Detection and localization of faces on digital images

Gilles Burel, Dominique Carel

▶ To cite this version:

Gilles Burel, Dominique Carel. Detection and localization of faces on digital images. Pattern Recognition Letters, 1994, 15 (10), pp.963-967. 10.1016/0167-8655(94)90027-2 . hal-03222631

HAL Id: hal-03222631
https://hal.univ-brest.fr/hal-03222631

Submitted on 18 Mar 2023

Pattern Recognition Letters, Vol. 15, n^o 10, October 1994

Detection and localization of faces on digital images

Gilles BUREL & Dominique CAREL

Thomson CSF, Laboratoires Electroniques de Rennes
Avenue de Belle Fontaine, 35510 Cesson-Sévigné, France

Abstract

A method for automatic detection and localization of faces on digital images is proposed. The method is based on learning by example and multi-resolution analysis of digital images. Special emphasis is put on the management of the learning data, in order to improve the performances. Various experimental results, obtained by using a Multi-Layer Perceptron (MLP) as a classifier, are provided.

KEYWORDS : Detection and Localization of Faces, Multi-Resolution, Classification, Multi-Layers Perceptron, Back-Propagation.

1 Introduction

Many authors have studied the problem of face characterization. The goal is generally identification of faces, that means providing the name of the person when the face is already localized (identity photography, hand made location). The techniques used are for instance analysis of isodensity maps (Nakamura et al. (1991)), measurement of anthropomorphic characteristics (Craw et al. (1992)), analysis of symmetries (Edelman et al. (1992)), or Principal Component Analysis (Turk and Pentland (1991)). A review is provided in Samal and Iyengar (1992).

The objective of detection and localization of faces is to provide the coordinates of the frames including the faces in an image. As mentioned in Chetverikov and Lerch (1993), few authors have addressed this problem : usually, it is assumed that the location of a face is known a priori. However, this is not often the case, and face detection and localization is needed prior to identification. The recent work reported in Chetverikov and Lerch (1993) is an attempt to address this problem by means of flexible matching of specific configurations of blobs and streaks. Here, we propose an alternative approach based on machine learning. The advantage of such an approach is the possibility to adapt to any special condition (people, environment), and to improve the performances by increasing the size of the data-base.

The current lack of papers as well as the lack of efficient industrial devices related to detection and localization of faces is mainly due to some difficulties issued from the problem itself :

- It is difficult to modelize “what a face is”, because of the large diversity among people, and the non-rigidity of a face.
- In the context of realistic industrial applications, the distance between the face and the camera is unknown.
- In the context of realistic industrial applications, the lighting conditions are only partially controlled.
- A face may be partially rotated. The system must be able to detect it, even if the face is slightly rotated.
- The system must be able to deal with uncontrolled backgrounds.

To deal with these problems, we have developped a method which is based on the following ideas :

- Learning by example, in order to avoid the problem of lack of a priori model.
- Multi-resolution analysis of digital images, in order to deal with the fact that the distance between the face and the camera is undetermined.
- Local normalization in order to reduce the effect of lighting conditions.
- Creation of a big training base, containing various faces under diverse orientations.

The paper is composed of two parts : in the first part, we explain the proposed method, and in the second part, we comment on experimental results.

2 The proposed method

The method can be decomposed in three main parts :

- A general training phase, during which the system tunes its internal parameters.
- A local training phase, during which the system adapts its internal parameters to the particular environment of a local site.
- A detection-localization phase during which the internal parameters are frozen while the system receives on input a sequence of images and provides on output the coordinates of the including frames of the faces detected.

The first two phases can be considered as preliminary, since their objective is to tune the system. The last phase is the normal operating mode of the system.

Figure 1 shows an overview of the method. The general training phase is performed on the basis of a huge set of faces and backgrounds (counter-examples). The classifier is an MLP (Rumelhart et al. (1986)).

For many realistic applications, the system is placed at a local site. Furthermore, the people who may appear on the image sometimes belong to a group known a priori. In that case, it is interesting to perform a further training on a learning base that takes this knowledge into account. To achieve fast training we proceed as follows :

1. The general training base is compressed using a Kohonen network (Kohonen (1984)), taking profit of the Vector Quantization properties of this algorithm.
2. The compressed base is completed by examples of faces and backgrounds related to a specific environment.
3. The classifier (MLP) is shortly trained on this base, starting from its state at the end of the general learning phase.

The backgrounds related to a specific environment can be chosen at random, but better performances are achieved if we choose with a higher priority backgrounds which cause false detections (using the network trained on the general learning base to process local sequences).

The detection-localisation phase consists in scanning each image of the sequence at various resolutions. For each location and size of the scanning window, the window content is normalized to a standard size (15x20 pixels for our experiments), then normalized in mean and variance (to reduce the sensibility to lighting conditions), and finally propagated through a MLP. The MLP provides a class (face/background) and a confidence measure.

All the local decisions (class+confidence) are then fused spatially. The aim of the spatial fusion is mainly to suppress multi-detections at close resolutions or close spatial locations. To achieve this goal, we search for the strongly overlapped detections. Two detections are considered as strongly overlapped if the center of a window is included in the other window. For such configurations, we suppress the detection which has the lower confidence.

Another goal of spatial fusion is to suppress incoherent detections, according to elementary geometrical reasoning. Anyone can be convinced that a situation where a small face is under a very bigger one is impossible, because in such a case, the smaller face is the farthest, and so, it should have been occluded by the body of the other person. Hence, we suppress all detections which are under a very larger one with higher confidence.

3 Experimental Results

3.1 Counting attendance to a conference, and creation of the general training base

For this experimentation, we use a CCD camera, which provides a black and white image of 720x576 pixels. We want to detect faces between 1.5m and 5m. Taken into account the characteristics of our camera, the size of a face corresponds approximately to 18x25 pixels (for a distance of 5m) and 60x80 pixels (for a distance of 1.5m).

We have disposed the camera in a conference room. Many people (more than forty) have been invited to attend the conference without any behaviour constraint. The average attendance to the conference was around 8 people. People were free to go out or come in the room according to their topics of interest. Various sequences have been registered, with changing lighting conditions and backgrounds. The sequences have then been divided in two sets : a set for training, and a set for testing, in such a way that a person or a background can be in only one set.

Each test sequence is scanned at 7 resolutions. The sizes of the scanning windows are 18x25, 23x31, 30x40, 37x50, 47x63, 60x80 (the change of scale between two successive resolutions is $2^{\frac{1}{3}}$). This allows the detection of faces at a distance between 1.5m and 5m from the camera. For each window location and size, the content of the window is normalized to 15x20 pixels (size of the MLP input layer). The image is scanned by steps of two pixels.

The general training base has been created by interactively defining the including frame of the faces. A total of 121 windows surrounding faces corresponding to 20 different persons, has been extracted. For each window, 27 sub-images have been extracted by translating (ΔX, ΔY) and scaling (ρ) the window :

$$\Delta X = -\frac{T_X}{10}, 0, \frac{T_X}{10}$$

$$\Delta Y = -\frac{T_Y}{10}, 0, \frac{T_Y}{10}$$

$$\rho = 2^{\frac{-1}{6}}, 1, 2^{\frac{1}{6}}$$

where (T_X, T_Y) stands for the size of the window, (ΔX, ΔY) for the translation, and ρ for the scaling factor. The values of the scaling factor have been chosen in order to fill the space between two successive resolutions.

We obtain finally a total of 121x27=3267 examples of faces. The same number of background examples is automatically extracted on sequences without people on them. Thus, the total number of examples is around 6500. A first network has been trained on this base. Then, we have completed the base with 3000 other backgrounds mistakenly detected as faces on various images, and performed some further learning iterations, to finally obtain a general training base of 9500 examples.

We have realized the learning experiment described above with 2- and 3-layer MLPs (resp. 300+2 neurons, and 300+10+2 neurons). The performances of these nets when used to examine new sequences are very close. Hence, since the 2-layer network is the fastest, it is better to use it.

Figure 2 shows a typical example of result on an image of the test base. One face is not detected, but this is probably due to the orientation of the face, and the presence of the arm partially overlapping the face.

This kind of result is confirmed when we process whole sequences. One or two false detections can appear stealthily, but may be removed by temporal filtering. The non-detection of faces is generally due to faces being strongly rotated, or to partial occlusion.

3.2 Surveillance of a room and creation of a local training base

Figure 3 shows a result obtained in another room with another CCD camera (providing images of 512x356 pixels). The camera is carried by a 6-axis industrial robot, which must watch over the room and detect faces.

The general data-base of 9500 examples previously mentionned has been compressed with a Kohonen algorithm, in order to obtain 256 faces and 256 backgrounds. Then, the robot has recorded images of the room seen from random locations. The false detections obtained on these images have been included in the compressed data-base, and some learning iterations have been performed.

Here, we had no a priori knowledge on the people coming in the room, which explains why no face has been added to the compressed base.

It is amazing to note that the MLP seems to have reached some level of abstraction, as proved by figure 4.

4 Conclusion

We have proposed a method for automatic detection and localization of faces on digital images. The method is based on learning by example on neural networks (an MLP is used for classification, and a Kohonen network for data-base compression). It uses a special management of the training base : there is a general training base as well as a base designed for a specific environment. Multi-resolution analysis is used to allow detection at various distances, and spatial fusion is performed to filter the results. Experimental results have been discussed to illustrate the approach. Further work may consist in increasing the size of the general training base (our training base currently includes the faces of 20 people).

REFERENCES

Chetverikov, D., & Lerch, A. (1993). Multiresolution Face Detection. In R. Klette and W.G. Kropatsch “Theoretical Foundations of Computer Vision”, Series : Mathematical Research, vol 69, Akademie Verlag.

Craw, I., Tock, D., & Bennet, A. (1992). Finding face features. *ECCV92 (European Conference on Computer Vision)*, Genova, Italy, 1992.

Edelman, S., Reisfeld, D., & Yeshurun, Y. (1992). Learning to recognize faces from examples. *ECCV92 (European Conference on Computer Vision)*, Genova, Italy, 1992.

Kohonen, T., (1984). Self-Organization and Associative Memory. Springer-Verlag, 1984.

Nakamura, O., Mathur, S., & Minami, T., (1991). Identification of human face based on isodensity maps. *Pattern Recognition*, vol. 24, n^o 3, 1991.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J., (1986). Learning internal representations by error backpropagation. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel Distributed Processing* (Vol. I, pp. 318-362). Cambridge, MA : The MIT Press.

Samal, A., & Iyengar, P.A., (1992). Automatic recognition and analysis of human face and facial expression : a survey. *Pattern Recognition*, vol 25, n^o1, 1992.

Turk, M., & Pentland, A., (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, vol 3, n^o1, 1991.

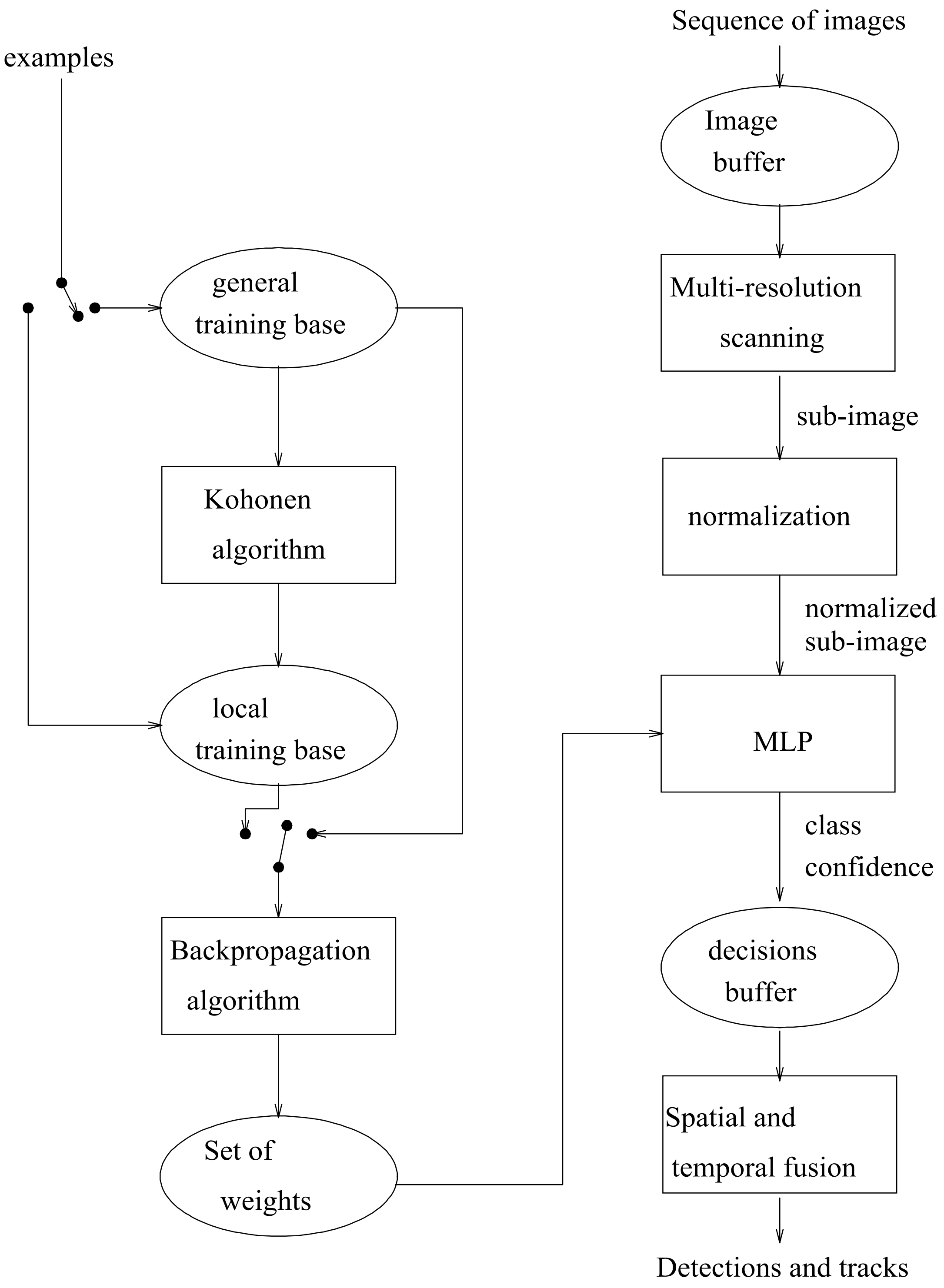

FIG. 1: Overview of the approach

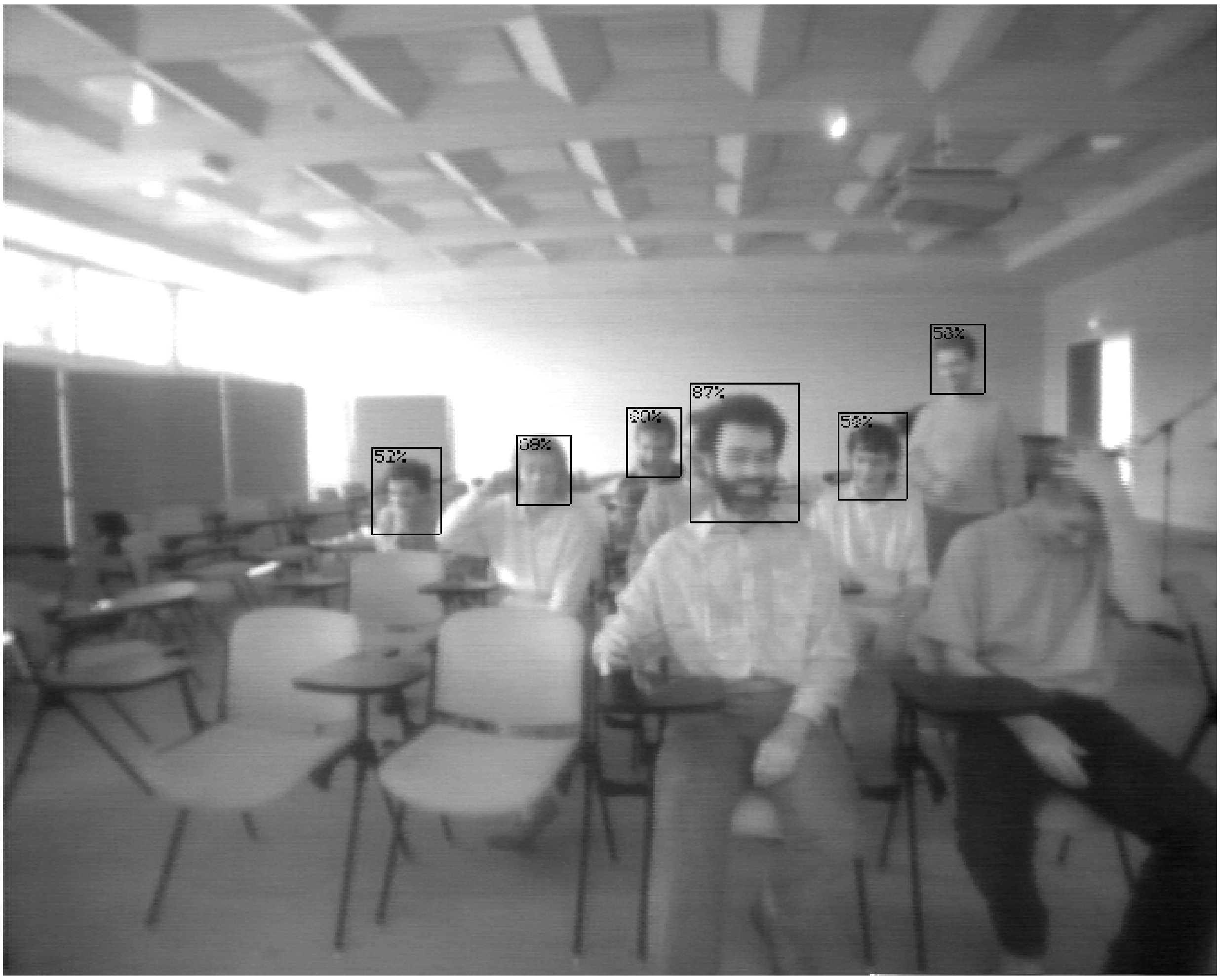

FIG. 2: Example of result (conference room)

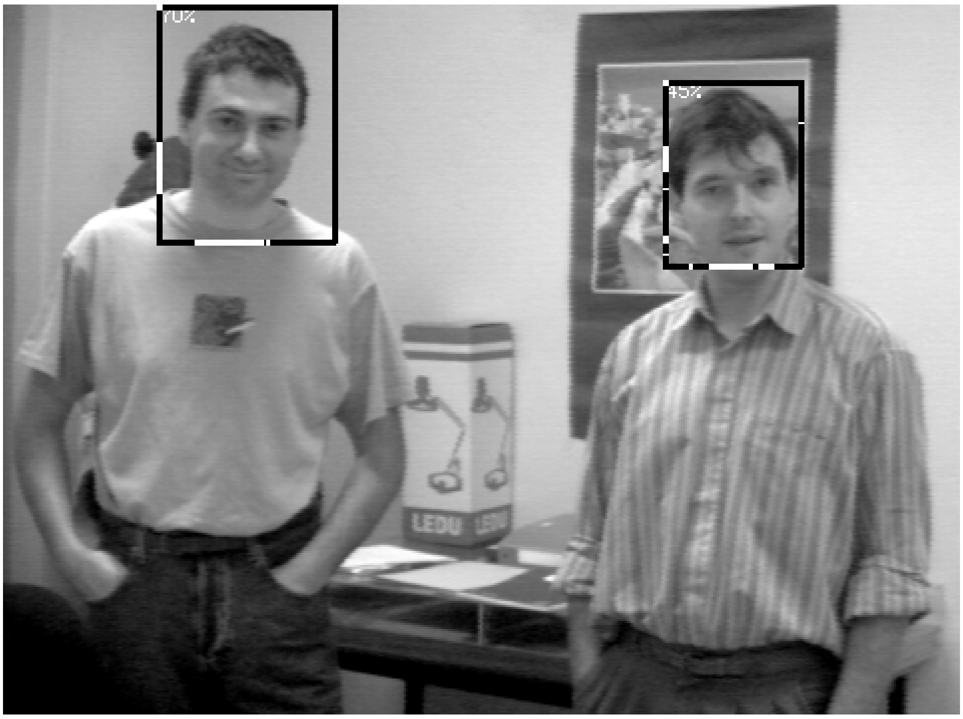

FIG. 3: Example of result (robot room)

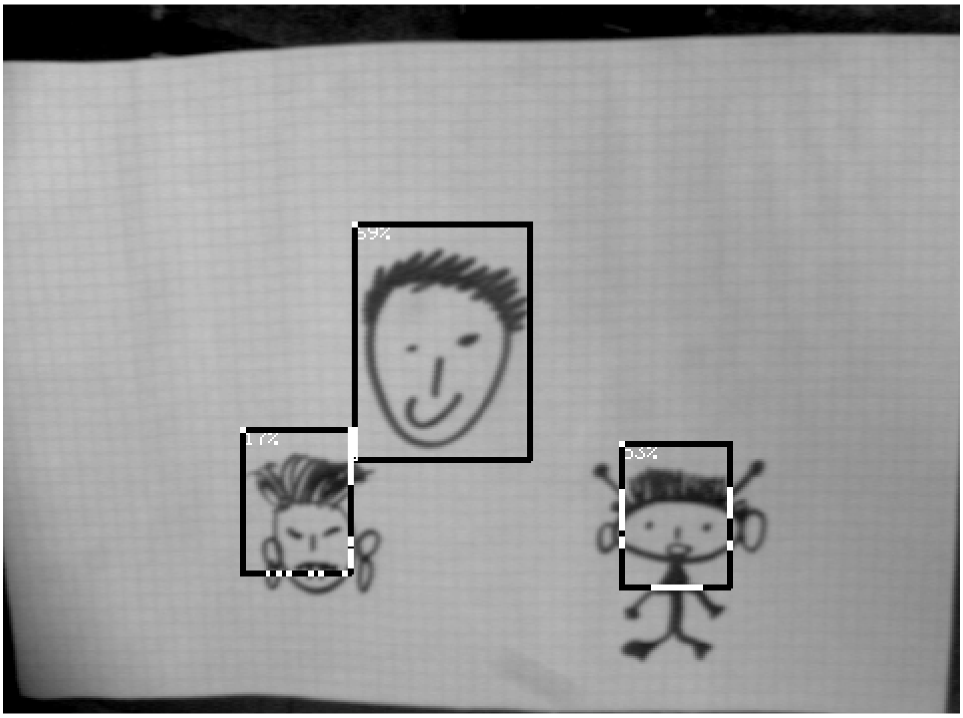

FIG. 4: Example of result