Multiple Neural Networks for Facial Feature Localization in Orientation-Free Face Images

Lionel Prevost, Rachid Belaroussi, and Maurice Milgram

Université Pierre et Marie Curie-Paris 6, EA2385 PRC
BC252 4, Place Jussieu, F-75005 France
{lionel.prevost, rachid.belaroussi, maurice.milgram}@upmc.fr

Abstract. We present in this paper a new facial feature localizer. It uses a kind of auto-associative neural network trained to localize specific facial features (like eyes and mouth corners) in orientation-free faces. One possible extension is presented where several specialized detectors are trained to deal with each face orientation. To select the best localization hypothesis, we combine radiometric and probabilistic information. The method is quite fast and accurate. The mean localization error (estimated on more than 700 test images) is lower than 9%.

1 Introduction

Automatic facial feature detection is becoming a very important task in applications such as model-based video coding, facial image animation, face recognition, facial emotion recognition, visual speech understanding, and intelligent human-computer interaction. Many face recognition systems are based on facial features, such as eyes, nose and mouth, and their spatial relationship, called the constituted approach [3]. Many feature detection methods have been developed in the last decade, but a wide majority concentrates on eye detection. The existing methods can be divided into several categories. A first classification is based on the acquisition device: active infrared-based approaches [13] and passive image-based approaches. Another one depends on the processed images: pre-focused images where rough feature regions have already been located or cluttered images where face detection is proceeded before feature detection. A third category is based on the detection algorithm: image-based approach using one or several low-level detectors to find specific properties (such as edge, colour, symmetry…) [7, 10, 11], statistical appearance-based approach [12], active appearance models [4], deformable templates [15]…

We present in this paper a neural-based facial feature localizer able to deal with orientation-free face images. As we already developed in our lab a face localizer [1], we assume that face has been already roughly localized in a cluttered image. The system uses a kind of auto-associative neural network trained to output a feature map, which maxima correspond to facial feature position.

The communication is organized as follows. In section 2, we describe the database used in the experiments. Section 3 is devoted to the hybrid auto-associative network used to localize facial features. In section 4, we study experimentally this orientation-free localizer and propose an alternate method where several networks are trained to

F. Schwenker and S. Marinai (Eds.): ANNPR 2006, LNAI 4087, pp. 188 – 197, 2006.

deal with specific face pose, in order to increase the system accuracy. Concluding remarks and future works are discussed in section 6.

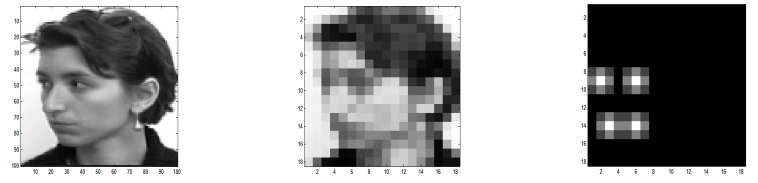
2 Database and Pre-processings

We collected in our lab a face database. It contains images of 40 people with various ages, genders and ethnicities. For each person, we took 36 images (resolution 100x100 pixels) with several facial orientations, expressions and "accessories" like beard or glasses (Fig. 9). In order to increase the number of data, we computed each mirroring image. This procedure results in a 2750 example dataset.

We clicked manually four facial features, respectively left eye (1st feature), right eye (2nd feature), left mouth corner (3rd feature) and right mouth corner (4th feature) to create one feature map F for each face image. This feature map had the size of the face image and its pixels have the following value (where x_{iT} and y_{iT} denote the true feature coordinates):

- At the feature location: $F(x_{iT},y_{iT}) = +1$
- Anywhere else: $F(i,j) = -1$

To normalize input images (Fig. 1), we performed histogram equalization. To normalize feature maps, we convolved these images with a 3x3 gaussian filter, which results in smoothing feature maps. Several sub-sampling were tested to reduce the data dimension and, thus the number of parameters to be trained.



(a) (b) (c)

Fig. 1. Normalization process: original image (a), sub-sampled input image (b), sub-sampled and smoothed feature map (c)

Facial feature are not randomly organized (except in Picasso's paintings perhaps). So, we can get anthropomorphic information about their spatial organization by analyzing feature map. Assuming the feature coordinates joint density distribution is gaussian, we can evaluate its parameters (means and covariance matrix) by using *Maximum a Posteriori* estimator. Assuming this density is monovariate, this estimation can be done on the whole dataset and leads to orientation-free parameters. To take into account the face orientation, we assume that feature density distribution is a mixture of gaussians, one for each face orientation. In this latter case, we estimate parameters on a given cluster. To perform self-supervised orientation clustering, we assumed there existed a unique relationship between 2D facial feature location and 3D face pose. So, knowing the facial feature localization allowed predicting the face

orientation. We used a simple K-means algorithm [2] with euclidean distance to get the best center of each subset. Then, we estimated parameters for each subset. We applied this procedure considering up to six face orientations. As can be seen (Fig. 2), the clustering had roughly separated the whole database in subsets, each one corresponding to a certain orientation.

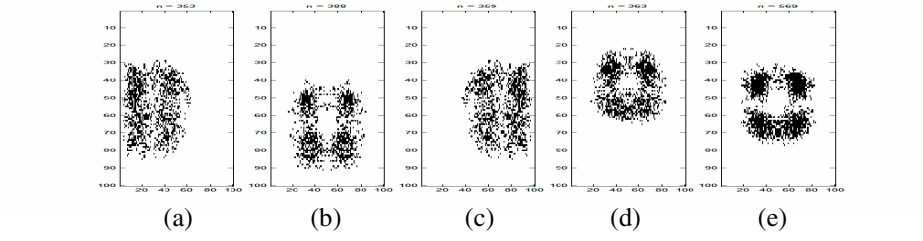


Fig. 2. Facial feature position for five clusters: left-sided (a), frontal down (b), right-sided (c), frontal up (d), frontal (e)

3 Hybrid Diabolos Networks

The Diabolos network is an auto-associative neural network. It is a completely connected two-layered perceptron. The input and output layers have the same side as the desired output is equal to the input. So, the network is trained to reconstruct an output identical to its input. It implements a specialized compression (quite similar to non-linear principal component analysis) as its hidden layer has much less units than input or output does. This network was successfully used for compression [5], handwritten character recognition [14], and face detection [1, 8]. In this latter application, the network is used to modelize the "face-class" and trained to reconstruct face images. So a non-face image should be badly compressed and the reconstruction error would be higher than for a face image. Here, we do not want to reconstruct a specific pattern class (the "face-class" for example) but to localize specific features within these patterns (eyes and/or mouth corners in the face case). In other words, we want to associate an image of face (input) with a facial feature map (output). So, we used as desired output, the normalized images containing the feature positions described in §2.

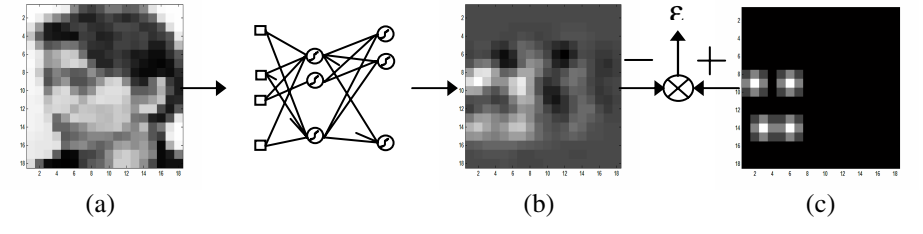


Fig. 3. Training process: input image (a) feeds the network. The mean squared error ε between network output(b) and feature map (c) is used as the cost function.

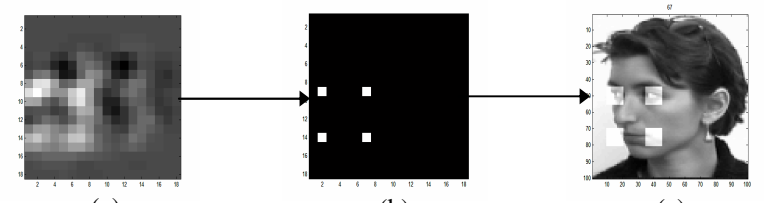


Fig. 4. Decision process: network produces the output image (a) where local maxima are detected (b) and back-projected onto the original image(c)

The network is trained using the back-propagation algorithm with adaptive momentum. The cost function is the mean squared error between network output and desired output (Fig. 3). Training parameters (number of epochs, hidden layer size) are tuned by exhaustive research. Once trained, the network is able to localize facial feature on unknown test images. The feature positions can directly be inferred by simply searching the maxima in the output image and back-projected onto the original image (Fig. 4). Let (x_{iD},y_{iD}) be the coordinates of these detected features.

4 Experimental Results

To evaluate the localization accuracy, we compute for each image the normalized error i.e the mean euclidean distance d between the detected feature position and the true feature position normalized with respect to the inter-ocular distance.

4.1 Orientation-Free Localizer

First, we trained a single neural network to localize facial feature on the whole database and perform orientation-free localization. We divided the whole dataset into two sets: training set (three fourth) and test (one fourth). Several experiments were made with different training and test sets.

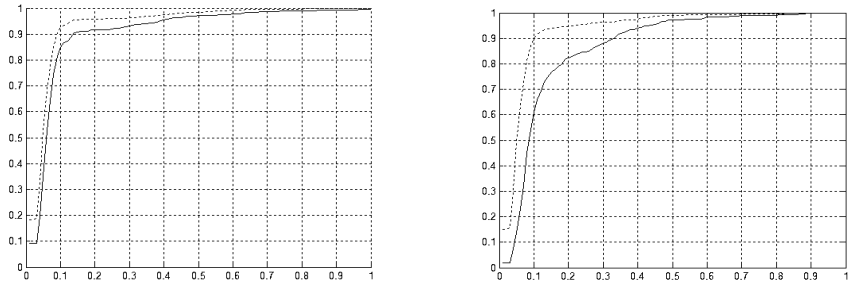
In the first experiment, we tested the localizer sensitivity to feature number and position. We dispatched the same people in both training and test sets with slightly different orientations. Then, we trained several localizers. The first one (SFL) consisted of four single feature localizers; each one specialized on one facial feature. The second (DFL) used two double feature localizers and each localizer dealed with a couple of features. Finally, (QFL) was a quadruple feature localizer (Fig. 3 & 4). Table 1 summarizes results in term of mean normalized error. These results are very interesting: the mean normalized error decreases as the number of feature to localize increases. This was quite predictable as the localizer associates a facial feature map with a face image. The more structured the feature map is, the more reliable the association will be. Note that when training an under-dimensioned QFL localizer (with a small number of hidden cells), this always outputs the same map that is the mean feature map whatever the input image is. Owing to these conclusions, we decided to make a thorough study on the QFL localizer. We can summarize its localization results on the test set (Fig 5.a) as follows: 35% of the images have a normalized error

lower than 0.05 (5%), 85% of the images have a normalized error lower than 0.1 (10%) and the mean normalized error is 0.096. We can validate localization hypothesis by computing the log-likelihood of the detected features coordinates (x_{iD},y_{iD}). We tuned a threshold on the training set to reject up to 10% of the poorest localization. This decreases the mean normalized error to 0.065.

Table 1. Mean normalized error of the single (SFL), double (DFL) and quadruple (QFL) feature localizers on the test set

Localizer	Mean normalized error
SFL	0.163
DFL	0.133
QFL	0.096

In the second experiment, we tested the QFL localizer sensitivity to identity. We dispatched different people in the training and test sets and trained a quadruple feature localizer. Compared to the first experiment, localization results (Fig. 5.b) are quite disappointing though predictable. Mean normalized errors on the training set are nearly the same for both experiments while they are very different on the test set showing that identity influences greatly the localizer accuracy. Only 15% of the images have a normalized error lower than 0.05 (5%), 60% of the images have a normalized error lower than 0.1 (10%) and the mean normalized error is 0.138.



(a) (b)

Fig. 5. Ratio of face images versus normalized error on training (dotted) and test (solid) databases. Sensitivity analysis: face orientation (a) and identity (b).

4.2 Multiple Localizer

Training. To improve the localizer accuracy, we decided to use several localizers; each one specialized on a given orientation. The clustering procedure described in §2 could separate the initial dataset into several subsets corresponding to a given face pose. Given N the number of considered orientations, the corresponding multiple localizer consists in N networks. So, for an input image, we have now N output images and N localization hypothesis corresponding to the four local maxima of each output image (Fig. 6). To compare the accuracy of the multiple localizers, we

compute the normalized error for each hypothesis and apply the WTA (Winner Takes All) criterion to select the best one. We have considered up to $N=6$ orientations. As can be seen (Fig. 7) the mean normalized error decreases continuously on both training and test sets when N increases. Such results are quite logical: as the number of specialized networks increases, the range of face orientations each network has to deal with decreases. The association process between face image and feature map becomes easier and the normalized error decreases.

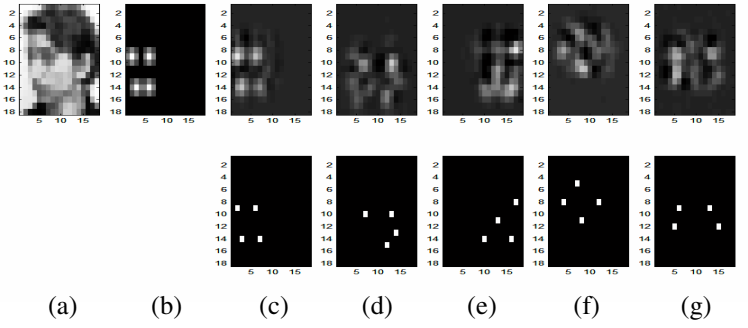


Fig. 6. Multiple localizers: Input image (a), target image (b), output image for the five networks and localization hypothesis (c to g)

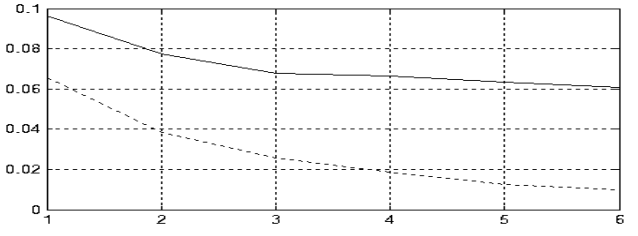


Fig. 7. Mean normalized error on the training (dotted) and test (solid) sets versus number of orientations considered

Decision. Latter result gave us a lower bound for localization error as it was produced by using the true feature position. As this information is not available, we have to find other criteria to select the best hypothesis.

Visual inspection of the networks output drives us to define a radiometric criterion (RC). As can be seen (Fig. 6), the best hypothesis corresponds to an output image O closed to the feature map, in terms of pixel intensities. This is an advantage of multiple auto-associative networks. As they are trained to localize features for a specific face pose, they perform well on this given orientation and poorly on the others leading to "noisy" outputs. So we define an "ideal" output image I as follows:

- At each maximum position: $I(x_{iD},y_{iD}) = O(x_{iD},y_{iD})$
- Anywhere else: $I(i,j) = median(O)$

Then, we compute the distance between the ideal output I and the real output O. We get a set of N distances D_j for the N output maps and use the WTA criterion to select the best one.

Another probabilistic criterion (PC) is obvious: the log-likelihood of each hypothesis. Given the coordinates (x_{iD},y_{iD}) of the detected features and the coordinate joint probability distribution for each face orientation, we can compute a set of N likelihoods L_j for the N hypothesis and use the WTA criterion to select the best one. As for the orientation-free localizer, we can reject a poor hypothesis while considering its likelihood (PCR criterion).

Finally, in order to get the best of these radiometric and probabilistic information, we can combine the two criteria. We normalized the distance vector $D = \{ D_1, \ldots, D_N\}$ and the likelihood vector $L = \{L_1, \ldots, L_N\}$ on [0;1] and use the sum rule [9] to combine them. Note that we experimented several normalization processes and combination operators (weighted sum, neural combination …) leading to quite similar results.

Table 2. Mean normalized error of the multiple localizer using radiometric (RC) criterion, probabilistic (PC) criterion and their combination on the test set

Criterion	Mean normalized error	
	N=3	N=5
RC	0.088	0.089
PC	0.121	0.146
PCR	0.055	0.058
Combination	0.091	**0.082**

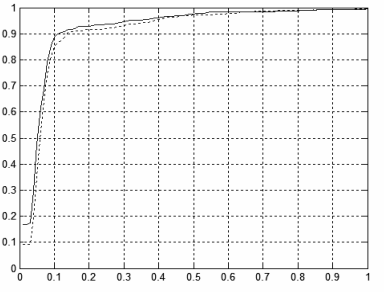


Fig. 8. Normalized localization error on the test set: orientation-free localizer (dotted) and multiple localizer combining five specialized networks (solid)

Table 2 summarizes the results for the two criteria and their combination, focusing on two multiple localizers respectively combining three and five specialized networks. The RC criterion outperforms slightly the orientation-free localizer (mean normalized error: 0.096). We can explain the poor results of the PC criterion by reminding the main drawback of auto-associative networks. As these latter are specialized on a specific face pose, they always produce an output that is close to the mean

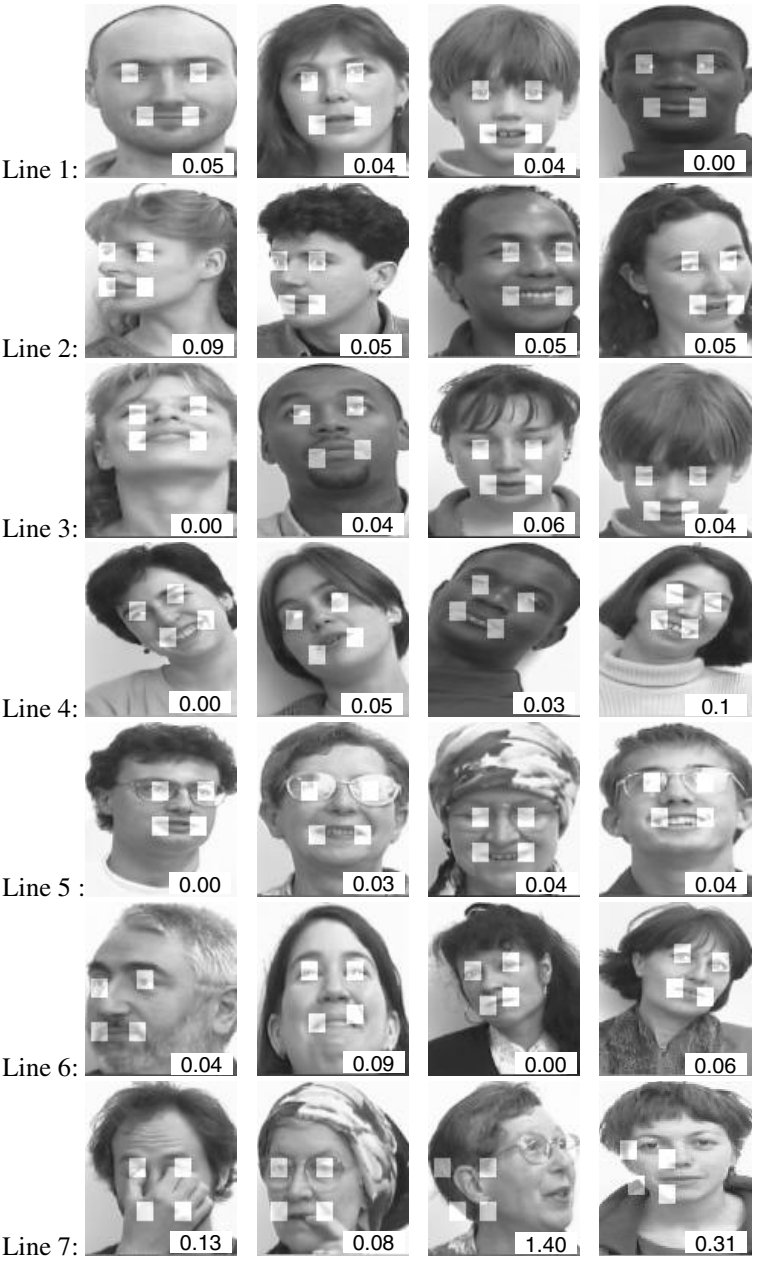


Fig. 9. Localization results on some test images. The normalized error is indicated bellow.

feature map of this face pose. This leads to high likelihood value whatever be the face image. Meanwhile, the PCR criterion is quite accurate. The information combination outperforms the orientation-free localizer. The higher accuracy happens when using five specialized networks. We can summarize the multiple localizer results on the test set (Fig. 8) as follows: 50% of the images have a normalized error lower than 0.05 (5%), 90% of the images have a normalized error lower than 0.1 (10%) and the mean normalized error is 0.082. Finally, we present some localization results on test images (Fig. 9): frontal faces (1st line), left-sided and right-sided faces (2nd line), up-sided and down-sided faces (3rd line) and tilted faces (4th line). Localizer sensitivities to glasses (5th line), scale (6th line) and partial occlusions (7th line) are shown. The association procedure makes the system less sensitive to partial occlusions and noise: e.g. if one feature is not visible, its position is inferred by the positions of other visible features. Two localization errors are presented (7th line). Note that, in both cases, an accurate localization hypothesis was found but the combination method failed to select it.

5 Conclusions and Future Works

We have presented a novel algorithm for the detection of facial features in a pre-focused face image. It is based on a particular neural network trained to associate a feature map with a face image. We studied thoroughly the single, orientation-free localizer and show that its accuracy increases with the number of features to detect. We proposed an alternate method where several specialized networks were trained to deal with specific face pose. The best localization hypothesis is then selected by combining radiometric and probabilistic information. This multiple localizer is more accurate than the orientation-free localizer: the mean normalized error decreases from 9.6% to 8.2%. Note that the whole system is quite fast: more than 14 images/second (on a Pentium IV 2.8 Ghz with MathLab).

Training the system on a larger dataset will validate all these results and should increase the localization accuracy in two ways. First, we hope it will reduce the sensitivity to identity. Secondly, it will increase the generalization ability in the multiple networks case and, by the way, the whole accuracy. To deal with such generalization problems, classical methods like bootstrapping and shared weight networks [6] are under study.

We have to evaluate the accuracy of the complete localizer, by cascading the face localizer [1], with the facial feature localizer. Finally, the cascade should be extended to perform coarse-to-fine localization and deal with finer facial feature (like eye corners or iris for example).

References

[1] Belaroussi, R., Prevost, L., Milgram, M., Classifier combination for face localization in color images. International Conference on Image Analysis and Processing, Lecture Notes in Computer Sciences, Vol. 3617, (2005), 1043-1050

[2] Bishop, C. M., Neural Networks for Pattern Recognition. Oxford University Press, (1995)

[3] Chellappa, R., Wilson, C.L., Sirohey, S.: Human and machine recognition of faces: a survey, Proceedings of IEEE, Vol. 83, no. 5, (1995), 705-740

[4] Cristinacce, D., Cootes., T.: A comparison of shape constrained facial feature detectors, International Conference on Automatic Face and Gesture Recognition, (2004), 375-380
[5] DeMers, D., Cottrell, G.: Non-linear dimensionality reduction. Neural Information Processing Systems 5, (1993) 580–587hould
[6] Duffner, S., Garcia, C., A Connexionist Approach for Robust and Precise Facial Feature Detection in Complex Scenes, IEEE International Symposium on Image and Signal Processing and Analysis, (2005), 316-321
[7] Feng, G.C., Yuen P.C.: Multi-cues eye detection on gray intensity image, Pattern Recognition Vol. 34, (2001), 1033-1046
[8] Féraud, R., Bernier, O., Viallet, J., Collobert, M.: A fast and accurate face detector based on neural networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, no. 1, (2002) 42-53
[9] Fumera, G.; Roli, F.;A theoretical and experimental analysis of linear combiners for multiple classifier systems, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, no. 6, (2005), 942-956
[10] Ioannou, S., Wallace, M., Karpouzis, K., Raouzaiou, A., Kollias S.: Combination of Multiple Extraction Algorithms in the Detection of Facial Features, International Conference on Image Processing, (2005),
[11] Milgram, M., Belaroussi, R., Prevost L.: Multi-stage combination of geometric and colorimetric detectors for eyes localization, International Conference on Image Analysis and Processing), Lecture Notes in Computer Sciences, Vol. 3617, (2005), 1010-1017
[12] Moghaddam, B. Pentland, A.: Probabilistic visual learning for object representation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, no. 7, (1997), 696-710
[13] Peng P., Chen, L., Ruan, S., Kukharev, G.: A Robust and Efficient Algorithm for Eye Detection on Gray Intensity Face. International Conference on Advances in Pattern Recognition, Lecture Notes in Computer Sciences, Vol. 3687, (2005), 302-308
[14] Schwenk, H., Milgram, M.: Transformation invariant auto-association with application to handwritten character recognition. Neural Information Processing Systems 7, (1995) 991-998
[15] Yuille, A., Hallinan, P., Cohen, D.: Feature extraction from faces using deformable templates. International Journal of Computer Vision, Vol. 8, no. 2, (1992), 99-111