MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan

11-1

Face Recognition Using SVM Fed with Intermediate Output of CNN for Face Detection

Katsuhiko Mori, Masakazu Matsugu, Takashi Suzuki

Canon Inc., Intelligent I/F Project., 5-1, Morinosato-Wakamiya, Atsugi, 243-0193 Japan

Abstract

We present a face recognition method using support vector machines which utilize intermediate output of convolutional neural networks for face detection. Face detection process carry out the functions both of detecting faces in input image with complex background and of generating the feature vector implicitly representing shape information and spatial arrangement of facial components to supply to support vector machines. Our experiments show that robustness to size variability from 0.8 to 1.2 (relative size in units of area for reference face), demonstrating 100% recognition with 0% false acceptance rate for 600 images of 20 people.

1 Introduction

Face recognition algorithms have been extensively explored [1]-[3], [5]-[7], [9], [12]-[14] and most of which address the problem separately from object detection, which is associated with image segmentation, and many assume the existence of objects to be recognized without background. Some approaches, in the domain of high-level object recognition, address economical use of visual features extracted in the early stage for object detection. However, only a few object recognition algorithms proposed so far explored efficiency in the combined use of object detection and recognition [9].

For example, in the dynamic link matching (DLM) [14], Gabor wavelet coefficient features are used in face recognition and detection as well. However, we cannot extract shape as well as spatial arrangement information on facial components directly from those features since, for a set of nodes of the elastic graph, they do not contain such information. This necessitated to device the graph matching technique, a computationally expensive procedure, which requires quite different processing from feature detection stage. Convolutional neural networks (CNN) [8] have been exploited in face recognition and hand-written character recognition. In [10], we proposed a CNN model for robust face detection. SVM has also been used for face recognition [5]-[7], [9], [13]. In particular, in [6], [7], SVM classification was used for face recognition in the component-based approach.

This study, in the domain of face recognition as a case study for general object recognition with object detection, explores the direct use of intermediate as well as low level features obtained in the process of face detection. Specifically, we explore the combined use of convolutional neural networks (CNN) and support vector machines (SVM), the former used for feature vector generation, the latter for classification. Proposed algorithm is one of component-based approaches [6], [7] with appearance models represented by a set of local, area-based features. The direct use of intermediate feature distributions obtained in face detection, for face recognition, brings unified and economical process that involves simple weighted summation of signals, implemented both in face detection and recognition.

2 Feature Vectors Extracted from Low and Intermediate CNN Outputs

Convolutional neural networks, with hierarchical feed-forward structure, consist of feature detecting (FD) layers, each of which followed with a feature pooling (FP) layer or sub-sampling layer [8], [10], [11].

This architecture comes with the property of robustness in object recognition such as translation and deformation invariance as in well-known *neocognitrons* [4], which also have similar architecture. Figure 1 shows our CNN model [10], [11] for face detection, which is slightly different from traditional ones in that it has only FD modules in the bottom and top layers.

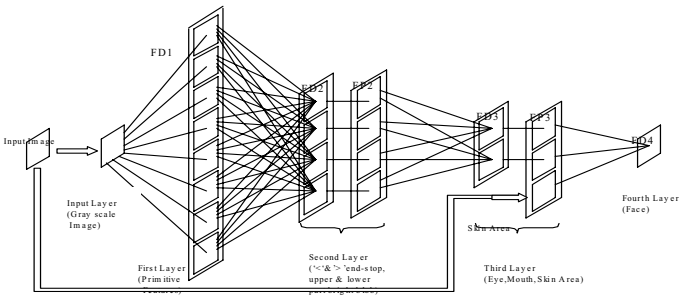

Fig.1 Convolutional neural networks architecture for face detection

In our CNN model, local features to be detected in each layer are edge-like features in the first layer, '<' and '>' end-stop, upper part bright blob, and lower part bright blob in the second layer, which we call alphabetical local features, eye and mouth in the third layer. Finally, using

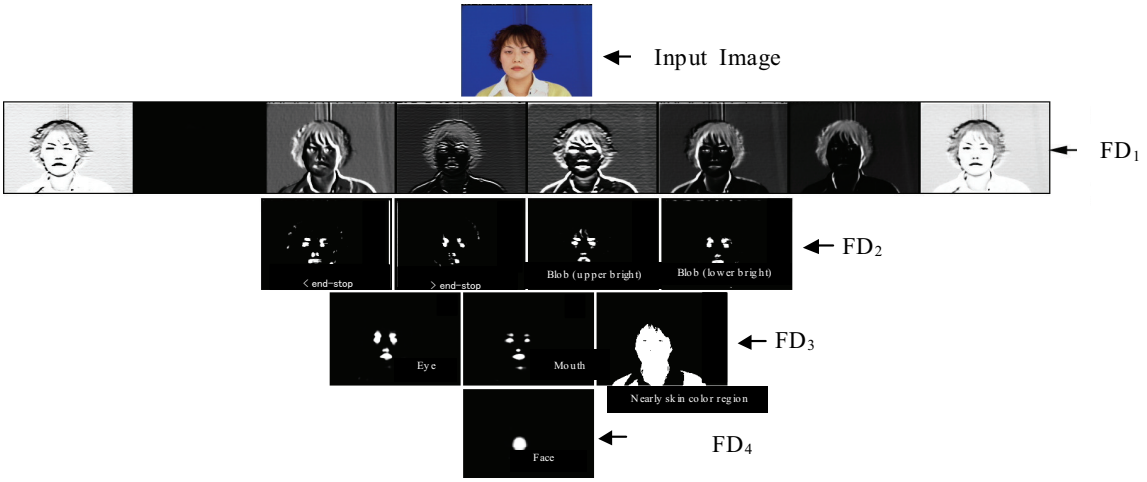

Figure 2: Output of each module in each layer

eye and mouth feature detected in the third layer and the skin area data defined by some restricted range of hue and saturation values, some face is detected in the forth layer.

Figure 2 shows the detected features in our CNN. In this figure, the output value of neurons in the range of (-1.0 – 1.0) is converted to (0(black) – 255(white)).

When some faces are detected in the CNN, we also have some intermediate feature distributions available for face recognition. So, we can use the result of facial detection process for face recognition without detecting additional new features.

We describe feature vectors and the procedure for their generation in face recognition.

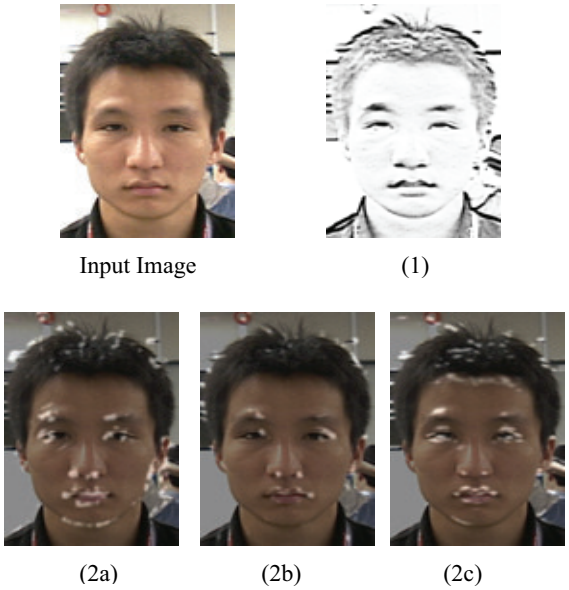

Fig.3 Input image and intermediate feature distributions for face recognition. (1) the 8th edge-like feature in FD1 layer (2a) '<' end-stop feature in FD2 layer (2b) '>'end-stop feature (2c) upper part bright blob feature (In (2a)-(2c), input image is superimposed on each intermediate feature distributions.)

Figure 3 shows four intermediate features distributions used for face recognition among ones of face detection. Output destributions in a module detecting 8th feature in FD1 layer presents especially shape information of eye, mouth and nose without graduation of brightness in cheek area which is seen in input image. And the output distributions in three modules detecting '<' end-stop feature, '>'end-stop feature and upper part bright blob feature in FD2 layer lie on the points corresponding to eye and mouth corners, and upper line of eye. Setting the opportune areas, we can produce a feature vector which implicitly show shapes of eye, the distance of eyes and spatial arrangement of eyes, mouth and nose.

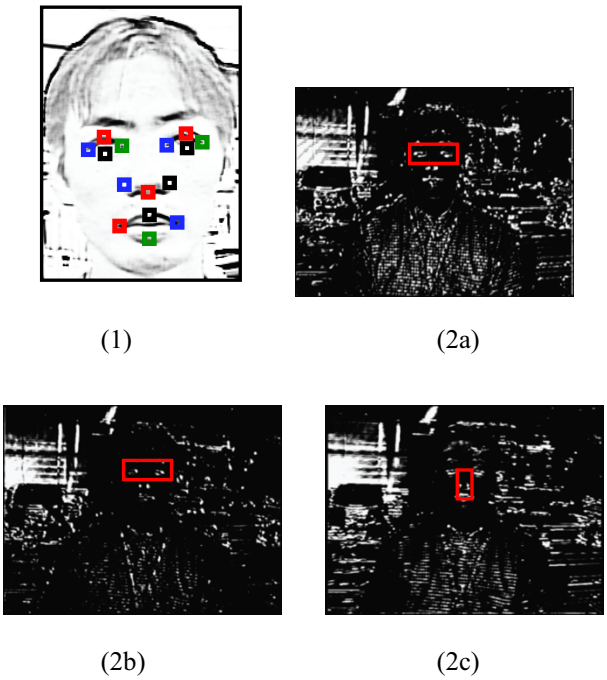

Fig.4 Local areas for feature vector extraction. (1) rectangle areas (15 x 15) set around eye-corners, mouth corners (2) rectangle areas (125 x 65: a,b; 45 x 65:c) defined for FD2 output.

A feature vector, ***F***, used in SVM for face recognition is an *N* dimensional vector, synthesized from a set of local

output distributions, $\boldsymbol{F}_1$ (as shown in Fig.4(1)), in a module detecting edge-like feature (8th feature) in FD1 layer in addition to output distributions, $\boldsymbol{F}_2$, (as shown in Fig.4(2)) of three intermediate-level modules detecting '<' end-stop feature, '>'end-stop feature and upper part bright blob feature in FD2 layer. Thus, $\boldsymbol{F} = (\boldsymbol{F}_1, \boldsymbol{F}_2)$ where $\boldsymbol{F}_1 = (\boldsymbol{F}_{11}, \ldots, \boldsymbol{F}_{1m})$ and $\boldsymbol{F}_2 = (\boldsymbol{F}_{21}, \ldots, \boldsymbol{F}_{2n})$ are synthesized vectors formed by component vectors, $\boldsymbol{F}_{1k}$ ($k=1, \ldots, m$) and $\boldsymbol{F}_{2k}$ ($k=1, \ldots, n$). Here, m and n are the number of local areas for $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ component vectors, respectively. Each component vector represents possibility or presence of specific class of local feature in an assigned local area. Dimension of a component vector is the area of a rectangular region as in Fig.4. Thus dimension of feature vector, N, is the total summation of respective dimensions of component vectors. In particular, $\boldsymbol{F}_1 = (\boldsymbol{F}_{11}, \boldsymbol{F}_{12}, \ldots, \boldsymbol{F}_{1,15})$, and local areas, total number of assigned areas being 15 as in Fig.4 (1), for component vectors are set around eye, nose, and mouth, using the detected eye location from the CNN. The results of detected eye and mouth location are shown in figure 5 by high intensity pixels, respective points for eye location are determined from maximum point of distribution of internal value of neurons in eye detection module. In our experiment, the average accuracy of positions of detected eyes and mouth was within 1pixel.

Fig 5. Detected positions of eyes and mouth decided with those intermediate features distributions in face detection.

$\boldsymbol{F}_1$ reflects shape information of eye, mouth, and nose. $\boldsymbol{F}_2 = (\boldsymbol{F}_{21}, \boldsymbol{F}_{22}, \boldsymbol{F}_{23})$, and each component vector reflects spatial arrangement of eye and mouth or nose, etc., depending on how local areas in FD2 (e.g., positions and size) are set.

Positions of local areas in FD1 module are set around specific facial components (i.e., eyes, mouth) as illustrated in Fig. 4 (1). The size of respective local areas in the output plane of FD1 module is set relatively small (e.g., 11 x 11) so that local shape information of figural alphabets can be retained in the output distribution, while the local area in the FD2 plane is relatively larger (e.g., 125 x 65) so that information concerning spatial arrangement of facial components (e.g., eye) is reflected in the distribution of FD2 outputs.

3 Results

As in [10], [11], training of the CNN is performed module by module using fragment images as positive data extracted from database (e.g., Softpia Japan) of more than 100 persons. Other irrelevant fragment images extracted from background images are used as negative samples. The size of partial images for the training is set so that only one class of specific local feature is contained.

For face recognition, we use an array of linear SVMs, each trained for one-against-one multi-class recognition of faces. The SVM library used in the simulation is *libsvm2.5*, available in the public domain. In the SVM training, we used a dataset of feature vectors (FVs) extracted from pictures took at various places under varying image capturing conditions, for example, in the laboratory or at the cafeteria etc.

The size of input image is VGA, and as illustrated in Fig.4, the size of local areas for FVs is 15 x15, 125 x 65, or 45 x 65 depending on the class of local features. As indicated in Fig.4, the number of local areas for FD1 feature and FD2 feature is fourteen and two, respectively. The number of FVs for one person is 30, which are obtained at various places under varying image capturing conditions so that size, pose, and lightning conditions of respective faces are slightly different. In our experiments, the differences of face size are from 0.5 to 1.5 (relative size in units of area for reference face), rotation in plane is from minus 15 degree to plus 15 degree, rotation in depth is almost zero degree, the differences of average brightness in face area is in about twice.

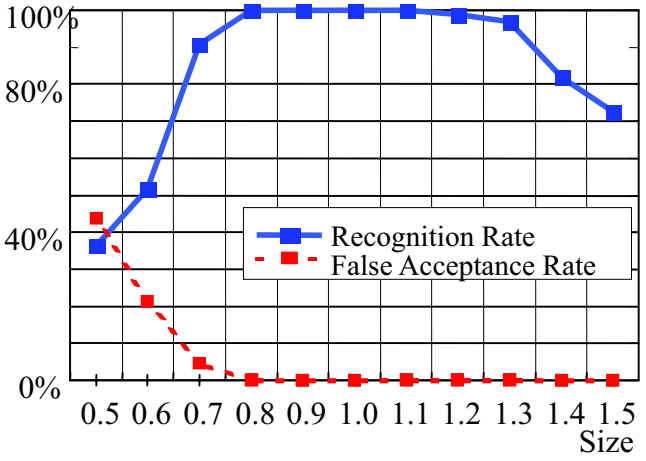

Fig.6 Robust face recognition performance for 20 people under varying sizes of face

The result shown in Fig.6, obtained using test images, different from training data, indicates robustness to size variability from 0.8 to 1.2 (relative size in units of area for reference face), demonstrating 100% recognition with 0% false acceptance rate.

4 Summary

We presented a new model of object recognition with economical use of low or intermediate level features. The preliminary results demonstrated robustness and efficiency in face recognition combined with detection, with 100 % recognition rate and 0% F.A.R. for 600 images of 20 people. Test images were with complex background, captured at the various places under varying conditions, including size variability.

The novelty of proposed model lies in simple and effi-

cient mechanism of object recognition that involves extracting alphabetical local features in CNN for face detection and generating support vectors in SVM from local area-based-feature vectors from intermediate outputs in CNN, both require relatively simple linear operation. This mechanism is in contrast with renowned method [14] in that our proposed model does not require computationally expensive and biologically implausible mechanism, a graph matching in DLM.

The computational procedure proposed mainly involves only weighted summation of inputs, implemented both in CNN and in linear SVM. Thus the approach presented in this study can be described in a common framework of relatively simple neuronal computation, weighted summation of inputs. This can be directed to incorporate as substrate for general model of object recognition and detection.

Acknowledgements : Some facial data in this paper are used by permission of Softpia Japan, Reseach and Development Division, HOIP Laboratory. It is strictly prohibited to copy, use or distribute the facial data without permission.

References

1. Belhumeur, P., Hesoanha, P., Kriegman, D.: Eigenfaces vs fisherfaces: recognition using class specific linear projection. IEEE Trans. on Pattern Analysis and Machine Intelligence **19** (1997) 711-720
2. Brunelli, R., Poggio T.: Face recognition: features versus templates. IEEE Trans. on Pattern Analysis and Machine Intelligence **15** (1993) 1042-1052
3. Turk, M., Pentland, A.: Face recognition using eigenfaces. Proc. IEEE Conf. On Computer Vision and Pattern Recognition (1991) 586-591
4. Fukushima, K.: Neocognitron: a self-organizing neural networks for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics **36** (1980) 193-202
5. Guodong, G., Li, S., Kapluk, C.: Face recognition by support vector machines. Proc. IEEE International Conf. On Automatic Face and Gesture Recognition (2000) 196-201
6. Heisele, B.,Ho, P., Poggio, T.: Face recognition with support vector machines: global versus component-based approach. Proc. International Conf. on Computer Vision (2001) 688-694
7. Heisele, B., Koshizen, T.: Components for Face Recognition Proc. IEEE International Conf. on Automatic Face and Gesture Recognition (2004)
8. Le Cun, Y., Bengio, T.: Convolutional networks for images, speech, and time series. In: Arbib, M.A. (ed.): The handbook of brain theory and neural networks, MIT Press, Cambridge (1995) 255-258
9. Li, Y., Gong, S., Liddel, H.: Support vector regression and classification based multi-view face detection and recognition. Proc. IEEE International Conf. on Automatic Face and Gesture Recognition (2000) 300-305
10. Matsugu, M., Mori, K., Ishii, M., Mitarai, Y.: Convolutional spiking neural network model for robust face detection. Proc. International Conf. on Neural Information Processing (2002) 660-664
11. Mitarai, Y.,Mori, K., Matsugu, M.: Robust Face Detection System Based on Convolutional Neural Networks Using Selective Activation of Modules (In Japanese). Proc. Forum in Information Technology (2003) 191-193
12. Moghaddam, B., Wahid, W., Pentland, A.: Beyond eigenfaces: probabilistic matching for face recognition. Proc. IEEE International Conf. on Automatic Face and Gesture Recognition (1998) 30-35
13. Pontil, M., Verri, A.: Support vector machines for 3-d object recognition. IEEE Trans.on Pattern Analysis and Machine Intelligence **20** (1998) 637-646
14. Wiskott, L., Fellous, J.-M., Krüger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Trans. on Pattern Analysis and Machine Intelligence **19** (1997) 775-779