Head-pose invariant Facial Expression Recognition using Convolutional Neural Networks

Beat Fasel [a]

IDIAP–RR 02-51

November 2002

published in
Proceedings of ICMI 2002, Pittsburgh, USA

Dalle Molle Institute
for Perceptual Artificial
Intelligence • P.O.Box 592 •
Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP - Institut Dalle Molle d'Intelligence Artificielle Perceptive Rue du Simplon 4, CP592 - 1920 Martigny, Switzerland Beat.Fasel@idiap.ch

IDIAP Research Report 02-51

Head-pose invariant Facial Expression Recognition using Convolutional Neural Networks

Beat Fasel

November 2002

published in
Proceedings of ICMI 2002, Pittsburgh, USA

Abstract. Automatic face analysis has to cope with pose and lighting variations. Especially pose variations are difficult to tackle and many face analysis methods require the use of sophisticated normalization and initialization procedures. We propose a data-driven face analysis approach that is not only capable of extracting features relevant to a given face analysis task, but is also more robust with regard to face location changes and scale variations when compared to classical methods such as e.g. MLPs. Our approach is based on convolutional neural networks that use multi-scale feature extractors, which allow for improved facial expression recognition results with faces subject to in-plane pose variations.

1 Introduction

Most automatic facial expression analysis approaches presented in the literature need some kind of manual intervention during training, such as the construction of face models [11][3][1] or during testing due to necessary initialization procedures, such as the precise localization of facial features, e.g. see [10] in order to perform reliably. Several data-driven face analysis methods have been described in the literature and comprise among others neural network based approaches, e.g. [9] and PCA-based methods, e.g. [2]. However, numerous data-driven face analysis approaches need accurate face normalization preprocessing stages. In this paper, we propose convolutional neural network (CNN)[6] based approaches for the task of facial expression recognition and compare them to classical MLPs using the same database. CNNs, as well as the similar neocognitrons [4], are bio-inspired hierarchical multi-layered neural network approaches that model to some degree characteristics of the human visual cortex and encompass scale and translation invariant feature detection layers. Convolutional neural networks have been successfully applied for character recognition [7], object detection [7] and more specifically for the task of face recognition [5].

2 Convolutional Neural Networks

Figure 1 shows the architecture of a convolutional neural networks we trained for the task of facial expression recognition. Its layers alternate between convolution layers with feature maps $C^i_{k,l}$

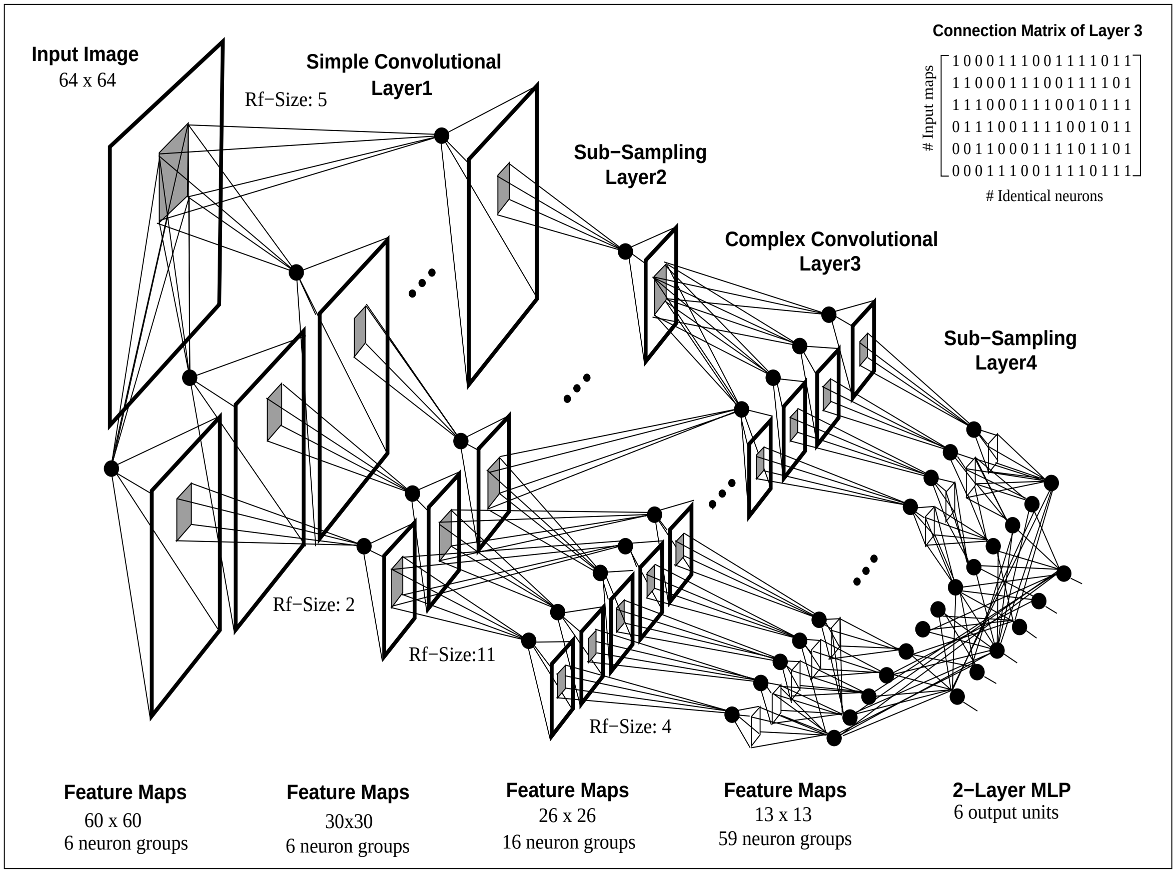
$$C^i_{k,l} = g(I^i_{k,l} \otimes W_{k,l} + B_{k,l})$$

and non-overlapping sub-sampling layers with feature maps $S^i_{k,l}$

$$S^i_{k,l} = g(I \downarrow^i_{k,l} w_{k,l} + Eb_{k,l})$$

where $g(x) = \tanh(x)$ is a sigmoidal activation function, B, respectively b the biases, W and w the weights, $I^i_{k,l}$ the ith input and $I \downarrow^i_{k,l}$ the down-sampled i'th input of the neuron group k of layer l. E is a matrix whose elements are all one and $\otimes$ denotes a 2-dimensional convolution. Note that upper case letters represent matrices, while lower case letters denominate scalars. In our first network setup, as shown in Figure 1, we chose receptive fields sizes of 5×5 pixels for the groups of neurons in the first simple feature extraction layer and a receptive field size of 11×11 pixels in the third complex feature extraction layer, respectively sizes of 2×2 and 4×4 pixels for the receptive fields of the sub-sampling layers. We implemented larger receptive fields in the complex layer in order to allow for an integration of features found by the preceding convolutional layer. Our second convolutional neural network setup is shown in Figure 4. It features receptive fields of different sizes (5×5, 7×7 and 9×9) that operate at slightly different scales on the input image. This allows for the extraction of features of different sizes within a given object of interest. The architecture of the convolutional neural network shown in Figure 1 can be described as follows: A6*5x5-B6*2x2-C16*11x11-B59*4x4-mlp2. Hereby, simple convolutional layers are denoted as A, complex convolutional layers with C and sub-sampling layers with B. The expression A6*5x5 means that there are six neuron groups with receptive fields sizes of 5x5 present in the convolutional layer A.

The learned weights of the convolutional layers allow for problem-at-hand dependent feature extraction, whereas the sub-sampling layers increase the invariance of the object of interest's location dependence. Weight sharing allows to significantly reduce the number of free parameters, which in turn improves the generalization ability [6]. For example, the number of neuron interconnections in the feature extraction layers of the network architecture shown in Figure 1 is 3367308, while the number of weights amounts only to 1902 and the number of neurons to 47787. The number of neuron interconnections in the 2-layer MLP following the feature extraction layers is however 997700 with the same number of weights for 100 neurons in the first layer and 6 neurons in the output layer. This shows that even though the feature extraction layers are huge with regard to the number of neuron



$$\begin{bmatrix}
1&0&0&0&1&1&1&0&0&1&1&1&1&0&1&1\\
1&1&0&0&0&1&1&1&0&0&1&1&1&1&0&1\\
1&1&1&0&0&0&1&1&1&0&0&1&0&1&1&1\\
0&1&1&1&0&0&1&1&1&1&0&0&1&0&1&1\\
0&0&1&1&0&0&0&1&1&1&1&0&1&1&0&1\\
0&0&0&1&1&1&0&0&1&1&1&1&0&1&1&1
\end{bmatrix}$$

Figure 1: Depicted is the architecture of a 5-layer convolutional neural network (with 2 feature extraction, 2 sub-sampling and two fully connected MLP layers), which we applied for head-pose independent facial expression analysis. Note that the larger dots represent groups of identical neurons. The complex convolutional layer three is not fully connected, its connection matrix is shown in the upper corner on the right-hand side.

interconnections, only a few weights need to be trained. Most weights have to be adapted in the MLP connected to the feature layers.

Face images I_{in} at the input of the CNNs were not pose-normalized, but only global lighting changes were addressed by removing the mean value $\overline{I_{in}}$ of the images contained in the training base. In order to increase the learning speed, we normalized also the variances of the input variables by dividing them by their standard deviation σ_{in} of the images of the training set: $I_{norm} = \frac{I_{in}-\overline{I_{in}}}{\sigma_{in}}$. No attempts were taken to reduce image dimensionality by using e.g. holistic PCA as demonstrated in [5]. Instead, we relied on the kernels of the feature extraction layers to perform decorrelation of the input data. Holistically applied PCA without using sophisticated pose normalization procedures would attempt to represent pose information, which is not desired, as there are too many pose variations present in natural face images (due to translation, rotation and scale).

Training of our CNNs was achieved in a supervised manner by using the standard back-propagation algorithm, adapted for convolutional neural networks. The weight and bias deltas for the feature extraction kernels in the convolutional layers (C) are

$$\Delta W^C_{t,k} = l_R \sum_{i=1}^{F} (I^L_i \otimes D^H_i) + m_R \Delta W^C_{t-1,k}$$

$$\Delta B^C_{t,k} = l_R \sum_{i=1}^{F} D^H_i + m_R \Delta B^C_{t-1,k}$$

while the weight and bias deltas for the sub-sampling layers (S) are as follows

$$\Delta w^S_{t,k} = l_R \sum_{i=1}^{F} \sum_{m=1}^{M_i} \sum_{n=1}^{N_i} (I \downarrow^L_i \times D^H_i) + m_R \Delta w^S_{t-1,k}$$

$$\Delta b^S_{t,k} = l_R \sum_{i=1}^{F} \sum_{m=1}^{M_i} \sum_{n=1}^{N_i} D^H_i + m_R \Delta b^S_{t-1,k}$$

I^L_i is the input image i, $I \downarrow^L_i$ a down-sampled version of the input image i of the lower layer L, D^H_i is the error delta coming from the higher layer H. $\otimes$ denotes a 2-dimensional convolution and $\times$ a component-vise matrix multiplication. F is the number of connected input feature maps of the current neuron group k, M_i and N_i the number or rows, respectively columns of the feature map i. l_R is the learning rate and m_R the moment rate.

3 Experiments and Results

We tested our neural network architectures on three different databases. Database set 1 consisted of the JAFFE facial expression database [8], which contains posed emotional facial expression images of 10 Japanese female subjects (6 different emotion displays and neutral face displays), see Figure 2.

The gray-scale images originally of size 256 × 256 pixels were reduced in scale to 64 × 64 pixels (in order to lower the information content that has to be learned by the networks and also in order to make training of the CNN networks faster and less memory consuming). We used a total of 140 images to train our neural networks and 70 images for testing (database set 1). In order to demonstrate the capability of CNNs when classifying facial expressions also in situations, where in-plane head pose variations come into play, we artificially increased the JAFFE database by shifting the images (up, down, left, right), zoomed in and out as well as rotated the images both in clockwise and counter-clock wise orientation, see Figure 3. Database set 2 contains 1260 training images created from the 140 training images contained in database set 1 by applying the afore mentioned affine transformations. Furthermore, 630 images were created by not only applying the same affine transformations to the 70 test images of database set 1, but shifting all test images by 5 pixels to the right. Database set 3

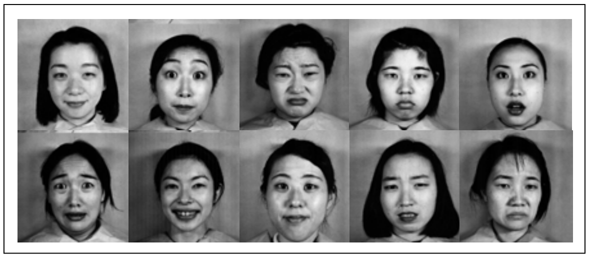


Figure 2: Sample images of the employed JAFFE facial expression database [8]. Note slight variations of the head position, scale and rotation.

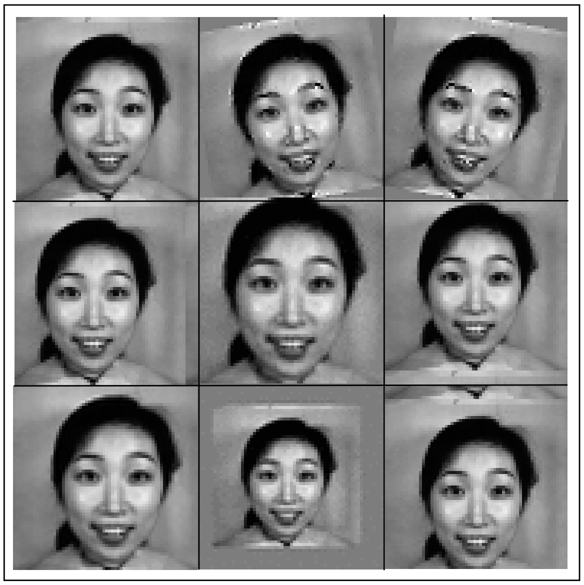


Figure 3: Above are shown some examples of database set 2/3 (Basic emotions labeled database), artificially extended database by introducing translational shifts (up, down, left, right), zoom in and out, as well as clockwise and counter-clockwise rotations.

CNN Network Architectures	**Corr. Recog.**	**Corr. Recog.**	**Corr. Recog.**
layer type - rf. sizes - classifier	Set 1 (140/70)	Set 2 (1260/630)	Set 3 (1260/630)
(1) A6*5x5-B6*2x2-C16*11x11-...	84.3%	53.7%	55.9%
B59*4x4-mlp2		(T49% R57% S49%)	(T50% R54% S58%)
(2) A6*5x5-6*7x7-6*9x9...	**91.4%**	**54.0%**	**57.0%**
B18*2x2-C16*11x11-B59*4x4-mlp2		(T49% R55% S51%)	(T51% R59% S54%)
(3) MLP1 (6)	88.6%	46.5%	45.7%
		(T40% R42% S44%)	(T40% R36% S48%)
(4) MLP3 (100-20-6)	90%	46.5%	45.4%
		(T41% R42% S44%)	(T35% R40% S49%)

Table 1: Facial expression recognition rates resulting by using two different convolutional neural network and MLP architectures on three different test sets. Architectures: A stands for simple feature extraction layers, B for sub-sampling layers, C for complex layers. Recognition Results: T stands for translated, R for rotated and S for scaled face images.

was constructed in a similar way as database set 2, but instead of shifting the test images, a scale change was applied (zooming out). The test images of the database set 1 and 2 allowed us to test our networks with faces appearing with previously unseen poses.

Table 1 lists the facial expression results obtained on the afore mentioned databases. As can be seen, classic MLPs as in network setup 3 (1-layer MLP) and 4 (3-layer MLP) performed almost equally well, also in comparison to the convolutional neural networks of setup 1 and 2. However, the featured MLPs lead to lower recognition results than the convolutional neural networks in setup 1 and 2 when applied on database set 2 and 3, where previously unseen face poses (zoom and translation) occur. For these database sets, we obtained the best recognition results with the convolutional neural network setup 2, which operates at three different resolutions, see also Figure 4.

Unfortunately, we cannot directly compare our facial expression recognition results with the ones Lyons et al. [8] obtained on the same database, as they computed facial expression similarities using semantic values stemming from human ratings, resulting in a mixture of facial expressions per analyzed face, while we used one category per facial expression. Lyons et al. argued that it is more appropriate to compare similarity spaces instead of measuring a categorization performance in order to avoid the problem of posed facial expressions that lead not always to pure facial expressions of one category. Lyons et al. reported an average rank correlation between their Gabor model and the ground truth semantic rating of 0.57 when including the fear stimuli and 0.68 without.

4 Conclusions

In order to analyze the capability of convolutional neuronal networks with regard to copying with in-plane head-pose variations, we artificially increased the size of the latter by shifting, zooming and rotating images of the original database. We were able to demonstrate that the employed CNN architectures recognized facial expressions with better results, when previously unseen pose variations occur. Our approach does not require extensive pose normalization, face segmentation or facial feature tracking initialization procedures. The only assumptions we made was, that the input image is coarsely centered around a single face to be analyzed. Furthermore, we were able to demonstrate that using receptive fields of different sizes in the input layer (in the simple feature extraction layer A) can lead to improved facial expression recognition results, especially in the context of in-plane pose variations. Further research has to be done with regard to the ability of convolutional neural networks to handle also larger databases and especially with regard to the number of test subjects involved.

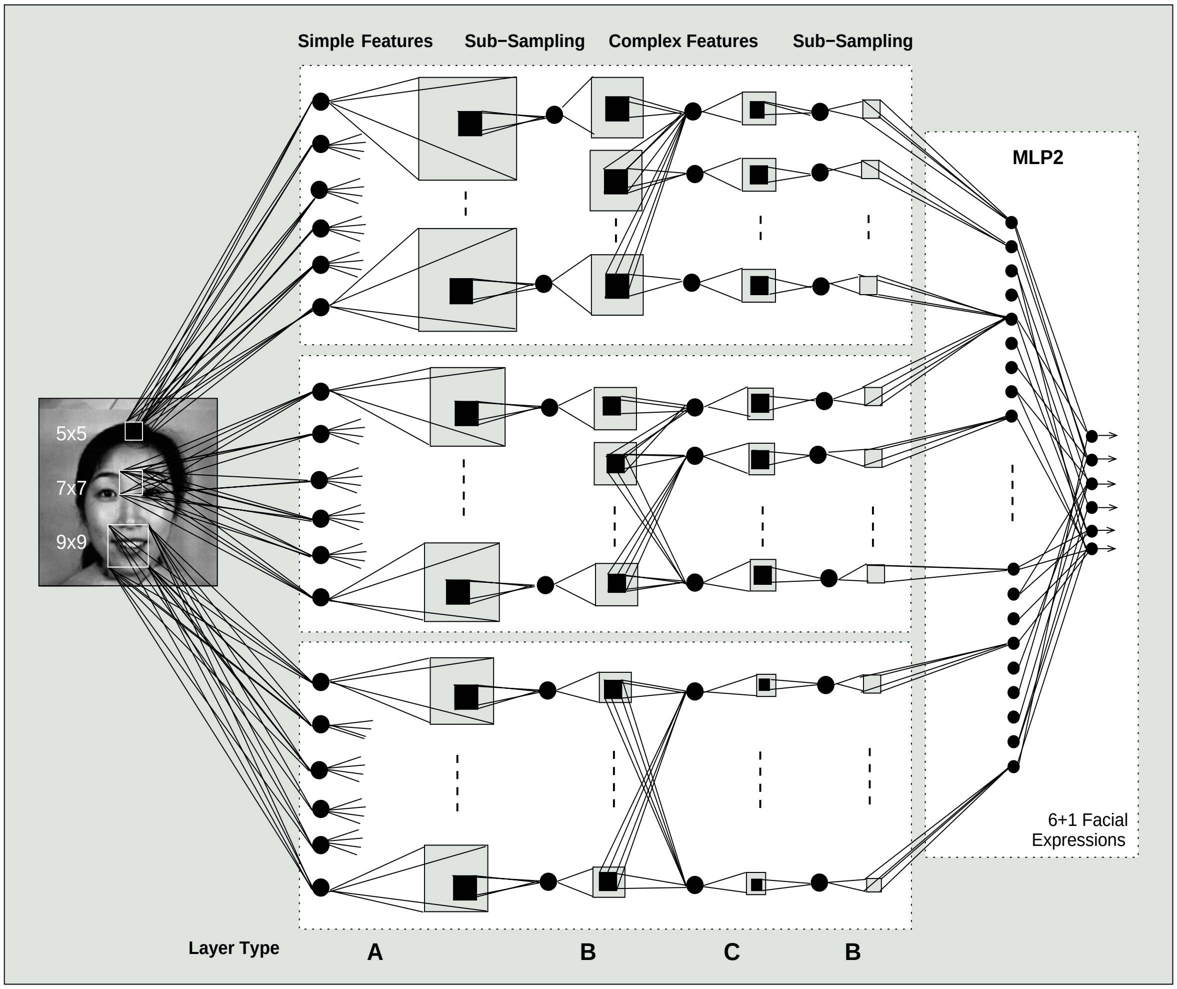


Figure 4: Multi-scale feature extraction convolutional neural network. The three highlighted subregions of the network operate at different resolutions and there is no connection in-between them. Integration only occurs in the MLP. Note also that the complex neurons in the third network layer integrate information stemming from several input feature maps into single output maps (summation).

5 Acknowledgments

Thanks go to the SNSF (Swiss National Science Foundation) for funding this project within the framework IM(2) under project number 21-54000.98.

References

[1] A. Lanitis, C.J. Taylor, and T. F. Cootes. Automatic Interpretation and Coding of Face Images using Flexible Models. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[2] Marian Stewart Bartlett. *Face Image Analysis by Unsupervised Learning and Redundancy Reduction*. PhD thesis, University of California, San Diego, 1998.

[3] Irfan A. Essa and Alex P. Pentland. Facial Expression Recognition using a Dynamic Model and Motion Energy. In *ICCV95*, 1995.

[4] Fukushima K. Neocognitron: A Self-Organizing Neural Network for a Mechanism of Pattern Recognition Unaffected by Sift in Position. *Biol Cybern*, 36:193–202, 1980.

[5] Steve Lawrence, C. Lee Giles, A.C. Tsoi, and A.D. Back. Face recognition: A convolutional neural network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, 1997.

[6] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time-series. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, 1995.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[8] Lyons M., Akamatsu S., Kamachi M., and Gyoba J. Coding Facial Expressions with Gabor Wavelets. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, April 1998.

[9] C. Padgett and G. W. Cottrell. A Simple Neural Network Models Categorical Perception of Facial Expressions. In *Proceedings of the Twentieth Annual Cognitive Science Conference*, 1998.

[10] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing Action Units for Facial Expression Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2), February 2001.

[11] Yaser Yacoob and Larry Davis. Computing Spatio-Temporal Representations of Human Faces. Technical report, Computer Vision Labratory, University of Maryland, 1994.