A Neural Architecture for Fast and Robust Face Detection

Christophe Garcia and Manolis Delakis
Department of Computer Science, University of Crete, P.O. Box 2208, 71409 Heraklion, Greece

Abstract

In this paper, we present a connectionist approach for detecting and precisely localizing semi-frontal human faces in complex images, making no assumption about the content or the lighting conditions of the scene, or about the size or the appearance of the faces. We propose a convolutional neural network architecture designed to recognize strongly variable face patterns directly from pixel images with no preprocessing, by automatically synthesizing its own set of feature extractors from a large training set of faces. We present in details the optimized design of our architecture, our learning strategy and the resulting process of face detection. We also provide experimental results to demonstrate the robustness of our approach and its capability to precisely detect extremely variable faces in uncontrolled environments.

1. Introduction

Human face detection is becoming a very important research topic, due to its wide range of applications, like security access control, model-based video coding or content-based video indexing, advanced human and computer interaction. It is also a required preliminary step to face recognition and expression analysis. Many different approaches for face detection have been proposed in the last years. Most methods are based on local facial features detection by low-level computer vision algorithms and classification using statistical models of human face [2,3,10]. Other approaches are based on template matching where several correlation templates are used to detect local sub-features, considered as rigid in appearance (eigenfaces [5]) or deformable [2,9]. The main drawback of these approaches is that either little global constraints are applied on the face template or extracted features are strongly influenced by noise or change in facial expression or viewpoint. Generally, the use of skin color information is an important cue for constraining the search space. In [1], we proposed a fast method using skin color filtering and probabilistic classification of facial textures based on statistical measures extracted from a wavelet packet decomposition.

In the general case of grey level images, unlike other systems depending on a hand crafted feature detection stage, followed by a feature classification stage, some techniques based on neural networks have been proposed. These techniques have the clear advantage of learning underlying rules contained in the highly variable face patterns from large training sets of images. They proved to be very tolerant to noise and distorsions. The first advanced neural approach that reported results on a large and difficult dataset was by Rowley *et al.* [7]. Their system incorporates face knowledge in a retinally connected neural network, looking at windows of 20x20 pixels. In their single neural network implementation (referred as system 5), there are two copies of a hidden layer with 26 units, where 4 units look at 10x10 pixel subregions, 16 look at 5x5 subregions, and 6 look at 20x5 pixels overlapping horizontal stripes. A large number of adjustable weights (2,905) are learnt through standard backpropagation. The input window is pre-processed through lighting correction (a best fit linear function is subtracted) and histogram equalization, like in the Sung and Poggio's system [8]. The image is scanned with a moving 20x20 window at every possible position and scale (with a subsampling factor of 1.2). To reduce the number of false alarms, they combine multiple neural networks with an arbitration strategy. Osuna *et al*. [6] developed a support vector machine (SVM) approach to face detection. The proposed system uses the same pre-processing stage for lighting correction and scan input images over scales with a 19 x 19 window. A SVM with a 2nd-degree polynomial as a kernel function is trained with a decomposition algorithm that guarantees global optimality. Approximately 2,500 support vectors are obtained and use for face detection.

In this article, we propose a novel scheme based on convolutional neural networks that have been introduced by Le Cun *et al*. and successfully applied to handwritten character recognition [4]. In comparison to the two methods mentioned above, our system automatically derives optimal convolution filters that act as feature extractors. Therefore, the use of receptive fields, shared weights and spatial subsampling in such a neural model provides much higher degrees of invariance to translation, rotation, scale, and deformation of the face patterns, while strongly reducing the number of adjustable weights to learn, aiding generalization. Moreover, no preprocessing on the input image is required and fast processing is automatically provided by successive simple convolutional and subsampling operations.

We first present in details the design of our architecture, our learning strategy. Then, we present the process of face detection using this architecture. Finally, we provide experimental results and a comparison to the technique proposed in [7] to demonstrate the robustness of our approach and its capability to precisely detect extremely variable faces in uncontrolled environment.

1051-4651/02 $17.00 (c) 2002 IEEE

2. The Proposed Approach

2.1. Neural network architecture

The convolutional neural network, shown in Fig.1, consists of a set of three different kinds of layers. Layers C_i are called convolutional layers, which contain a certain number of planes. Layer C_1 is connected to the retina, receiving the image area to classify as face or non face. Each unit in a plane receives input from a small neighborhood (biological local receptive field) in the planes of the previous layer. The trainable weights (convolutional mask) forming the receptive field for a plane are forced to be equal at all points in the plane (weight sharing). Each plane can be considered as a feature map that has a fixed feature detector that corresponds to a pure convolution with a trainable mask, applied over the planes in the previous layer. A trainable bias is added to the results of each convolutional mask. Multiple planes are used in each layer so that multiple features can be detected.

Once a feature has been detected, its exact location is less important. Hence, each convolutional layer C_i is typically followed by another layer S_i that performs a local averaging and subsampling operation. More precisely, a local averaging over a neighborhood of four inputs is performed followed by a multiplication by a trainable coefficient and the addition of a trainable bias. This subsampling operation reduces by 2 the dimensionality of the input and increases the degrees of invariance to translation, rotation, scale, and deformation of the face patterns.

In our implementation, layers C_1 and C_2 perform convolutions with trainable masks of dimension 5x5 and 3x3 respectively. Layer C_1 contains 4 feature maps and therefore performs 4 convolutions on the input image. Layers S_1 and C_2 are partially connected. Mixing the outputs of feature maps helps in combining different features, thus in extracting more complex information. In our system, layer C_2 has 14 feature maps. Each of the 4 subsampled feature maps of S_1 is convolved by 2 different trainable masks 3x3, providing 8 feature maps in C_2. The other 6 feature maps of C_2 are obtained by fusing the results of 2 convolutions on each possible pair of feature maps of S_1.

Layers N_1 and N_2 contain simple sigmoid neurons. The role of these layers is to perform classification, after feature extraction and input dimensionality reduction are performed. In layer N_1, each neuron is fully connected to every points of one feature map only of layer S_2. The unique neuron of layer N_2 is fully connected to all the neurons of the layer N_1. The output of this neuron is used to classify the input image as face or non face. For training the network, we used the classical backpropagation algorithm with momentum modified for being used in convolutional networks as described in [4]. Desired responses are set to –1 for non-faces and to +1 for faces.

In our system, the dimension of the retina is 32x36. Because of weight sharing, the network has only 897 trainable parameters, despite the 127,093 connections it uses. Local receptive fields, weight sharing and subsampling provide many advantages to solve two important problems at the same time: the problem of robustness and the problem of good generalization, which is critical given the impossibility of gathering in one finite-sized training set all the possible variations of the face pattern. This topology has another decisive advantage. In order to search for faces, the network must be replicated (or scanned) at all locations in the input image, as done in the above mentioned approaches [6,7]. In our approach, since each layer essentially performs a convolution (with a small-size kernel), a very large part of the computation is in common between two neighboring locations in the input images. This redundancy is naturally eliminated by performing the convolutions corresponding to each layer *on the entire input image at once*. The overall computation amounts to a succession of convolutions and non-linear transformations over the entire images.

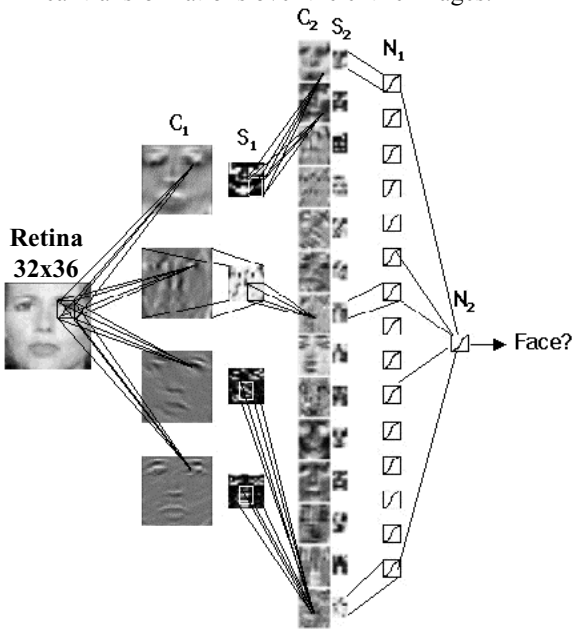

Fig. 1: Convolutional neural network architecture

2.2. Training Methodology

We built our training set by manually cropping 2,146 highly variable face areas in a large collection of images obtained from various sources over the Internet. Most of the neural network-based approaches in the literature [6,7] use an input window of dimension around 20x20, reported as being the smallest window one can use without loosing critical information. Usually, this window is the very central part of the face, excluding the border of the face and any background. We have chosen approximately the same window for the central part of the face but we have added in the input the border of the face and in some cases some portions of background. By doing so, we give the network some additional information, which can help in

characterizing the face pattern and canceling some border effects that may arise in the convolutions. Finally, the cropped faces have a size of 32x36 in order to account for the face aspect ratio. No intensity normalization is applied on the cropped faces, such as histogram equalization and overall brightness correction that are performed in [6,7]. In addition, we have no need to perform the tedious task of spatial normalization so that the eyes, mouth and other parts of the faces remain exactly at the same position [6,7]. Moreover, systems in [6,7] are only tolerant to small rotations of ±5 degrees. As mentioned earlier, our network topology is quite robust in scale and position, and we aim at enforcing this robustness by providing examples that are not normalized. In order to create more examples and to enhance the capabilities of invariance to rotation and variation of intensity, some transformations such as rotation of ±30 degrees and contrast reduction are applied to all the examples, leading to a final training set of 12,976 faces. Some samples are shown in Fig. 2.

Fig. 2: Some samples of the training set.

We collect non-face examples via an iterative bootstrapping procedure. We first build an initial training set of non face examples by producing random images. The network is then trained with face and non face examples. The iterative bootstrapping procedure acts as follows. For the first iteration, the trained network is used for scanning a set of 120 various highly textured images containing no face. Areas where the response of the network is greater than a threshold *thr=0.8* are added to the set of non face examples. Then, the same network is retrained with the set of face examples and the updated set of non face examples. The procedure of scanning for false alarms and training the network is repeated for 4 more iterations reducing the threshold *thr* by 0.2 at each iteration until it reaches 0.0, which is the separating value between face and non faces. By doing so, we gather iteratively false examples which are close to the boundaries of the cluster of "faces" in network space, without gathering to many false alarms in the early stages of training. We finally obtain about 15,000 false examples.

2.3. Face Localization

In order to detect faces of different sizes, the input image is repeatedly subsampled via a factor of 1.2, resulting in a pyramid of images. Each image of the pyramid is filtered by our network. In [6,7], the neural filter is applied at every pixel of each image of the pyramid, after some operations of lighting corrections, given that it has very small invariance in intensity, position and scale. In our approach, as mentioned earlier, each image of the pyramid is entirely convolved at once by the network. For each image of the pyramid, an image containing the network results is obtained. Because of the successive convolutions and subsampling operations, this image has a size approximately four times smaller than the original one. This fast procedure corresponds to the application of the network retina at every location of the input image with a step 4 in both dimensions, without computational redundancy. This search may be seen as a very fast rough localization, where the positive answers of the network correspond to candidate faces.

Then, candidate faces in each scale are mapped back to the input image scale. They are iteratively grouped according to their proximity in image and scale spaces. Each group of candidate faces is fused in a representative face whose center and size are computed as the average of the centers and sizes of the grouped faces weighted by their network responses. After applying this grouping algorithm, the representative face candidates serve as a basis for the next stage of the algorithm in charge of fine face localization and false alarm dismissal.

A fine search is performed in an area around each rough face candidate center in image-scale space. A search space centered at the face candidate position is defined in image-scale space for precise localization of the candidate face. It corresponds to a small pyramid centered at the face candidate position covering 5 scales varying from 0.8 to 1.4 of the scale of the face candidate. For every scale, the presence of a face is evaluated on a grid of 6x6 pixels around the corresponding face candidate center position. Usually true faces give positive responses in 2 or 3 consecutive scales, but non-faces not so often. We therefore count the number *nok* of positive responses in the fine search space. Face candidates are accepted if *nok>6*. Fig. 3 shows different steps of the detection process for an image containing 3 faces at different scales. The first line presents the feature maps computed by layer C_1, at the scale corresponding to the central face. The second line presents the final responses of the network at all scales. The black points correspond to positive responses. The third line shows the positions and sizes of the faces detected during fine search, and the final results. One can notice that one false alarm has been detected, with only 2 votes in fine search and removed according to the criterion *nok>6*.

3. Experimental Results

The proposed method has been evaluated using the test data set used in [1], which contains images kindly provided by the Institut National Audiovisuel (INA), France and by ERT Television, Greece. This test data of 100 images contains 124 faces (of minimal size 19x22 pixels) that present large variability in size, illumination, facial expression, orientation, and partial occlusions. In Fig. 4., we present some results of the proposed face detection scheme on this test data set. These examples include images with multiple faces of different sizes and different poses.

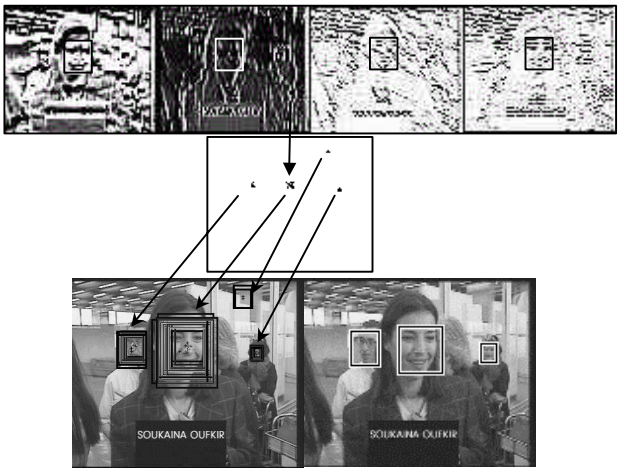

Fig. 3: The process of detection

False alarms and false dismissals examples are presented as well. On this test set we obtained a good detection rate of 97.5% with 3 false alarms for *nok>6*. It should be noted that the number of false alarms is very small. This may illustrate the capability of the convolutional network architecture to highly separate face from non-face examples. As a comparison, with our previous approach [1] we obtained 94.23% of good detection rate with 20 false alarms when 104 faces (of size greater than 48x80 pixels which was the minimal size for this approach) are considered. Considering this subset of 104 faces, the CMU's system [7] resulted in 85.57% of good detection and 15 false alarms and the approach proposed in this paper in 98% of good detection and 1 false alarm. An interactive demonstration of our system is available on the Web at www.csd.uoc.gr/~cgarcia/FaceDetectDemo.html, allowing anyone to submit images for processing and to see the detection results for pictures submitted by other people.

4. Conclusion

Our experiments have shown that using convolutional neural networks for face detection is a very promising approach. The robustness of the system to varying poses, lighting conditions, and facial expressions was evaluated using a set of difficult images. In addition, the stability of responses in consecutive scales and a precise localization of faces were noticed. Because of its convolutional nature, our system is faster than the other approaches [6,7] which require a dense scanning of the input image at all scales and positions. It processes a 352x288 image in less than 4 sec. on a PC (PIII 933Mhz with 256M memory). Moreover, the proposed approach is not restricted to vertical semi-frontal faces. It is able to detect faces tilted up to ±30 degrees.

As an extension of this work, we plan to use the information contained in the convolution layers of the network at the end of the face detection step for other purposes related to face analysis.

Fig. 4: Some results of the proposed method

References

[1] C. Garcia and G. Tziritas, Face Detection Using Quantized Skin Color Region Merging and Wavelet Packet Analysis. *IEEE Trans. Multimedia*, 1(3):264-277, 1999.

[2] C. Garcia, G. Simandiris and G. Tziritas. A Feature-based Face Detector using Wavelet Frames. In: *Proc. of Intern. Workshop on Very Low Bit Coding*, pp. 71-76, Athens, Ooctober 2001.

[3] S.-H. Jeng, H. Y. M. Yao, C. C. Han, M. Y. Chern and Y. T. Liu. Facial Feature Detection Using Geometrical Face Model: An Efficient Approach. *Pattern Recognition,* 31(3):273-282, 1998.

[4] Y. Le Cun, B. Boser, J.S. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel.. Handwritten digit recognition with a backpropagation neural network. In D. Touretzky editor, *Advances in Neural Information Processing Systems* 2, pp.396–404. 1990.

[5] B. Moghaddam, A. Pentland. Probabilistic Visual Learning for Object Recognition, *IEEE Trans. PAMI*, 19(7):696-710, 1997.

[6] E. Osuna, R. Freund, F. Girosi. Training Support Vector Machines: an application to face detection, In: *Proc. of CVPR*, Puerto Rico, pp.130-136, 1997.

[7] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. PAMI*, 20(1):23-28, 1998.

[8] K. K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. PAMI*., 20(1):39–51, 1998.

[9] L. Wiskott, JM. Fellous, N. Kruger, C. Von der Malsburg. Face Recognition by Elastic Bunch Graph Matching. *IEEE Trans. PAMI,* 19(7):775-779, 1997.

[10] K. C. Yow, C. Cipolla. Feature-based human face detection. *Image and Vision Computing*, 15, pp. 713-735, 1997.