Neural Network Cascade for Facial Feature Localization

Thibaud Senechal, Lionel Prevost, and Shehzad Muhammad Hanif

Universite Pierre and Marie Curie-Paris 6
ISIR, CNRS UMR7222, BC173
4 Place Jussieu, 75252 Paris Cedex 5, France

Abstract. We present here a complete system for the localization of facial features in frontal face images. In the first step, face detection is performed using Viola & Jones state of art algorithm. Then, a cascade of neural networks localizes precisely 28 facial features. The first network performs a coarse detection of three areas in the image corresponding roughly to left and right eyes and mouths. Then, three local networks localize, in these areas, 9 key points per eye and 10 key points on the mouth. Thorough experiments on 3500 images from standard databases (Feret, BioID) show the detector accuracy, its generalization ability and speed.

1 Introduction

Localizing facial features (like mouth and eye corners or eyebrow) is usually the first step in applications like face recognition, expression analysis or action unit identification [1,2]. These key-points are also very useful for model alignment.

Active shape model [3] [4] and active appearance model [5] are commonly used to perform the detection. Unfortunately, they rely on an unstable optimization procedure which depends on hundreds of parameters encoding shape (and texture) variations. Other statistical methods include Neural Networks [6][7], Bayesian Networks [8], Support Vector Machines [9] and Cascade Of Boosted Ensembles (using either Haar filter [10] or Gabor jet [11]). Though most of these algorithms are able to detect precisely a small number of facial features (typically four, including eye centres, mouth and nose), their accuracy on a large number of key-points is rarely stated.

In our previous works (described in [7]), manually cropped images are fed to a neural network trained to output a probability map. Facial feature hypothetic locations (eye centers and mouth corners) corresponded to local maxima in this map.

We present in this paper a fast and an accurate method for precise facial feature localization. A facial detector is used to extract the face in the image. Then, a neural network performs coarse detection, defining 3 regions of interest (left and right eyes and mouth) within the face image. In the second stage, another network is applied on each region to detect 28 points (mouth contour,

F. Schwenker and N. El Gayar (Eds.): ANNPR 2010, LNAI 5998, pp. 141–148, 2010.

eye contour and eyebrow). Such a cascade was already explored in [12] but they only detected 10 points.

The paper is organized as follows. Section 2 details the 3 stages: face detection, coarse to fine facial feature localization. Section 3 is devoted to sensitivity analysis and experimental results on several benchmark datasets. Conclusion and prospects are presented in section 4.

2 Overview

2.1 Face Detector

To extract automatically face in images, we use the OpenCV's face detector which provides an implementation of the Viola-Jones algorithm [13]. It uses Haar-like filters as weak classifiers. The AdaBoost algorithm makes a forward selection of the best features and trains the weak learners. To run in real-time, strong classifiers are arranged in a cascade in order of complexity, where each classifier is trained only on examples which pass through the previous classifiers.

2.2 Coarse Localization

This step detects three Regions Of Interest (ROI) corresponding roughly to eyes and mouth in the detection window. To achieve, we train a fully connected multilayer perceptron using back-propagation algorithm to detect five points: eye centers, nose tip and mouth corners.

Neural network inputs can be either the gray levels of sub-sampled extracted faces. To improve the robustness of our detector, we synthesized new face images by translating the detection and modifying the scale factor of the face detector by 10% of the inter-ocular distance (this corresponds to the standard deviation of the eye position in the detection image). To obtain face images invariant to illumination effects, the images are normalized. Statistical mean and standard

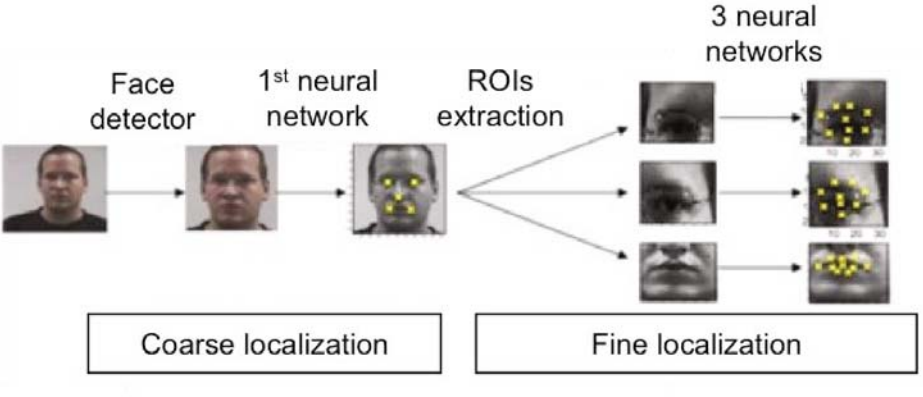

Fig. 1. System overview

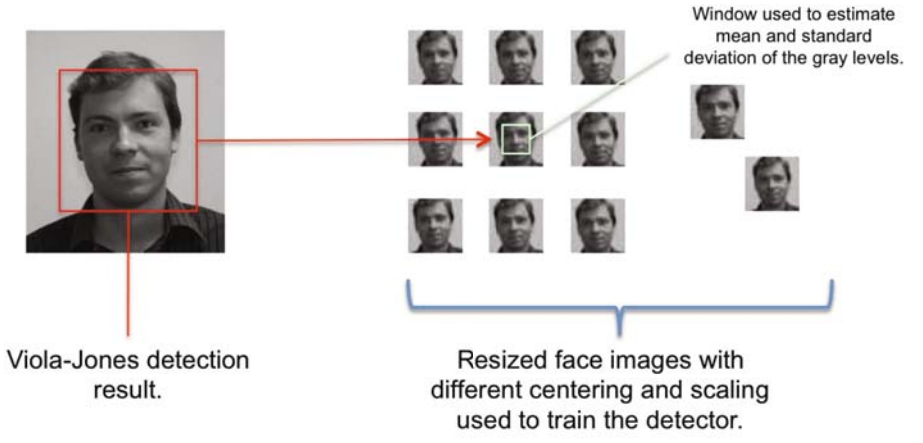

Fig. 2. Train database

deviation are estimated on a window centered on the center of the detection window with a size equal to the half of the detection window (fig. 2. The 10 outputs are the (X, Y) coordinates normalized between -1 and 1 of the five facial features we try lo localize.

2.3 Precise Localization

In this last step, three fully connected multilayer perceptrons are used to deal with eyes and mouth area. They detect 28 points: 9 for each eye and eyebrow contours and 10 for the two lip contours (fig. 1). The eye localizers have 9x2=18 outputs and the mouth localizer has 20 outputs.

Coarse localization determines three ROIs bounding the eye centres and the two mouth corners. We chose the size of these ROIs as a fraction of the detection window size to obtain images including all the points that we try to localize. We made a statistical study on face pattern coordinates in detection images to find the smallest window that contains these points in most of the cases. Finally, to compensate the imprecision of the ROI centre coordinates, we increased ROI dimensions by the mean localization error of the coarse localizer. As before (Section 2.2), to train these three neural networks, we synthesized new images.

3 Setup and Sensitivity Analysis

In order to estimate the parameters of the detector, we manually labeled the ground truth of 320 images from a homemade database (so called ISIR database) containing frontal faces with small expression and natural rotation changes of men and women with different specificities (facial hair, glasses). Training dataset includes 256 peoples and the other 64 images are used to stop training. The

size of the training set is artificially expanded to 3700 examples using basic image transformations like small translation, rotation and scaling. For each set of parameters, we performed 3-fold cross-validation. The localization error E for an example is the mean Euclidian distance between the detected feature positions (x_i, y_i) and the true (labelled) feature positions $(\tilde{x}_i, \tilde{x}_i)$, normalized with respect to the inter-ocular distance D. The mean localization error E_m is computed over the whole dataset.

$$E = \frac{1}{N} \sum_{i=1}^{N} \frac{\sqrt{(x_i - \tilde{x}_i)^2 + (y_i - \tilde{y}_i)^2}}{D} \tag{1}$$

3.1 Coarse Localization

We evaluated several image coding schemes: gray-level sub-sampled images of size 20x20 and 30x30 pixels and principal component analysis on images of size 20x20, 30x30 and 60x60 pixels resulting into 70, 90 and 100 eigenvalues respectively (corresponding to 90% of the explained variance). Mean localization results for the 5 points are reported table 1. We get quite similar results using gray level 30x30 images (900 input cells) and eigenvalues (100 input cells) and 100 hidden cells. In both cases, the mean localization error E_{mv} on the validation set is lower than 6% for the five points we try to localize.

Table 1. Mean localization error on training set (E_{mt}) and validation set (E_{mv}) for five points with the coarse detector

Inputs	Number of input cells	E_{mt}	E_{mv}
20x20 images	400	6.0%	7.1%
ACP on 20x20 images	70	8.9%	9.2%
30x30 images	900	4.3%	5.9%
ACP on 30x30 images	90	6.0%	6.9%
ACP on 60x60 images	100	4.3%	5.8%

3.2 Precise Eye Localization

We want to localize 4 points on each eyebrow and 5 points on each eye as shown in fig 1. We have already defined the sizes and position of the ROI in section 2.3. We evaluated several resolutions (6x10, 10x15, 13x20, 16x25, 19x29, 22x34 pixels) for precise eye detection.

Lowest resolution (6x10 pixels) results in a high localization error (higher than 5%) on the training set while the highest resolution (22x34 pixels) generalizes poorly. Best results correspond to the following parameters: 13x20=260 input cells and 10 hidden neurons. The mean localization error $E_m v$ on these 18 feature points is 4.8%.

3.3 Precise Mouth Localization

We have evaluated two sizes for the mouth region of interest: small size (including mouth only) and large size (including the mouth and some parts of nose and chin). For each size, three image resolutions were tested. Small ROIs lead to 4.5% as mean localization error while large regions give better results: the mean localization error is 4%. The sensitivity to resolution is quite low. Best results correspond to the following parameters: 21x23 input cells and 20 hidden neurons. The mean localization error E_mv on these 10 feature points is 4%.

4 Experimental Results

We trained and tune the parameters of the cascade localizer on the ISIR dataset divided into independent training and cross-validation sets. Then, to evaluate the localizer generalization ability, we tested it extensively, without any retraining on several benchmark databases.

4.1 Results on the ISIR Database

Table 2 details performances (mean localization error) for all the face patterns that we want to localize. Center and contour of the eye are localized with a high precision (the mean error is less than 3%). This is partially due to the face detector that produces eye-centered detections. Mouth features are precisely localized too (the mean error is equal to 4%) though mouth position is more variable. Precision on eyebrow is a little poorer (6.4%) but performance evaluation is biased due to ground truth imprecision. The percentage of correctly detected images with a localization error lower than 10% is 98% for the eye and eyebrow regions and 93% for the mouth contour (fig. 3).

Table 2. Mean localization error on training set (E_{mt}) and validation set (E_{mv}) for different facial features compared to the standard deviation of their position in detection image (SD)

	SD	E_{mt}	E_{mv}
Mouth (10 points)	10.5%	3.3%	4.0%
Eye centers (2 points)	6.4%	2.7%	3.0%
Eyes contour (8 points)	6.9%	3.1%	3.3%
Eyebrows (8 points)	9.4%	5.7%	6.4%
28 facial points	8.9%	3.8%	4.4%

Table 3 compares test error of each step of our architecture for 4 points (eyes and mouth corners). First column reports the localization error on eye centers and mouth corners using only the Viola-Jones detector: we use as localization hypothesis the mean position of each feature (estimated on the training dataset) inside the detection windows. Second column reports the mean error after the

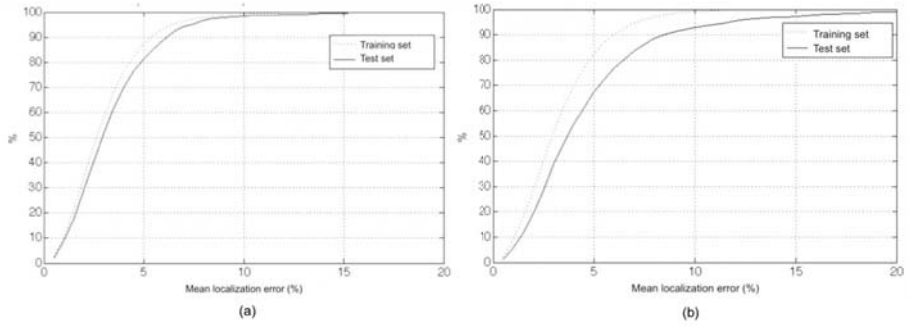

Fig. 3. Percentage of images with a mean localization error lower to the X axis value for eyes points (a) and mouth points (b)

coarse localization. The third column reports the mean error for the same points after the fine localization step. This proves that reducing the research area with a coarse step allows us to increase locally the image resolution without having too many input cell. Moreover it shows that cascading localizers increases drastically the system accuracy.

Table 3. Mean localization error on validation set after each step of the system

	VJ detector	Coarse localization	Fine localization
Eye centers (2 points)	6.4%	4.8%	3.0%
Mouth corners (2 points)	10.5%	7.2%	3.7%

4.2 Results on the Inrialpes, BioID and Feret Test Databases

The system has been successively evaluated on the 60 frontal faces from Inrialpes database, 1500 images from BioID database [14] and the 1918 images FERET Duplicate I dataset [15] without retraining the neural networks. (fig 4) shows some results we obtain on these bases.

Inrialpes database includes frontal faces with poor luminosity conditions and the system performs a mean localization error of 7% (fig 4b).

On BioID database that contains large changes in expressions and has very poor luminosity conditions (emphasis has been laid on "real world" conditions), the system gives a mean localization error of 11% on 28 facial features (fig 4c).

Finally, Feret database contains multi-ethnic subjects with small facial expression changes and some subjects wear glasses. We only have the eyes, mouth centres and the nose tip manually labelled. Although the learning database that we used only includes European subjects without glasses, the mean localization error on eyes and mouth centre is lower than 6% (fig 4a).

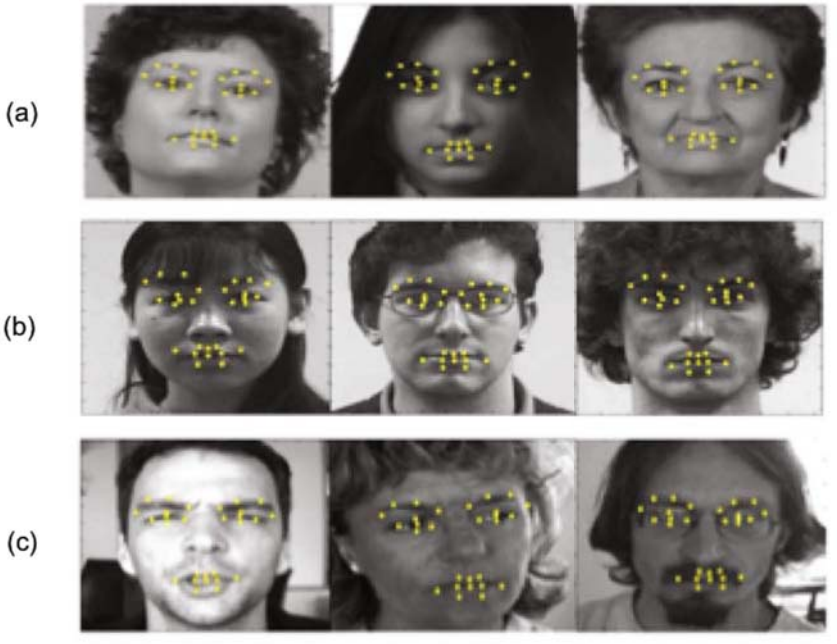

Fig. 4. Localisations on Feret database (a), Inrialpes database (b) and BioID database (c)

5 Conclusion and Prospects

We have presented here a new localizer able to detect precisely the eyes, eyebrows and mouth contours. It uses a cascade of neural networks. The face is first detected using a standard algorithm. The first stage performs a coarse detection of three regions of interest corresponding roughly to eyes and mouth. The second stage localizes precisely 28 facial features on the eye contour, the eyebrow and the mouth contour. The mean localization error is lower than 5% on the validation set. To show the generalization ability, we evaluate the system on three standard databases, namely Inrialpes, bioId and Feret, where faces are sometime slightly expressive, multi-ethnic or poorly illuminated. Results are really encouraging as the overall localization error is lower than 8%. Moreover, the computation speed (including face detection, coarse and fine localization) is nearly 20 images per second. In our previous works [7], we already showed the system ability to localize coarsely facial features in orientation-free images by combining several experts dedicated to each facial pose. So, we can easily combine parallel and cascade approaches to build an orientation-free fine localizer. Other future works include active appearance model initialization and action unit detection for facial expression labelling.

References

1. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys 35(4), 399–458 (2003)
2. Pantic, M., Rothkrantz, L.: Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1424–1445 (2000)
3. Cootes, T., Taylor, C., Cooper, D., Graham, J., et al.: Active shape models-their training and application. Computer Vision and Image Understanding 61(1), 38–59 (1995)
4. Milborrow, S.: Locating Facial Features with Active Shape Models. PhD thesis, Faculty of Engineering, University of Cape Town (2007)
5. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision 60(2), 135–164 (2004)
6. Duffner, S., Garcia, C.: A connexionist approach for robust and precise facial feature detection in complex scenes. In: Image and Signal Processing and Analysis, pp. 316–321 (2005)
7. Hanif, S.M., Prevost, L., Belaroussi, R., Milgram, M.: Real-time facial feature localization by combining space displacement neural networks. Pattern Recognition Letters 29(8), 1094–1104 (2008)
8. Yan, S., Li, M., Zhang, H., Cheng, Q.: Ranking prior likelihood distributions for bayesian shape localization framework. In: International Conference on Computer Vision, pp. 51–58 (2003)
9. Nguyen, M., Perez, J., De la Torre Frade, F.: Facial feature detection with optimal pixel reduction svms. In: International Conference on Automatic Face and Gesture Recognition (2008)
10. Cristinacce, D., Cootes, T.: Facial feature detection using adaboost with shape constraints. In: British Machine Vision Conference, vol. 1, pp. 231–240 (2003)
11. Vukadinovic, D., Pantic, M.: Fully automatic facial feature point detection using gabor feature based boosted classifiers. In: International Conference on Systems, Man and Cybernetics, vol. 2 (2005)
12. Duffner, S., Garcia, C.: A hierarchical approach for precise facial feature detection. In: Compression et Representation des Signaux Audiovisuels (2005)
13. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
14. Jesorsky, O., Kirchberg, K., Frischholz, R., et al.: Robust face detection using the hausdorff distance. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 90–95. Springer, Heidelberg (2001)
15. Phillips, P., Wechsler, H., Huang, J., Rauss, P.: The feret database and evaluation procedure for face-recognition algorithms. Image and Vision Computing 16(5), 295–306 (1998)