Facial Expression Analysis using Shape and Motion Information Extracted by Convolutional Neural Networks

Beat Fasel [a]

IDIAP–RR 01-49

December 2001

PUBLISHED IN
Proceedings of NNSP 2002, Martigny, Switzerland

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 01-49

Facial Expression Analysis using Shape and Motion Information Extracted by Convolutional Neural Networks

Beat Fasel

December 2001

published in
Proceedings of NNSP 2002, Martigny, Switzerland

Abstract. In this paper we discuss a neural networks-based face analysis approach that is able to cope with faces subject to pose and lighting variations. Especially head pose variations are difficult to tackle and many face analysis methods require the use of sophisticated normalization procedures. Data-driven shape and motion-based face analysis approaches are introduced that are not only capable of extracting features relevant to a given face analysis task at hand, but are also robust with regard to translation and scale variations. This is achieved by deploying convolutional and time-delayed neural networks, which are either trained for face shape deformation or facial motion analysis.

1 Introduction

Many face analysis approaches require manual intervention during training, such as the construction of face models or during deployment, due to necessary initialization, such as the precise localization of facial features. Several data-driven face analysis methods have been described in the literature and comprise among others neural network and PCA-based approaches. However, numerous data-driven face analysis approaches need accurate face normalization preprocessing stages. In this paper, we propose a convolutional neural network (CNN)[3] based approach that improves a specific face analysis task by combining the output of differently trained convolutional neural networks in a fusion-MLP (multi-layer perceptron). CNNs, as well as the similar neocognitrons [1], 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 [4], object detection [4] and more specifically, for the task of face recognition [2]. As facial expressions can be characterized not only by shape information (facial deformation), we investigated two-dimensional time-delay neural networks (TDNN) for the extraction of facial motion as well. TDNN have been successfully employed to recognize e.g. pedestrians [6].

2 Face Analysis Systems

2.1 Convolutional Neural Network

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

$$C^i_{k,l} = g(I^i_{k,l} \otimes W_{k,l} + B_{k,l}) \tag{1}$$

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

$$S^i_{k,l} = g(I \downarrow^i_{k,l} w_{k,l} + Eb_{k,l}) \tag{2}$$

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. We obtained good results by choosing receptive fields sizes of 11×11 pixels for the groups of neurons in the first feature extraction layer and 8×8 pixels in the third feature extraction layer, respectively 2×2 pixels for the receptive fields of the sub-sampling layers. 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 [3]. This can also be seen in Figure 2, where the number of neuron-interconnections in the CNN is much greater than the number of weights to be learned.

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 $\overline{I_{in}}$. In order to increase the learning speed, we norm also the variances of the input variables by dividing them by their standard deviation σ_{in}: $I_{norm} = \frac{I_{in} - \overline{I_{in}}}{\sigma_{in}}$. No attempts were taken to reduce image dimensionality by using e.g. holistic PCA as demonstrated in [2]. 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 changes).

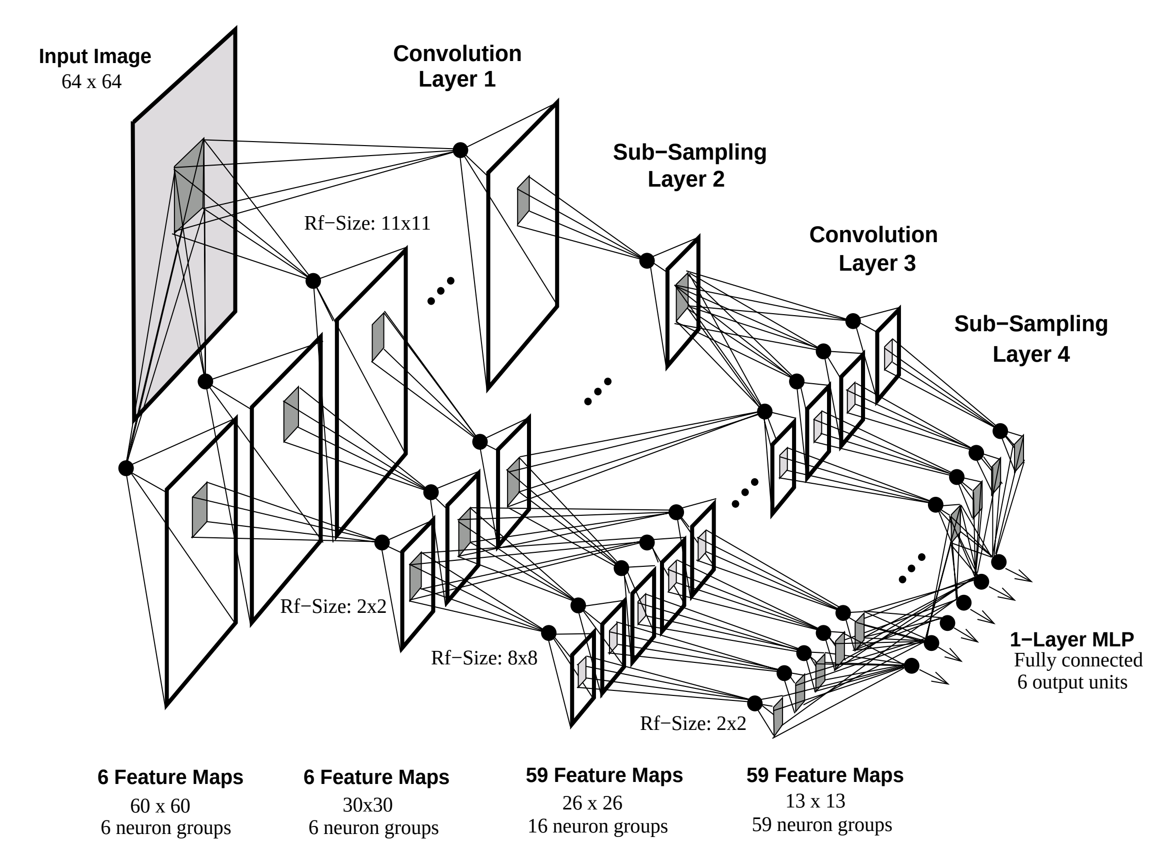

Figure 1: Depicted is the architecture of a 5-layer convolutional neural network (with 2 feature extraction, 2 sub-sampling and one fully connected MLP layer), which we applied for facial shape analysis. Note that the larger dots represent groups of identical neurons. The shown architecture does not feature full connectivity in the feature extraction layer 3, see also Figure 2.

Connection Matrix of Layer 3

# Input maps

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

# Identical neurons

CNN + MLP Characteristics

Nb. of neurons CNN:	47787
Nb. of connections CNN:	3367308
Nb. of weights in CNN:	1902
Nb. of neurons MLP:	6
Nb. of connections MLP:	59826
Nb. of weights in MLP:	59826

Figure 2: Depicted on the left hand side is the interconnection matrix of the third layer of the CNN we trained for shape as well as motion recognition in the setups 3 and 6 listed in Table 1. This interconnection matrix leads to 59 feature maps (number of ones). On the right hand side are given the corresponding number of weights, neurons and neuron inter-connections for the feature extraction (CNN) and feature combination part (MLP).

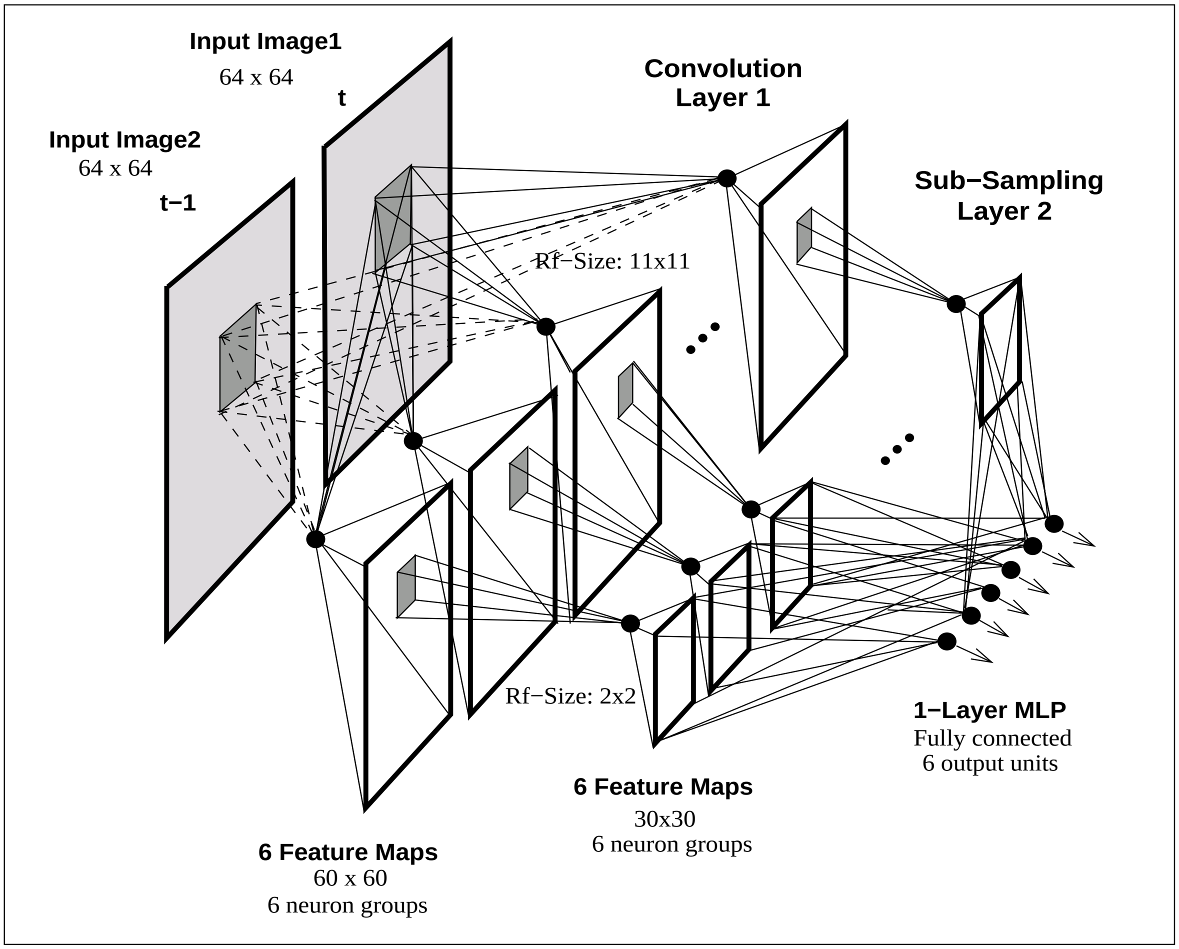

Figure 3: Sample Architecture of a 2-layer convolutional time delay neural network (TDNN), featuring a temporal horizon of 2 (two inputs: at time t-1 and t). Notice that the first layer is of spatio-temporal nature, while the second layer operates only in the spatial domain.

2.2 Time Delay Neural Network

Figure 3 shows a sample architecture of a convolutional two-dimensional time delay neural network that we have employed for the analysis of facial motion.

As can be seen, the network features two inputs at time *t-1* and *t* (temporal horizon of two). Hereby, the first network layer operates in the spatio-temporal domain, while the second layer (and all further layers) is purely spatial. The only difference to a convolutional neural network is thus the summation of of the inputs into a single feature map $C^i_{k,1}$

$$C^i_{k,1} = g(\sum_{i=1}^{G}(I^i_{k,l} \otimes W_{k,l} + B_{k,l})) \tag{3}$$

where G is the temporal horizon, being equal to 2 throughout this paper. This means that we chose a neutral face at time t-1 and a face showing facial expressions at time t (with only one facial expression intensity level). Note that this is of course a simplification of a real-world situation, where we would have both increasing and decreasing facial expressions intensities.

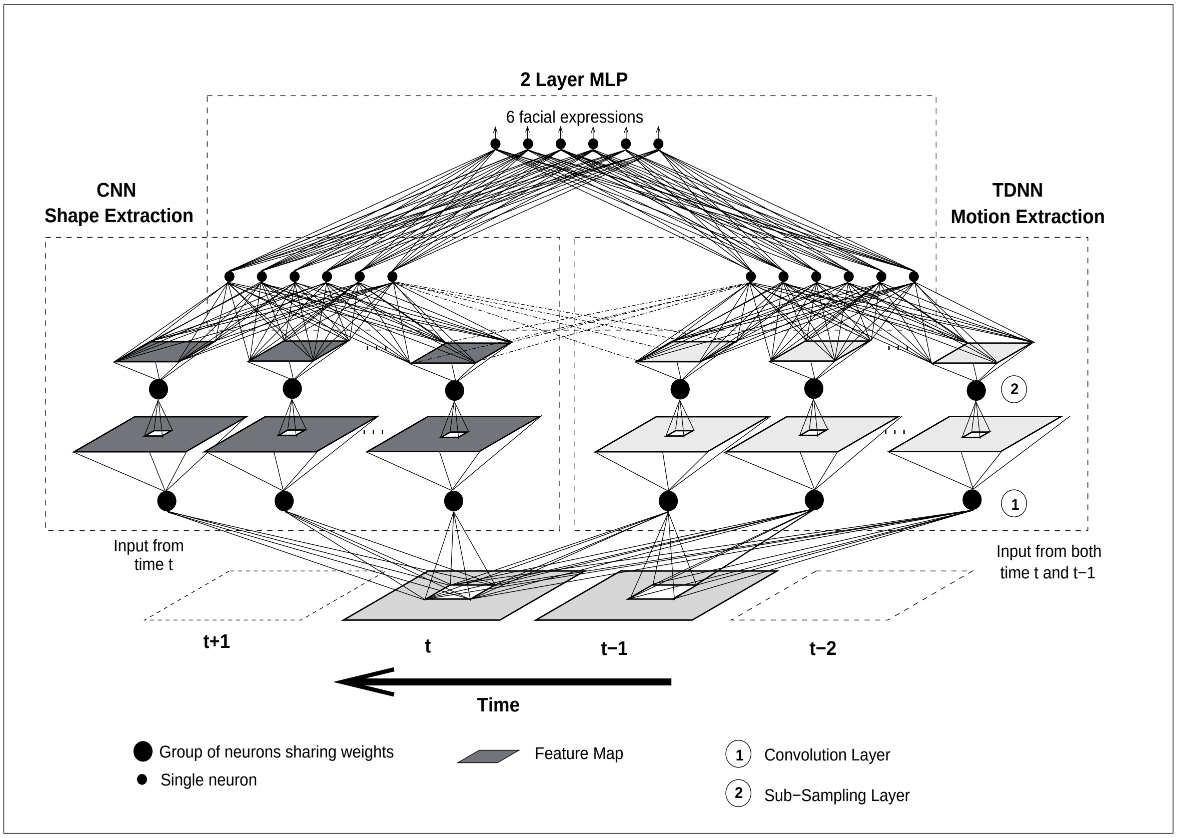

Figure 4: Combined motion and shape analysis for enhanced facial expression recognition: shown are three sub-networks, namely a shape extraction CNN, a motion extraction TDNN and an MLP for combined classification. This network architecture was trained as a whole or separately, according to the different sub-networks involved.

2.3 Network Combination

CNNs and TDNNs cannot only be used to analyze shape and motion independently, but they may be combined in order to improve facial expression recognition results by focusing on different features, respectively producing different error patterns, see Figure 4.

Network combinations can be achieved in two different ways:

- A CNN shape recognition and a TDNN motion recognition sub-network are trained independently from each other. In a second step, a MLP sub-network is trained on the output of the former two networks and thus allows to fuse the information stemming from the shape and the motion sub-networks, see network setup 7 in Table 1.

- The CNN, TDNN and the fusion MLP are trained at once, compare network setup 8 in Table 1.

Note that both setups lead to a difference in how motion features are extracted in the TDNN sub-network as the resulting target vector during training is not the same. In the first case we chose as target vectors $0 \rightarrow 0$ and $0 \rightarrow x$ for the TDNN sub-network, where 0 is a neutral face at time *t-1* and x is a face showing facial expressions at time *t.* The CNN sub-networks on the other hand were attributed target vectors of 0 for neutral faces and x for a faces showing facial expressions. Note that the CNN sub-networks operate in the present (at time *t).* The same labeling as for the previously mentioned sub-networks was chosen for the independent CNNs and TDNNs in the setups 1-3, respectively, setups 4-6 as well as in setup 7 listed in Table 1. In total we chose $6 + 1 = 7$ target classes (6 different facial expressions and 1 neutral face per subject).

2.4 Network Training

Training of our CNNs and TDNNs 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

$$\Delta W^C_{t,k} = l_R \sum_{i=1}^{F} (I^L_i \otimes D^H_i) + m_R \Delta W^C_{t-1,k} \tag{4}$$

$$\Delta B^C_{t,k} = l_R \sum_{i=1}^{F} D^H_i + m_R \Delta B^C_{t-1,k} \tag{5}$$

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

$$\Delta w^S_{t,k} = l_R \sum_{i=1}^{F} \sum_{m=1}^{M_i} \sum_{n=1}^{N_i} (I \downarrow^L_i \times D^H_i) + m_R \Delta w^S_{t-1,k} \tag{6}$$

$$\Delta b^S_{t,k} = l_R \sum_{i=1}^{F} \sum_{m=1}^{M_i} \sum_{n=1}^{N_i} D^H_i + m_R \Delta b^S_{t-1,k} \tag{7}$$

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 setups on the JAFFE facial expression database [5], which contains posed emotional facial expression images of 10 Japanese female subjects (6 different emotion displays), see Figure 5.

The grayscale 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 make training of the CNN networks faster). We used 140 images to train our neural networks and 70 images for testing. Furthermore, we created a second test set by using the afore mentioned test images and shifting them 3 pixels upwards, downwards, to the left and to the right, resulting in 280 additional images (thus 350 in total). The employed database was too small in order to allow for a validation set. Cross-validation was neither performed, as the training of the convolutional neural networks is time consuming (due to the important number of convolutions and sub-sampling operations taking place in the CNNs). Instead, we trained our convolutional neural networks until a small error was obtained on the training images (which occurred after about 250 epochs). This is of course not optimal, but our results should be more of a qualitative than quantitative nature.

Table 1 shows the facial expression recognition results obtained on the afore mentioned database. The neural network setups 1-3 use a single CNN as shown in Figure 1, while setup 4-6 are based on TDNNs. Finally, setup 7 and 8 correspond to the combined shape and motion recognition network architectures depicted in Figure 4. Note that setup 2 and 5 use different receptive field sizes within the same network layer. This allows for a multi-scale feature analysis within a given object of interest

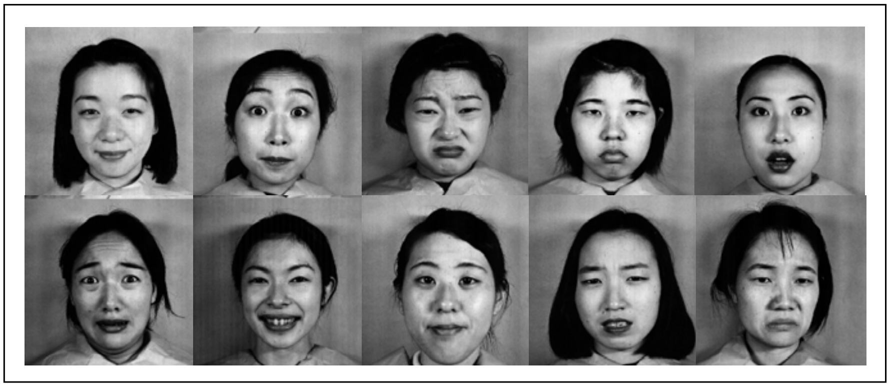

Figure 5: Sample images of the employed JAFFE facial expression database [5]. Note slight variations with regard to head positions, scale and rotation.

Network Setup	**Network Architecture**	**Task**	**Corr. Rec.**
(1) CNN	A-6x11-B-mlp1	Shape	90% (52%)
(2) CNN	A-6x5-6x7-6x11-B-mlp2	Shape	92% (50%)
(3) CNN	A-6x11-B-A-59x8-B-mlp1	Shape	91% (54%)
(4) TDNN	A-6x11-B-mlp1	Motion	84% (47%)
(5) TDNN	A-6x5-6x7-6x11-B-mlp2	Motion	84% (48%)
(6) TDNN	A-6x11-B-A-59x8-B-mlp1	Motion	83% (50%)
(7) CNN+TDNN	2 x (A-6x11-B-mlp1) + mlp1	Sh.+Mot.	93% (52%)
(8) CNN-TDNN	A-12x1-B-mlp1	Sh.+Mot.	92% (53%)
(10) MLP	mlp2-100	Shape	89% (40%)
(11) MLP	mlp2-100	Motion	73% (39%)

Table 1: This table shows facial expression recognition results based on shape, motion or a combined shape and motion analysis. Note that in the network column ***A*** stands for convolutional layer, ***B*** for a sub-sampling layer and ***a*** in ***axb*** for the number of square receptive fields, respectively their size ***b***. MLP stands for multi-layer perceptron, e.g., MLP1 for a 1-layer MLP and mlp2-100 for a 2-layer MLP with 100 hidden neurons. Recognition results are depicted for test set 1 (70 images) and in brackets for the shifted test images of test set 2 (350 images).

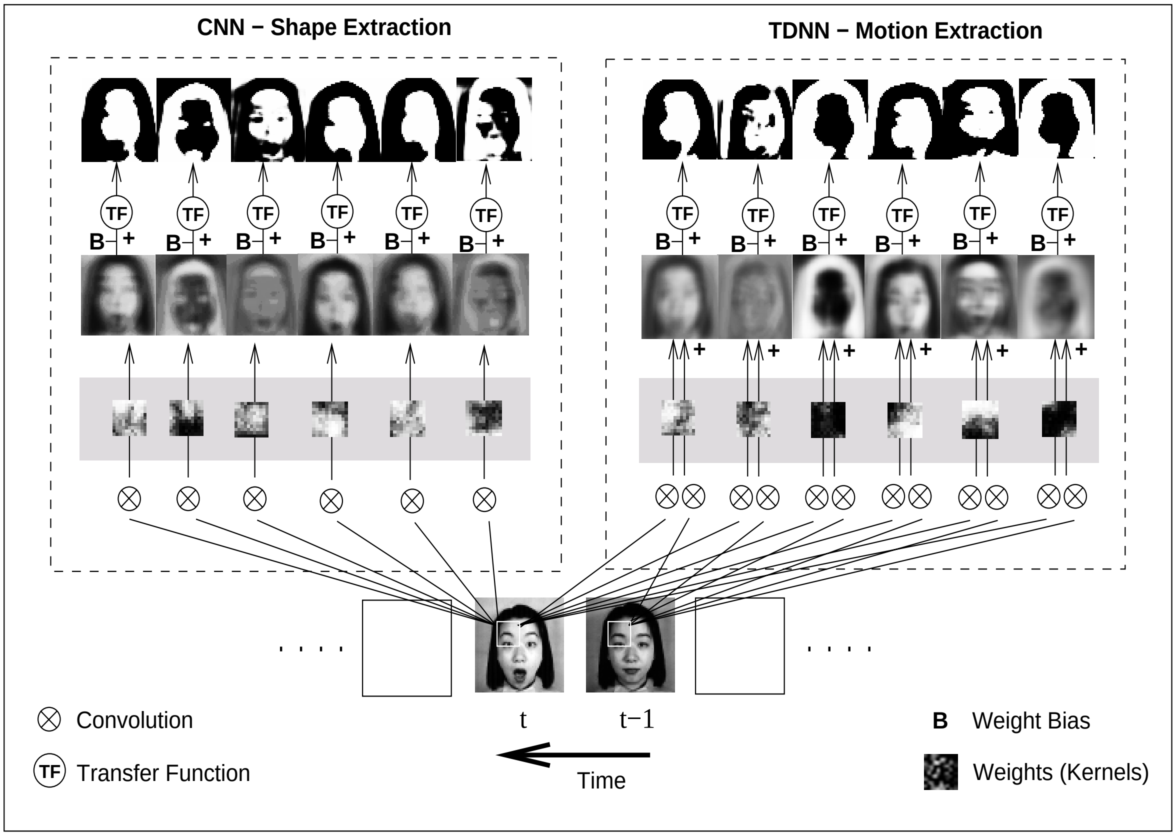

Figure 6: Data-driven feature extraction: Shown are the feature extraction layers of a 2-layer CNN on the left-hand side and a 2-layer TDNN on the right-hand side with task specific feature extraction kernels (of size 11 × 11) for shape-based and motion-based facial expression analysis. Note that we have enhanced the contrast of both the convoluted images and the weight kernels in order to improve readability.

(taken for granted that the receptive fields are smaller than the object of interest itself). The latter setups achieved slightly better recognition results than setups using a unique receptive field size per layer. However, the employed database is too small in order to obtain significant differences in recognition rates. The results suggest that facial shape deformations are more reliable than facial motion for identifying facial expressions. Clearly, both CNNs and TDNNs lead to better results than the MLPs employed in setup 10 and 11 and this especially for the second test set, which contains shifted images (recognition results for the second test set are given in brackets). Also note that the number of weights to be learned is considerably smaller in the CNNs than in comparable MLPs (e.g. 61728 weights for setup 3 versus 410200 weights for the MLP in setup 10). Unfortunately, we cannot compare our facial expression recognition results with the ones Lyons and Akamatsu [5] 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. Figure 6 illustrates different feature extraction kernels obtained for the tasks of face shape and motion analysis as well as their application onto two sample input face images.

4 Conclusion

In this paper we have shown that convolutional neural networks can be applied both in space and time for the analysis of facial expressions without relying on a complex and error-prone face pose normalization stage. Face deformation as well as facial motion are important indicators for facial expression and by combing a shape and motion extraction convolutional neural network, we were able to extract more information from a sequence of images. As we have seen, convolutional weight kernels are learned with regard to the task at hand and a combination of shape and motion extraction leads to slightly better facial expression recognition results, especially in the context of image transformations (here shown for shifted images) and when compared to e.g. MLPs. Shape information allowed for better recognition results when compared to motion extraction only. Using varying receptive fields sizes within the same layer also increased recognition results. The employed database is fairly small and therefore, further experiments have to be carried out in order to determine if our approach scales up reasonably well with regard to the number of different subjects involved as well as the number of employed facial expression classes by allowing also for facial expression intensity changes.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] C. Wohler, J. Aulanf, T. Portner, and U. Franke. A time delay neural network algorithm for real-time pedestrian recognition, 1998.