis of the pattern can be thought of as composed of two parts: an informative part that condenses most of the eye shape, plus a non-informative part that contains noise and irrelevant details. The goal of feature selection is to separate the two parts as much as possible in order to keep only the relevant features. This yields a better representation of the pattern because it reduces the dimensionality of the problem by removing the non-informative coefficients (realizing a sort of denoising of the data) and results in an easier separation of the positive examples (pattern) from the negative ones (non pattern).

Our feature selection procedure was inspired by the work of Papageorgiou et al. [29]. The size of our pattern examples is 16×16 pixels: such a dimension represents a trade off between the necessity to maintain low the computational cost and to have sufficient details to learn the pattern shape. All the wavelets have square support and are generated via the application of the one-dimensional



FIGURE 1. The shape of the bi-dimensional Haar wavelet coefficients. From left to right: horizontal, vertical, diagonal.

FWT (Fast Wavelet Transform)

by alternating the transform along the rows and the columns of the image for each level j of the wavelet decomposition (the level j=4 corresponds to the scaling coefficients of the highest detail level, that is to the original pixels of the image). This is a modified version of the original FWT because it skips the subsampling step over the wavelet coefficients $d_{j,k}$. In fact it is called a *denser* or *overcomplete* transform as it produces four times as many coefficients with respect to the original one. We desire this redundancy because we want to increase the variety of features among which to select the most significant.

A bi-dimensional wavelet coefficient $d^o_{j,k1,k2}$ is identified by four parameters: j is called the detail level and relates to the size of the window over which the coefficient is calculated (hence it regulates the frequency); (k1,k2) is called the shift and relates to the position of the coefficient within the image; $o \in \{horizontal, vertical, diagonal\}$ determines the orientation of the edge that is tested for presence. In Figure 1 we show the shape of the wavelet coefficients under consideration.

The coefficients $c_{j,k1,k2}$ are called the scaling coefficients and describe the mean illumination of the image; we discard them producing a sort of illumination normalization of the pattern examples.

In order to carry out the selection process, we need a way to assess the relative importance of the coefficients left. This is done via a normalization step: we take a set L of eye pattern images (|L| = 2152 in our case) and we decompose each $l \in L$ in its wavelet coefficients $d^o_{j,k_1,k_2}(l)$. For each coefficient d^o_{j,k_1,k_2} we calculate its mean value in the sample of patterns:

$$\overline{d^{o}}_{j,k_{1},k_{2}} = \frac{\sum_{l=1}^{|L|} |d^{o}_{j,k_{1},k_{2}}(l)|}{|L|}$$

and we normalize it with respect to the average mean of its band (i.e. of its detail level); if we call B_j the band (the set) of all d_{j,k_1,k_2}^o of level j then

$$\tilde{d}^{o}_{j,k_{1},k_{2}} = \frac{\overline{d}^{o}_{j,k_{1},k_{2}}}{m_{j}}, \text{ where } m_{j} = \frac{\sum_{k_{1}} \sum_{k_{2}} \sum_{o} \overline{d}^{o}_{j,k_{1},k_{2}}}{|B_{j}|},$$

represent the normalized coefficients that can be now ordered to assess their relative importance. Notice that the normalization is done here within the entire bands and not within each sub-band (a subset of a band whose coefficients have all the same orientation).

Since the expected value of the sum of all $\tilde{d}^{o}_{j,k_1,k_2}$ in the same band is approximately equal to their cardinality²

$$E\left[\sum_{k_1} \sum_{k_2} \sum_{o} \tilde{d^o}_{j,k_1,k_2}\right] = \sum_{k_1} \sum_{k_2} \sum_{o} E\left[\frac{\overline{d^o}_{j,k_1,k_2}}{m_j}\right]$$

$$\approx \frac{\sum_{k_1} \sum_{k_2} \sum_{o} E[\overline{d^o}_{j,k_1,k_2}] \cdot |B_j|}{E\left[\sum_{k_1} \sum_{k_2} \sum_{o} \overline{d^o}_{j,k_1,k_2}\right]} = |B_j|$$

the normalized coefficients $\tilde{d}^o{}_{j,k_1,k_2}>0$ can be interpreted as follows:

$$\tilde{d^o}_{j,k_1,k_2} \left\{ \begin{array}{ll} \sim 1 & \Rightarrow \text{no regularity} \\ \ll 1 & \Rightarrow \text{systematic uniformity} \\ \gg 1 & \Rightarrow \text{systematic variation.} \end{array} \right.$$

Hence the normalization allows us to distinguish two sub-categories of coefficients that can be ordered separately: C^+ , the coefficients that are systematically greater than 1, and C^- , those which are systematically smaller than 1. Both of them retain precious information because they respectively represent the edges of the pattern that are systematically strong or systematically absent.

The first feature selection is done by not considering the detail level j=3 because eye patterns are characterized by relatively small frequencies (filters with long support), while the d_{3,k_1,k_2}^o correspond to Haar wavelets of support 2×2 pixels; this first step reduces the number of wavelet coefficients from 849 to 174.

Once separated the 174 normalized coefficients in the ordered sets C^+ and C^- , we define an error function to drive the selection process. We can measure the expressive power of the coefficients by measuring how well they reconstruct the pattern that they represent. We choose as distance metrics the summation of the absolute value of the pixelwise difference. The coefficients in C^+ are meant to characterize the eye shape, thus we calculate the distance between the mean eye pattern and the pattern reconstructed by incrementally adding more and more

²The equality does not hold strictly because the random variables $\overline{d^o}_{j,k_1,k_2}$ and m_j are not independent. However they can be assumed to be so, as their correlation is very low $(m_j$ depends on the sum of many $\overline{d^o}_{j,k_1,k_2}$).

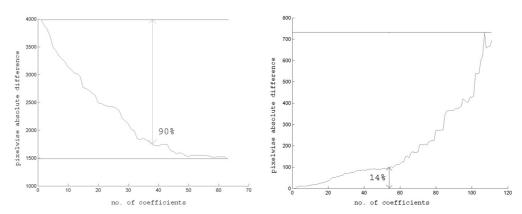


FIGURE 2. Feature selection of coefficients within C^+ and C^- .

coefficients according to their ordering (from the biggest to the smallest). The normalized coefficients in C^- are evaluated against a uniform pattern because they represent the absence of detail (now the ordering is from the smallest to the biggest).

In Figure 2 we plot the error functions as we vary the number of features maintained. By looking at the trend of the curves we devise an empirical selection principle: regarding C^+ we want to be able to reconstruct at least 90% of the pattern information while not including uninformative coefficients, therefore we stop at position 38 (out of 63) just before a "plateau" of the error function (the following coefficients would not add crucial information). Analogously, for what concerns C^- we intend to exclude about 90% of spurious information meanwhile stopping the selection just after a plateau; hence we keep 54 coefficients out of 111 (they account for 14% of the difference from the uniform pattern).

Figure 3 visualizes the decomposition of the original pattern and its reconstruction by only considering the selected features.

3. SVM EYE CLASSIFIER

The Support Vector Machines are the state-of-the-art model for many classification tasks. They can be used as a powerful statistical tool for supervised learning of visual patterns and generate strong classifiers with good discriminative capacity and high accuracy rate [9,36]. We put ourselves in the context of binary classification where we intend to design and train machines that separate a positive class from a negative one.

In order to train an SVM we collect a set of labelled example $T = \{(\mathbf{x}_i, y_i) | i = 1, ..., l\}$ where the label $y_i \in \{-1, +1\}$ determines the class of the example \mathbf{x}_i . Each example is formed by the selected wavelet coefficients extracted from the corresponding image. We intend to build a classifier that shows good generalization

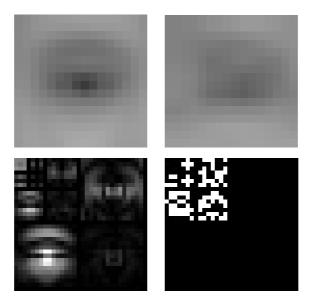


FIGURE 3. Upper left: the original pattern (16×16) ; lower left: its decomposition (32×32) ; upper right: the reconstructed pattern (16×16) ; lower right: the selected coefficients (32×32) .

properties without making any assumption on the probability distribution that generates the training set.

An SVM binary classifier is a linear machine because it is defined as the hyperplane that optimally separates the two classes of examples. It is optimal in the sense that it maximizes the separation margin of the classes (the distance between the hyperplane and the examples closest to it) by solving the dual quadratic program

$$\max_{\alpha} \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j$$
s.t.
$$\sum_{i=1}^{l} y_i \alpha_i = 0$$

$$0 \le \alpha_i \le C, \quad i = 1, \dots l.$$
(1)

The parameter C regulates the "hardness" of the separating surface, that is the number of training example that can be misclassified in the training phase: by increasing the value of C the machine accepts fewer and fewer exceptions to the decision rule (if C is too big there might not exist a separating hyperplane).

The function $K(\mathbf{x}_i, \mathbf{x}_j)$ is called the *kernel* function and is a measure of similarity of the training examples. Some kernel functions induce a non linear mapping from the example space to a highly dimensional space where the linear separation can be actually achieved. One function of this type is the Gaussian kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2}$$

that has been successfully used in many techniques devoted to solving difficult learning tasks. It is defined in terms of the parameter $\gamma = \frac{1}{2\sigma^2}$ that regulates the width of the Gaussian functions used for the similarity measurement (σ is the standard deviation of the Gaussian distribution).

The decision function is defined as follows

$$f(\mathbf{x}) = \operatorname{sign}\left(SVM(\mathbf{x})\right) = \operatorname{sign}\left(\sum_{\mathbf{x}_i \in SV} y_i \alpha_i^{opt} K(\mathbf{x}, \mathbf{x}_j) + b^{opt}\right)$$
(2)

where $SVM(\mathbf{x}) = 0$ is the equation of the separating hyperplane. It is important to notice that f only depends on those vectors \mathbf{x}_i that, when projected in the high dimensional space, lie closest to the decision surface (the only examples for which $\alpha_i \neq 0$). They represent the most difficult examples to learn and are called Support Vectors just because they alone support the construction of the decision function.

Apart from the choice of the kernel function, the machine definition is customized to the specific problem by tuning all its free parameters. The generalization skill of the classifier is then estimated by measuring the classification error on a test set that is disjoint from the training set, but that has been generated according to the same distribution. One suitable measure to evaluate the generalization capacity of the machine is the product of the recall times the precision³. These quantities are defined as follows

$$precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}$$

where TP = true positives, FN = false negatives, FP = false positives. The choice for the best parameters is done by choosing the combination of γ and C that maximizes $precision \times recall$.

The training was carried out on a training set of 8936 examples and a test set of 4357 examples, given as input to the SVMlight optimization algorithm [23]. The examples have been extracted from a subset of the BANCA database (to model the frontal face features under different illumination conditions) and from a custom database containing many heterogenous and uncontrolled pictures of various people (useful to model pose variations and non-neutral face expressions). The positive class contains eye images cropped to a size equal to the inter-ocular distance. The negative class is populated by the other facial features (nose, mouth, chin, cheeks, forehead, etc.) and by some examples extracted from the background of the images.

Figure 4 shows the trend of the generalization performance as we vary the parameters that define the machine; the best choice is done for C=11 and $\gamma=4\times 10^{-4}$ which corresponds to an error on the test set lower than 3%.

³These quantities relate to the *soundness* and the *completeness* of the classification method. In fact the *precision* estimates the probability that an example classified as positive is really an eye, while the *recall* estimates the probability that, given an eye, it is classified as a positive.

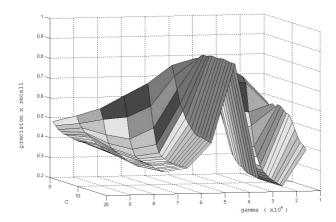


FIGURE 4. Parameters tuning: best generalization for C=11, $\gamma=4\times 10^{-4}$.

4. Detection technique

The detection of the eye features proceeds hierarchically from the detection of the face position, since a first extraction of the face candidate regions avoids an extensive eye search in the whole image. It is like doing an attention selective process which quickly discards the irrelevant background information. We do this step using our implementation of the Viola Jones face detector [37] or our skin map detector [5] for color images. Besides giving the position of the face, both detectors roughly estimate its extension (scale) on the image with a certain error distribution. This uncertainty makes it particularly unfeasible to rely on geometric relations in order to identify the eye region [34], but the range of the scale variation is not so wide to require a complete multi-resolution search of the eye pattern. We deal with this variability by considering a range of three scales fed in a strong classifier for eye detection.

The evaluation of a candidate point P comes to evaluating three examples centered in it: the one at the inferred scale (example \mathbf{x}_P), plus two examples extracted according to a small underestimation (80%) and a small overestimation (120%) of that scale (examples \mathbf{x}_P^- and \mathbf{x}_P^+).

We weaken the standard decision function (2) because the sign of the classification is not sufficient to obtain a robust detection. Since the margin of the hyperplane is proportional to the Euclidean distance of the support vectors from the decision function, we treat $SVM(\mathbf{x})$ as a "measure" of the confidence with which the SVM classifies the example \mathbf{x} . Thus we define the function

$$\rho(P) = SVM(\mathbf{x}_P) + SVM(\mathbf{x}_P^-) + SVM(\mathbf{x}_P^+)$$

as the strength of the candidate point P. Being the three scales quite close, we usually observe a good correlation among the margins on positive examples, and the definition of ρ is useful to prevent the exclusion of a good candidate due to a

wrong estimate of the face dimension. It also serves to weaken the strength of a pattern that looks similar to an eye only at a certain scale.

We proceed by evaluating $\rho(P)$ over a small subset of points in the face region: first we identify the points that lie on the edges, then we subsample them with a step that depends on the scale of the face region; we consider as point candidates the ones for which $\rho(P)$ is greater than 0, and we group them according to their proximity in the image; each group of point candidates is then represented by its centroid (the eye candidate) obtained weighting each point P with its $\rho(P)$. This last step strengthens the eye detection, making it more stable. Ideally we should have two eye candidates, however, sometimes it happens that the eye classifier detects also one or more false positives. In presence of false detections (if more than two eye candidates are present) we introduce a final selection criterion that exploits the margin of the classifier and assumes the substantial verticality of the face pose. Doing so, we manage to select the eye positions, and to discard the false detections, by choosing the candidates couple (c_i, c_j) that maximizes

$$\frac{SVM(c_i) \cdot SVM(c_j)}{1 + \sqrt{|(c_i)_y - (c_j)_y|}}$$

where $(c_i)_y$ is the y coordinate of the eye candidate c_i . As we do not want to enforce the perfect verticality of the face, the square root at denominator is introduced to give more importance to the strength of the eye candidates with respect to their horizontal alignment.

5. Localization technique

While the eye detector must distinguish the global eye shape from that of other facial patterns, the eye localizer is intended to exhibit a high sensitivity to the center of eyes, that is we want its response to peak sharply at the eye center and rapidly fade away within its neighborhood.

The construction of the localizer directly follows from these considerations. First of all, the examples are represented by the same overcomplete wavelet transform but this time we do not discard the highest detail level (j=3) as we are particularly interested in the high resolutive power of the localizer. After applying the same feature selection procedure of Section 3 we retain 380 coefficients out of 849 (197 in C^+ , 183 in C^-). Secondly, the eye localizer works on a smaller receptive field: it is presented with patterns that are half size of the eye patterns defined for the eye detector. Thirdly, the training examples are taken from the same image sample of the eye detector but with an important difference: the negative class is now populated with sub-images obtained by small, random displacements of the positive examples (in the ratio of one positive to ten negatives).

The eye localizer can be used in cascade to the eye detector in order to improve the precision of the eye localization. If correctly initialized, the eye localizer really manages to recover the detection. The initialization concerns two parameters: the eye position and scale. In the previous section we have described how we calculate

Database	# of images	% with $d_{eye} \leq 0.25$
XM2VTS	1180	99.0%
BANCA	416	96.9%
FRGC Controlled	473	99.6%
FRGC Uncontrolled	396	94.8%
FERET	1000	96.4%

Table 1. Eye detection results.

the eye positions. Regarding the eye scale we can infer quite a stable estimate by calculating the average of the scales of \mathbf{x}_P , \mathbf{x}_P^- and \mathbf{x}_P^+ weighted by their margins of classification ρ . The refinement procedure performed by the localizer proceeds analogously as in the eye detection module: the candidates that give the highest response are grouped together and averaged according to their margin to compute their centroid.

6. Experimental results

In order to assess the accuracy of both the eye detection and localization module we need to adopt a suitable protocol. Jesorsky and others [21] proposed to use the normalized measure

$$d_{eye} = \frac{\max(||C_l - \tilde{C}_l||, ||C_r - \tilde{C}_r||)}{||C_l - C_r||}$$

where the values $\tilde{C}_{r/l}$ stand for the positions output by the localization method, while the values $C_{r/l}$ are the ground truth of the right and left eye center respectively. This measure is scale independent and therefore it permits to compare data sets characterized by different resolution of the face regions. There is an agreement that $d_{eye} \leq 0.25$ is a good criterion to identify correct eye detection, that is to flag their presence [21, 25, 45]. Such a level of precision is considered sufficient for recognition algorithms which do not require a precise alignment, but they need an initialization step to work locally on the found regions [2].

In the following table we give the rate of the sole eye detector for various public databases (see appendix for description). We manually annotated ourselves the ground truth of the BANCA and FRGC databases, while for the XM2VTS and FERET we referred to the public ones available at the web pages [14, 20].

Concerning the XM2VTS and BANCA databases the highest eye detection rates have been reported by [19]; comparing their results with ours we observe that on the XM2VTS databases, the performance are approximately the same (99%), while on the BANCA database they reach the 88% of success against our 96.9%. Regarding the FERET and FRGC databases, to our knowledge the

Database	# of images	% with $d_{eye} \leq 0.1$
XM2VTS	1180	94.7%
BANCA	416	91.9%
FRGC Controlled	473	94.1%
FRGC Uncontrolled	396	85.8%
FERET	1000	89.5%

Table 2. Eye localization results after refinement.

best results have been reported in [39], which are globally worse than ours (they give a detection of the 94.5%).

However when we speak of eye localization the threshold $d_{eye} \leq 0.25$ is not sufficient: here we consider the threshold value $d_{eye} \leq 0.1$ as a reasonable accuracy (10% of the inter-ocular distance). In Table 2 we report the localization results after the refinement phase, that is after the application of the eye localizer initialized on the positions output by the eye detector.

The comparison between these results and those obtained in [19] shows that we have a greater percentage of success on both the XM2VTS and BANCA databases, being significantly better on the second one (91.9% against 72%).

The work described in [38] does not adopt the same error metrics as we do and it reports results only on the FRGC. They measure the normalized mean error (not the maximum, which is a worst case analysis) and give an error of 2.67% on the entire FRGC. By adopting this measure on our experiments, we observe an error of 2.65% and 3.88% on the controlled and uncontrolled images of our experiments (respectively). If we consider that the entire FRGC contains twice as many controlled images as uncontrolled ones, we see that these results are very similar.

The work [34] adopts the same evaluation criteria ($d_{eye} \leq 0.1$ and $d_{eye} \leq 0.25$) and reports good results on the BioID database [12]: 98.1% and 91.8% respectively. In the future we plan to try our method on this dataset as well, however here we stress the fact that our technique behaves equally good on a large variety of databases and it actually works also on images with complex background and on faces with certain expression and pose variations. Moreover we observe that the generalization capabilities of the statistical classifier allow to treat problematic situations like the presence of transparent or semi-transparent spectacles, or the eye closure.

In Figure 5 we show some outputs of our algorithm on a sample of custom images (here we use as face detector our skin color detector [5]). As we show in the figure, the algorithm is quite robust to moderate rotations in and out of the plane, as well as to a certain overestimation of the face extension; moreover it is quite insensitive to face expressions as they usually do not affect the eye patterns.



FIGURE 5. Some outputs of the eye localization algorithm. Boxes: detected faces; circles: detected eyes; crosses: localized eyes.

7. Summary and conclusions

We have presented an eye localization algorithm which assumes as input a rough localization of the face in the image. One or two SVM can be used as a function of the accuracy required by the further steps of the face recognition task. The first SVM localizes the eyes with a rough precision and, when an higher precision is required, a second classifier with higher discriminative power can be made to work in a small surround of the points found by the first one.

Higher precision, however, has a cost; indeed with our implementation, which does not focus on speed, the first SVM takes approximately 3 s on face regions 300×350 pixel wide on a Pentium 4, 2.8 GHz while the second takes 4 times more.

Experimental results show that, at both considered levels of accuracy, our method localizes the eyes with a detection rate that is higher than the current state of the art on most of the standard public databases.

Further research will concentrate on reducing the computational time and improving the results of the second SVM.

APPENDIX A: DATABASE DESCRIPTION

In this appendix we briefly describe the public databases used to test our method, specifying for each of them which images we selected for tests.

- **XM2VTS** [15]: it consists of 1180 high quality images of single faces acquired in frontal position, and with homogeneous background; some of the subjects wear spectacles. The tests refer to the whole database;
- BANCA [11]: it consists of three sub-databases referred to as Controlled, Adverse and Degraded. In this work we do not take into account the third one, since the images are particularly blurred, making the step of precise eye localization useless. The other two are: Controlled: it consists of 2080 images each one representing one person placed frontally to the camera and on a uniform background. The database collects pictures of 52 people of different ethnic groups (Caucasian, Indians, Japanese, Africans, South-Americans), acquired in 4 different sections (10 images per subject in each section). The illumination conditions vary from daylight to underexposed, while no evident light chromatic alteration is present. 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 not very good.

We built the BANCA test set taking for each subject the first image of each section, that is 416 images.

• FRGC [13] The FRGC databases collects 5658 high resolution images of 275 subjects in frontal position. The images are organized in *subject*

session, where each section consists of 4 images acquired in controlled conditions (uniform background and homogeneous illumination) and two in uncontrolled conditions (generic background and varying illumination conditions). In both situations half of the images represent subjects with neutral expression and half smiling. The number of sections varies from subject to subject, between 1 and 7. We built two different test sets, corresponding to the controlled and uncontrolled conditions. The first contains 473 images, and is obtained taking for each subject the first image of the first two sections (when the second is present). The set of uncontrolled images is built according to the same criterion, and contains 396 images.

• **FERET** [14] The FERET database consists of 10 gray level images per person organized according to the angle between the subjects and the camera (0°, ±15°, ±25°, ±40°, ±60°), and where two sets of frontal view images, respectively with neutral and smiling expression, are included. Our FERET test set consists of 1000 images selected randomly from the images with a rotation up to ±15°.

REFERENCES

- [1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla and T. Darrell, Face recognition with image sets using manifold density divergence, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 05)* (2005).
- [2] S. Arca, P. Campadelli and R. Lanzarotti, A face recognition system based on automatically determined facial fiducial points. *Pattern Recognition* (2006) 432–443.
- [3] J.R. Beveridge, D. Bolme, B.A. Draper and M. Teixeira, The CSU face identification evaluation system. its purpose, features, and structure. *Machine vision and applications* **16** (2005) 198–138
- [4] P. Campadelli and R. Lanzarotti, Fiducial point localization in color images of face fore-grounds. *Image Vision Comput. J.* 22 (2004) 863–872.
- [5] P. Campadelli, R. Lanzarotti and G. Lipori, Face localization in color images with complex background, in *Proceedings of the IEEE International Workshop on Computer Architecture for Machine Perception (CAMP 2005)*, Palermo, Italy (2005) 243–248.
- [6] H. Cevikalp, M. Neamtu, M. Wilkes and A. Barkaba, Discriminative common vectors for face recognition. IEEE Trans. Pattern Anal. Machine Intell. 27 (2005) 4–13.
- [7] K.I. Chang, K.W. Bowyer and P.J. Flynn, An evaluation of multimodal 2d+3d face biometrics. IEEE Trans. Pattern Anal. Machine Intell. 27 (2005) 619–624.
- [8] C. Liu, Gabor-based kernel PCA with fractional power polynomial models for face recognition. IEEE Trans. Pattern Anal. Machine Intell. 26 (2004) 572-581.
- [9] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines. Cambridge University Press (2000).
- [10] D. Cristinacce and T.F. Cootes, A comparison of shape constrained facial feature detectors. Proceedings International conference on Automatic Face and Gesture Recognition (2004).
- [11] The BANCA Database. http://www.ee.surrey.ac.uk/Research/VSSP/banca/.
- [12] The BioID Database. http://www.humanscan.de/support/downloads/facedb.php.
- [13] The Face Recogniton Grand Challenge Database. http://www.frvt.org/FRGC/.
- [14] The FERET Database. http://www.itl.nist.gov/iad/humanid/feret/.
- [15] The XM2VTS Database. http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/.

- [16] G. Givens, J.R. Beveridge, B.A. Draper, P. Grother and P.J. Phillips, How features of the human face affect recognition: a statistical comparison of three face recognition algorithms, in *Proceedings of IEEE on Computer Vision and Pattern Recognition* (2004).
- [17] Y. Gizatdinova and V. Surakka, Feature-based detection of facial landmarks from neutral and expressive facial images. IEEE Trans. Pattern Anal. Machine Intell. 28 (2006) 135–139.
- [18] V. Bruce, F. Fogelman Soulie, H. Wechsler, P.J. Phillips and T. Huang Eds. Face Recognition: from theory to applications. Springer-Verlag (1998).
- [19] M. Hamouz, J. Kittler, J.K. Kamarainen, P. Paalanen, H. Kälviäinen and J. Matas. Feature-based affine invariant localization of faces. *IEEE Trans. Pattern Anal. Machine Intell.* 27 (2005) 1490–1495.
- [20] INRIA. http://www-prima.inrialpes.fr/FGnet/data/07-XM2VTS/.
- [21] O. Jesorsky, K.J. Kirchberg and R.W. Frischholz, Robust face detection using the hausdorff distance. Lect. Notes Comput. Sci. 2091 (2001) 212–227.
- [22] Q. Ji, H. Wechsler, A. Duchowski and M. Flickner, Special issue: eye detection and tracking. Comput. Vision Image Understanding 98 (2005) 1–3.
- [23] T. Joachims, SVMlight package implementation. http://svmlight.joachims.org/.
- [24] J. Kim, J. Choi, J. Yi and M. Turk, Effective representation using ica for face recognition robust to local distortion and partial occlusion. *IEEE Trans. Pattern Anal. Machine Intell.* 27 (2005) 1977–1981.
- [25] Y. Ma, X. Ding, Z. Wang and N. Wang, Robust precise eye location under probabilistic framework, in *Proceedings of IEEE International Conference on Automatic Face and Ges*ture Recognition (FGR 04) (2004).
- [26] B. Moghaddam, C. Nastar and A. Pentland, A bayesian similarity measure for direct image matching, in *International Conference on Pattern Recognition*, Vienna, Austria (1996).
- [27] K. Okada, J. Steffens, T. Maurer, H. Hong E. Elagin, H. Neven and C. von der Malsburg, The bochum/usc face recognition system and how it fared in the feret phase iii test. In H. Wechsler and Huang [18] 186–205.
- [28] M. Osadchy, M.L. Miller and Y. LeCun, Synergistic face detection and pose estimation with energy-based models (2005) 1017–1024.
- [29] C.P. Papageorgiou and T. Poggio, Trainable pedestrian detection, in Proceedings of International Conference on Image Processing (1999) 35–39.
- [30] J. Phillips, H. Moon, S.A. Rizvi and P.J. Rauss, The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Machine Intell.* 22.
- [31] S. Romdhani and T. Vetter, Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 05) (2005).
- [32] H. Schneiderman and T. Kanade, Object detection using the statistic of parts. Int. J. Comput. Vision 56 (2004) 151–177.
- [33] S. Shan, Y. Chang, W. Gao and B. Cao, Curse of mis-alignment in face recognition: problem and a novel mis-alignment learning solution. *International Conference on Automatic Face* and Gesture Recognition (FG04) (2004) 314–320.
- [34] X. Tang, Z. Ou, T. Su, H. Sun and P. Zhao, Robust precise eye location by adaboost and svm techniques, in *Proceedings of the International Symposium on Neural Networks* (2005) 93–98.
- [35] M. Turker and A. Pentland, Face recognition using eigenfaces. J. Cognitive Neuroscience 3 (1991).
- [36] Vapnik, The nature of statistical learning theory. Springer (1995).
- [37] P. Viola and M. Jones, Robust real time object detection. Inter. J. Comput. Vision 57 (2004) 137–154.
- [38] P. Wang, M. Green, Q. Ji and J. Wayman, Automatic eye detection and its validation, in *Proceedings of the Workshop FRGC in the IEEE conference on Computer Vision and Pattern Recognition* (2005).
- [39] P. Wang and Q. Ji, Learning discriminant features for multi-view face and eye detection, in Proceedings of IEEE on Computer Vision and Pattern Recognition (2005).

- [40] X. Wang and X. Tang, A unified framework for subspace face recognition. IEEE Trans. Pattern Anal. Machine Intell. 26 (2004) 1222–1228.
- [41] B. Zhang, X. Chen, S. Shan and W. Gao, Nonlinear face recognition based on maximum average margin criterion, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 05)* (2005).
- [42] Y. Zhang and A.M. Martinez, Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class, in *Proceedings of International Conference on Pattern Recognition (ICPR 2004)* (2004).
- [43] W. Zhao, R. Chellappa and A. Krishnaswamy, Discriminant analysis of principal components for face recognition. In H. Wechsler and Huang [18] 73–85.
- [44] W. Zhao, R. Chellappa, P.J. Phillips and A.Rosenfeld, Face recognition: A literature survey. ACM, Computing Surveys 35 (2003) 399–458.
- [45] Z.H. Zhou and X. Geng, Projection functions for eye detection. Pattern Recognition Journal 37 (2004) 1049–1056.

To access this journal online: www.edpsciences.org