Face and facial feature localization

Paola Campadelli*, Raffaella Lanzarotti**, Giuseppe Lipori, and Eleonora Salvi

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

Abstract. In this paper we present a general technique for face and facial feature localization in 2D color images with arbitrary background. In a previous work we studied an eye localization module, while here we focus on mouth localization. Given in input an image that depicts a sole person, first we exploit the color information to limit the search area to candidate mouth regions, then we determine the exact mouth position by means of a SVM trained for the purpose. This component-based approach achieves the localization of both the faces and the corresponding facial features, being robust to partial occlusions, pose, scale and illumination variations. We report the results of the separate modules of the single feature classifiers and their combination on images of several public databases.

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

1 Introduction

Face localization[1] is a crucial first step for many applications such as face recognition, face expression analysis, and face tracking; moreover all these applications require the identification of the main facial features (eyes, mouth, nose) either to normalize the image or to further process them.

Both face and feature localization are challenging because of the face/feautures 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).

These tasks are often solved making some restrictions on the input images (uniform background, fixed scale and pose, etc.), and thus restricting the domain of applicability of the system [1], [7], [14]. Even more, sometimes the localization is accomplished in

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

** Work partially supported by the PASCAL Network of Excellence under EC grant no. 506778. This publication only reflects the author view.

[1] We speak about *face localization* when input images depict only one subject in the foreground and about *face detection* when no assumption is made regarding the number of faces in the images.

a manual or semi-manual way. For example Zhang and Martinez presented in [13] a face recognition system in which the face localization and normalization is manually done on the AR database [8]. To our knowledge the most significant face localization work has been presented by Smeraldi and Bigun [10]: they tested their application on the XM2VTS image collection [5], 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.

We designed a system for face and facial feature detection in color images consisting of two modules which try to exploit respectively the advantages of feature invariant approaches [11, 12] and appearance-based methods [6, 9]. The first module searches for skin regions within the image exploiting their peculiar colors and further information that helps to characterize faces. This step determines a *Skin-Map* which represents the restricted search area to be referred by the subsequent steps. In the second module, two different SVMs (trained to recognize eyes and mouths respectively) are applied only in correspondence to the skin regions with the objectives of both discriminating between faces and non faces and to localize the eyes and the mouth, if any. In particular, if at least one facial feature is localized within a skin region we validate it as a face. Of course, the detection of one feature can be enough to validate a skin region, but our final objective is to detect all three if they are all visible. In fact we observe that a good result would be to detect two features, since they would allow to determine the face scale and even to foresee the position of the lacking one; the third feature could be looked for in a subsequent step exploiting the knowledge given by the first two.

In [2] we presented the skin detection module, and the validation step based on the eye SVM only. At that stage we obtained high performance on images of very high quality, like those in the XM2VTS; on images with complex background, and which differ in illumination, scale, pose, and quality we obtained about 90.9% of face detection, with 125 false positives with respect to 783 eyes present.

In this paper we focus on the mouth localization (section 2), and we discuss (section 3) on how the conjunction of the eye and mouth SVM outputs can help in several directions such as to rise the detection rate, to achieve a lower acceptance of false positives and to deal with feature occlusions.

2 Mouth localization with SVM

The localization step is mainly based upon the output of a statistical classifier, without taking into account any strong geometric knowledge of what constitutes a face. The only *a-priori* knowledge we exploit regards the peculiar color of mouths: to greatly reduce the search area to give in input to the classifier, we select those sub-regions within the *Skin-Maps* which show the peculiar mouth chromaticity (see section 2.1).

By searching for the mouth, we account for certain problematic situations that can happen in generic scenes, as occlusions of features other than the mouth or significant rotations of the head around the three axis. In fact rotations may greatly modify the face pattern while leaving substantially untouched the mouth appearance[2]. Moreover our component based approach allows to treat some basic non-neutral expression, as anger, happiness and so on. We included in our treatment both mouth closed (neutral or angry) and slightly open (as if reading or smiling), as specified in the section 2.2, where we present the construction of the SVM classifier. In section 2.3 we give a brief description of the localization technique and finally in section 2.4 we summarize the results obtained on several data sets.

2.1 Mouth-Map

As we mentioned in the introduction, for each image we determined a *Skin-Map* (Fig. 2) which consists of one or more connected regions. Being focused on the mouth research, we reduce the search area furthermore on the basis of the peculiar mouth chromaticity. To this end we first apply a transformation to the pixels within the *Skin-Map* in order to simplify the segmentation task. Such transformation refers to the chrominance components in the *YPbPr* color space[3], and produces a *Mouth-Map* in which the mouth pixels assume high values:

$$\textit{Mouth-Map} = Pr^2 \cdot (255 - (Pr - Pb^2))$$

The first factor of the formula exploits the consideration that usually mouth pixels have high values in the *Pr* plane; the second factor penalizes skin pixels which also have high *Pr* values but low *Pb* ones in comparison to the mouth pixels.

The *Mouth-Map* is thresholded considering each *Skin* region separately. Experimentally we observed that the mouth area corresponds roughly to the 3% of the face area and thus we selected in each skin region the 3% highest *Mouth-Map* values (Fig. 3).

On the 2147 images of the test set (for details see section 2.4) the *Mouth-Map* usually contains at least one region in correspondence to the mouth; such regions are always strictly included in the portion of the image depicting the mouth. When the mouths are open, the lower and upper lips are very often disconnected, and the lower one is generally better defined, being the lower lip thicker on most subjects.

2.2 Training the classifier

The objective of gathering a rich and representative training sample is accomplished by considering different image databases, each contributing to the definition of a specific aspect of the mouth pattern:

1. 585 images from the **XM2VTS** [5]: model the general, well-defined mouth pattern;

[2] As specified in section 2.2, the SVM has some knowledge of examples extracted from portraits of people whose head is rotated to some extent along any degree of freedom.

[3] *YPbPr* color space is obtained from the RGB components by means of a linear transformation which allows to separate the luminance *Y* from the chrominance *Pb* and *Pr*. Comparing several color spaces we retain this one since it has shown to be the most adapt to highlight both the skin and the mouth regions.

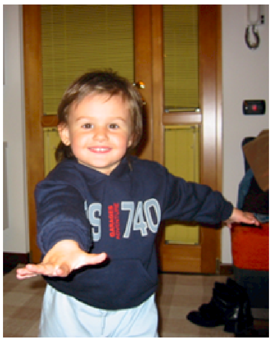

Fig. 1. the input image

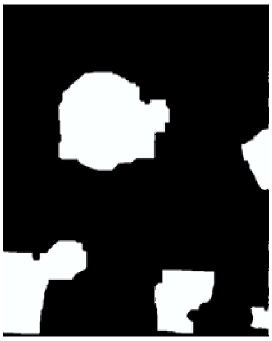

Fig. 2. The *Skin-Map*

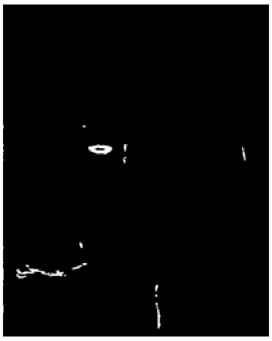

Fig. 3. The *Mouth-Map*

2. 525 images from the **AR** [8]: contain portraits under poor illumination conditions; we chose four sessions for each subject, referring to four different situations: neutral, anger, and smiling expressions, and non uniform illumination;
3. 890 images from the **Color FERET** [4]: suitable to model mouth belonging to faces rotated from $-45°$ up to $45°$ around the vertical axis;
4. 208 images from the **BANCA-Adverse** [3]: useful to model the class of mouths taken from people who are reading (hence bending down and slightly opening the mouth);
5. 480 images from the **DBLAIV**: this selection helps to include in our classifier some knowledge about real world pictures. For instance it allows to model mouths taken from tilted faces[4] and it enriches the class of negative examples due to the high complexity and variety of the backgrounds.

The mouth classifier is based exclusively on the intensity information of the patterns, therefore all these images have been converted to gray scale prior to example extraction. For each image we dispose of manually placed ground truth: the coordinates of the eyes' centers, the nose tip and the four corners of the mouth (see Fig. 4). The sample is built by extracting from each picture the mouth (labelled as positive example), four non-mouth components chosen randomly out of twelve (see Fig. 5) plus three random examples taken from the background (or generally speaking from the complement of the mouth bounding box). These latter seven examples are labelled as negatives. The dimensions of the window used for extraction are related to the mouth width. We centered the mouth pattern on the lower lip for two main reasons. Firstly, if we consider the exact mouth center we would experience a greater variability of the pattern appearance due to the high variability of mouth expressions. On the contrary our positive examples show good uniformity in the lower half of the pattern. The second reason comes from the properties of the *Mouth-Map* that, especially if the mouth is not tightly closed, tends to

[4] This subset of the DBLAIV contains faces whose *tilt* angle (defined as the angle between the vertical axis of the head and the horizontal plane) varies continuously from $-60°$ to $60°$, with mean around $0°$ and standard deviation $15°$.

be more accurate on the lower lip. However we did not include in our sample all mouth examples, since we wish to exclude from our model the subcase of open mouth[5]. By considering the distribution of the ratio width/height of each mouth, we observe that the vast majority of the examples we wish to treat falls above the value 2. This means that in general a mouth closed or slightly open is at least twice wider than tall. Hence, following our intentions, we characterized the positive class by discarding mouths whose ratio is under 2. This elimination step leaved us with 2353 positive examples.

The whole sample was then split into training and test set with proportion two third and one third respectively. This procedure gave rise to sets of cardinality 13340 and 6677. After the extraction, all sub-images have been contrast stretched and pyramidally down-sampled to the size of 16×16 pixels.

As in our previous work on eyes [2], we studied the positive class in terms of wavelet coefficients in order to reduce the number of features to consider, while retaining the essential pattern information, thus simplifying the training task. The feature selection process saved 95 wavelet coefficients for each example.

We trained a 1-norm soft margin SVM with Gaussian kernel and $\gamma = \frac{1}{2\sigma^2} = 4 \times 10^{-4}$, $C = 9$. Such parameters have been chosen as trade off between error reduction and generalization ability. By doing so, we selected a machine based on 1476 support vectors and achieving the 2.3% error on the test set. These results show a very accurate learning of the pattern by the SVM and make it suitable for the robustness of the following step.

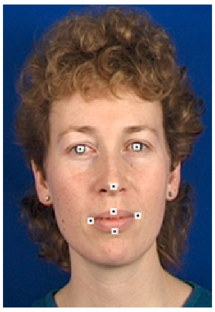

Fig. 4. The arrows indicate the ground truth

$\Longrightarrow$

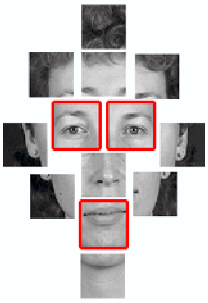

Fig. 5. Facial features

2.3 Localization technique

The localization technique searches for mouths within the *Skin-Map*, with the idea of classifying as ‘face’ the region that corresponds to the detected mouth.

The mouth research poses two major problems. First, it is necessary to reduce the number of points to consider for classification, while not excluding the ones corresponding to the mouth. Second, the absence of any assumption on the scale of the face

[5] In a preliminary experiment we trained a SVM on all mouth examples, but we obtained very poor results, meaning that the patter was too rich.

forces the mouth research on a range of possible dimensions. The two questions have implications both on the computational cost and on the accuracy of the technique.

Regarding the first issue, we have partially answered to it introducing the *Mouth-Map*, which reduces the research to the 3% of the *Skin-Map* area. A further reduction consists in selecting a proper set of candidates; to this end, we consider each region within the *Mouth-Map* separately: we determine its bounding box, and we enlarge it in the lower part of the 50% of its height[6]. Afterwards, on the basis of the bounding box width w, we determine the scan step: the candidates are the points included in the enlarged bounding box that correspond to the interceptions of a grid of lines, spaced horizontally and vertically according to the scan steps $s_h = w/(5+k)$ and $s_v = w/(3+k)$ respectively[7], being k a constant that regulates such detail (in our experiments $k = 4$).

In order to evaluate each candidate point, we need to extract an example at a scale which fits the mouth model used to train the SVM. To this end, we infer the size of the examples on the basis of both the *Skin-* and *Mouth-Map*; these two contributes are quite independent and allow to determine a reliable approximation. To make it even more robust, we account for possible errors of over or underestimation of them, which means to consider different possible dimensions for mouths (hypothetically) present in the region. For simplicity, besides the optimal size d, we extract only two additional examples of sizes $(0.7 \times d)$ and $(1.3 \times d)$.

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. This interpretation is not standard; usually the only output considered is the class label the SVM attributes to the example fed in, which corresponds to the sign of its margin. However, 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 $SVM(\mathbf{x}) = 0$ 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 candidate due to a wrong *Skin-* or *Mouth-Map* estimate and simultaneously to weaken the strength of a pattern that looks similar to a mouth only at a certain scale.

Finally, given all the $f(p)$, we localize the mouth in two steps: at first we compute the score of each region adding all the $f(p)$ corresponding to it; the region with the highest score is identified as the mouth; afterwards, the mouth position is determined computing the centroid of all the points corresponding to the validated mouth region whose $f(p)$ is positive.

2.4 Experimental results

We list here the results of face and feature localization; the experiments have been carried out on the same databases that we used to build the training sample but on disjoint

[6] This choice is driven by the fact that we trained the classifier to recognize the lower lip.

[7] We adopted two different scan steps for the horizontal and vertical position selections since we observed a higher sensitivity of the SVM to horizontal translations than to vertical ones.

image sets. Table 1 shows both the localization rate of eyes and mouth separately, and their combination which gives a better estimate of the overall behavior.

We observe that in the 10% of the data set one or more features are not available for classification, due to one of the following reasons: the incompleteness of the *Skin-Map* (if it does not cover all the features present); mouth occlusion by moustache or beard; eye occlusion because of strong rotations. This lack prevents to reach full localization of the three features on all images (the results in the second last column reflect this fact).

Since the loss of either one eye or the mouth can be easily recovered on the basis of the other two, in the last column we show the results obtained by relaxing the goal to the localization of at least two features.

Table 1. Face localization results

Localization results		**Eyes**		**Mouth**		**Face**	
Database	number of images	positive rate	false positives	positive rate	false positives	all three features	at most one feature missing
XM2VTS	583	97.9%	20	91.1%	15	87.8%	**99.0**%
AR	479	92.5%	75	86.6%	41	71.4%	**91.0**%
FERET Color	689	93.6%	163	88.3%	24	60.5%	**90.6**%
DBLAIV	189	80.2%	67	84.0%	19	62.4%	**84.1**%
BANCA Adverse	207	85.2%	58	86.3%	18	68.1%	**82.6**%

On the basis of the obtained results, we can conclude that the task of locating mouths seems more difficult than the one of detecting eyes; we think this is due to the fact that the eye pattern is more structured than the one associated to the lower lip, making easier and more robust the corresponding classification. Finally, we observe that if we consider the localization of just one feature as sufficient to validate the presence and position of the face, than we reach almost 100% of face localization on all databases; we notice that the addition of the mouth detector has increased of about the 10% the overall performance, considering that, using the eye detector only, we reached the 90% of face localization [2].

3 Conclusions

In this paper we presented a module for mouth localization based on a SVM classifier trained to recognize closed or slightly open mouths. This module is a part of a more general face and facial feature localization system which first localizes skin regions (on the basis of the peculiar skin color), and then searches for eyes and mouths. Such system is robust to partial occlusions and to changes in pose, expression and scale.

The results obtained prove that our system performs well: on the high quality images of the XM2VTS we obtained performances comparable to the ones presented in [10], while being more general and pose and scale-independent; on images which differ in illumination, scale, pose, quality and background we obtained good performance proving the generality and robustness of the system.

Differently from many scale-independent methods [6], which scan the image several times, we limit our search only to three different scales and on a small subset of points, exploiting the information given by color. Regarding execution times, our Java method, run on a Pentium 4 with clock 3.2 GHz, takes approximately 5*ms* for each candidate point, bringing to a mean time of roughly 6 seconds to localize the facial features in images of 800×600 pixels.

In order to achieve a lower acceptance of false positives and a more precise feature localization, we intend to add a further module which takes into account the outputs of the different SVMs. Moreover we are working on the training of a classifier which will recognize open mouths; this will allow to apply the system to a even larger domain of images.

References

1. 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.
2. P. Campadelli, R. Lanzarotti, and G. Lipori. Face localization in color images with complex background. *Proceedings of the IEEE International Workshop on Computer Architecture for Machine Perception (CAMP 2005), Palermo, Italy. To appear*, 2005.
3. The BANCA database. Web address: http://www.ee.surrey.ac.uk/Research/VSSP/banca/.
4. The FERET Database. Web address: http://www.itl.nist.gov/iad/humanid/feret/. 2001.
5. The XM2VTS Database. Web address:http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/. 2001.
6. 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.
7. O. Jesorsky, KJ Kirchberg, and RW Frischholz. Robust face detection using hausdorff distance. *Lecture Notes in Computer Science*, 2091:212 – 227, 2001.
8. A.M. Martinez and R. Benavente. The ar face database. CVC 24, June 1998.
9. 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.
10. F. Smeraldi and J. Bigun. Retinal vision applied to facial features detection and face authentication. *Pattern recognition letters*, 23:463–475, 2002.
11. 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.
12. 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.
13. 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.
14. J. Zhou, X. Lu, D. Zhang, and C.Wu. Orientation analysis for rotated human face detection. *Image and Vision Computing*, 20:239–246, 2002.