IDIAP RESEARCH REPORT

Robust Face Analysis using Convolutional Neural Networks

Beat Fasel [a]

IDIAP–RR 01-48

December 2001

published in
Proceedings of ICPR 2002, Quebec, Canada

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 01-48

Robust Face Analysis using Convolutional Neural Networks

Beat Fasel

December 2001

published in

Proceedings of ICPR 2002, Quebec, Canada

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 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 robust with regard to face location changes and scale variations. This is achieved by deploying convolutional neural networks, which are either trained for facial expression recognition or face identity recognition. Combining the outputs of these networks allows us to obtain a subject dependent or personalized recognition of facial expressions.

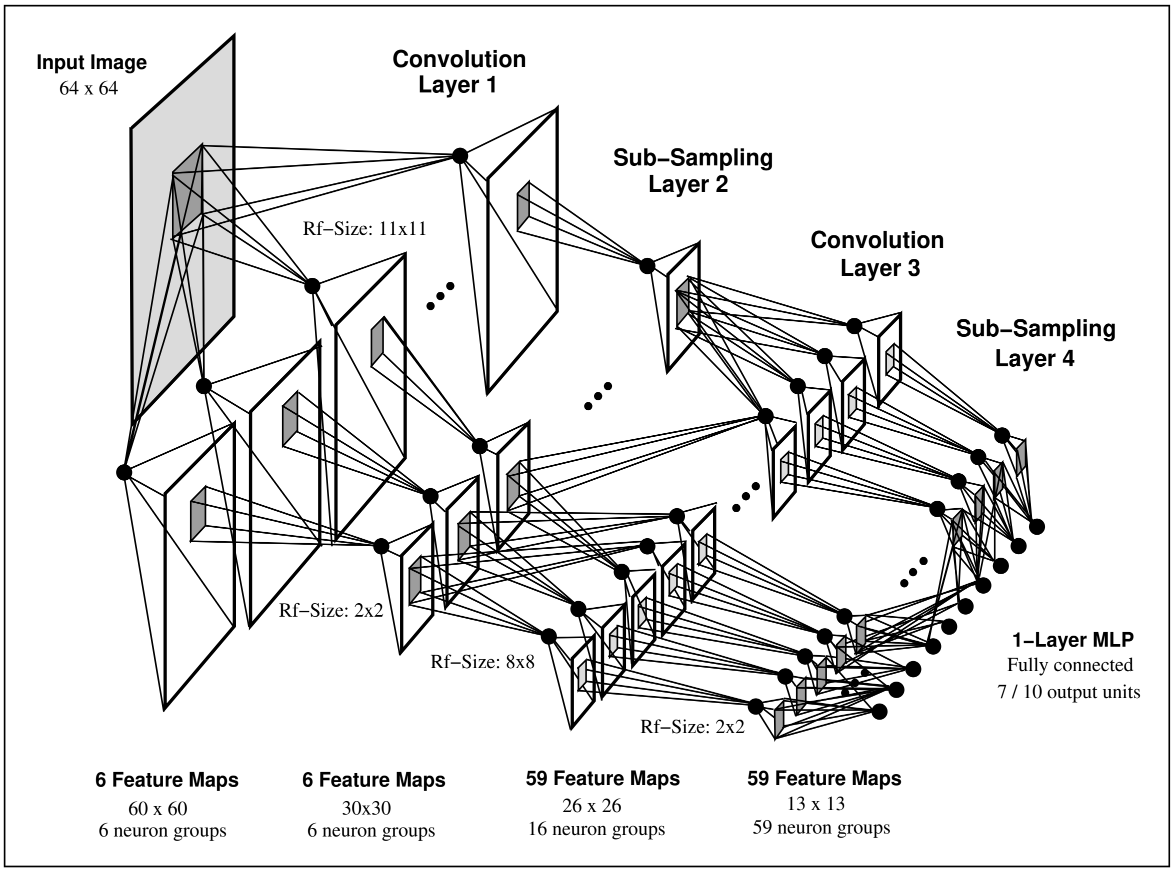

Figure 1: Depicted is the architecture of a 5-layer convolutional neural network (with 2 feature extraction, 2 sub-sampling and one fully connected MLP layer) which we applied for robust face identity and facial expression analysis. Note that the larger dots represent groups of identical neurons.

1 Introduction

Many face analysis approaches require manual intervention during training, such as the construction of face models [11][3][1] or during deployment due to necessary initialization, such as the precise location of facial features, e.g. see [10]. 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 a convolutional neural network (CNN)[6] 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 [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 Face Analysis Systems

Figure 1 shows the architecture of the convolutional neural networks we trained for the task of facial expression recognition and face identity recognition. Its layers alternate between convolution layers

Connection Matrix of Layer 3

$$\text{\# Input maps}\quad\begin{bmatrix} 1&0&0&0&1&1&1&0&0&1&1&1&1&0&1&1\\ 1&1&0&0&0&1&1&1&0&0&1&1&1&1&0&1\\ 1&1&1&0&0&0&1&1&1&0&0&1&0&1&1&1\\ 0&1&1&1&0&0&1&1&1&1&0&0&1&0&1&1\\ 0&0&1&1&0&0&0&1&1&1&1&0&1&1&0&1\\ 0&0&0&1&1&1&0&0&1&1&1&1&0&1&1&1 \end{bmatrix}$$

\# 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:	7
Nb. of connections MLP:	69797
Nb. of weights in MLP:	69797

Nb. of neurons MLP:	10
Nb. of connections MLP:	99710
Nb. of weights in MLP:	99710

Figure 2: Depicted on the left hand side is the interconnection matrix of the third layer of the CNN we trained for facial expression and face identity recognition, whereas on the right hand side are given the number of weights, neurons and neuron inter-connections.

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. 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 [6]. This can also be seen in Figure 2, where the number of CNN neuron-interconnections 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 normalized 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 [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).

Two or more CNNs trained for different face analysis tasks can be combined, so that they may complement each other by delivering context information. This is for example the case with facial expression recognition and face identity recognition. It is an advantage to know the identity of a given subject, when attempting facial expression recognition, as each individual has not only a different facial physiognomy leading to a specific facial action display, but furthermore also performs facial expressions with individual intensities. Therefore, we combined two convolutional neural networks trained for the task of face recognition and facial expression recognition by using a 2-layer MLP with

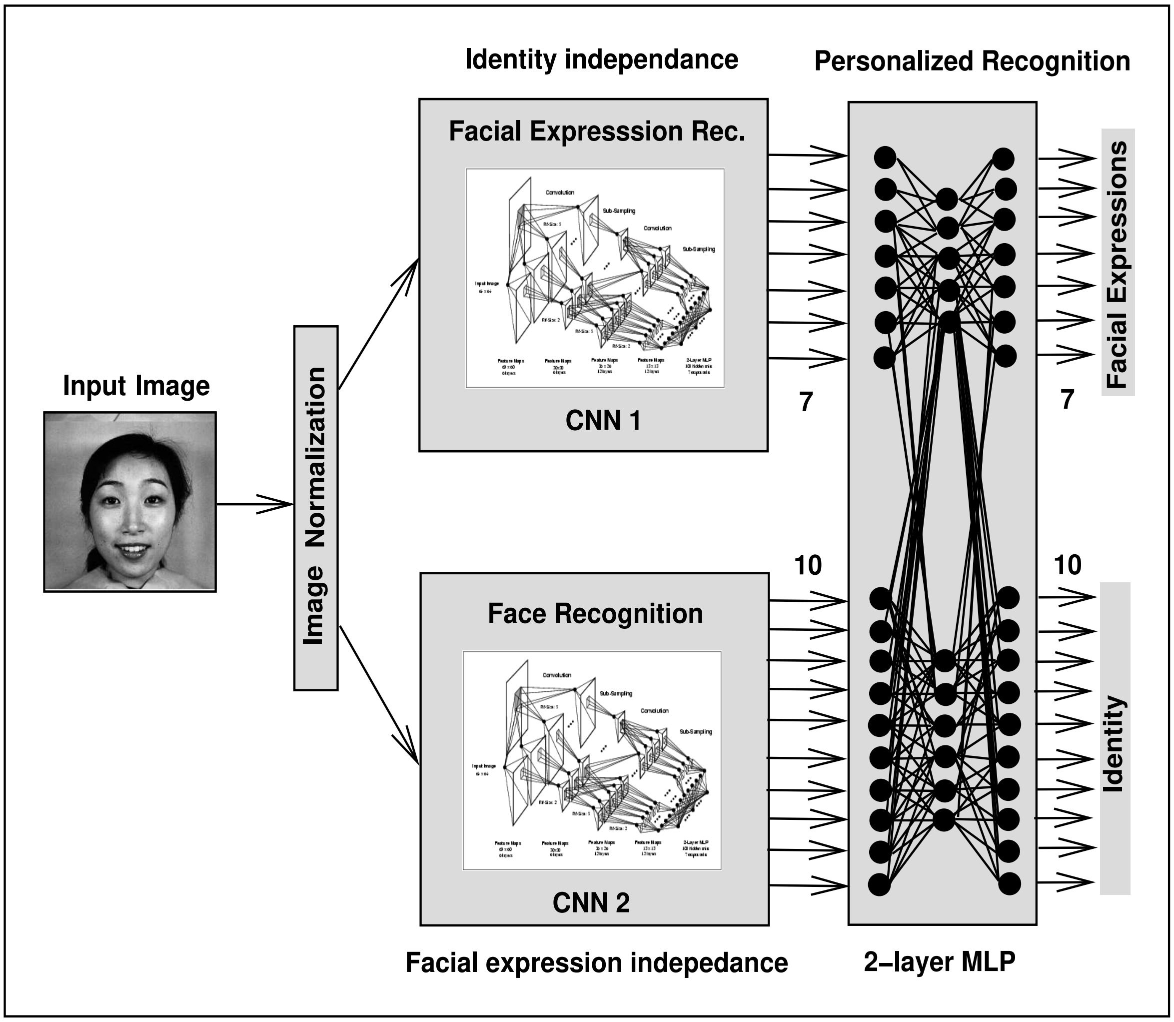

Figure 3: Enhanced facial expression and identity recognition is achieved by using two convolutional neural networks in combination with a 2-layer MLP trained in a standard way or as an auto-associative network.

50 hidden neurons in order to fuse the extracted data for improved and personalized facial expression recognition, see Figure 3. The fusion MLP was trained independently from the CNNs and in two different ways, namely once by training it's input based on the CNNs output in combination with a corresponding target value at its output, once by training it as an auto-associative network by putting at the input the same values as for its target. The MLP-fusion approach allows not only for personalized facial expression analysis, but should also improve face identity recognition of faces deformed by facial actions. However, we couldn't confirm the latter aspect by experiments, because our face database doesn't contain enough subjects, see further below. Note that auto-associative networks have the interesting capability of being able to reconstruct partially degraded inputs of known signals, thus our interest for assessing them.

One of the reasons why we investigated separated facial expression recognition and face identity recognition is, that a previously unseen face leads to a complete failure of the face recognition network. Facial expression recognition on the other hand should work also with previously unseen faces. If face identity and facial expression recognition are achieved in one single convolutional network, the

Figure 4: Sample images of the employed JAFFE facial expression database [8]. Note slight variations of the head position, scale and rotation.

learned feature extraction kernels are coupled, possibly leading to a degraded performance for the task of facial expression recognition with previously unseen faces. Furthermore, separating the task of facial expression recognition and face recognition probably leads to better feature extractors.

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 setups on the JAFFE facial expression database [8], which contains posed emotional facial expression images of 10 Japanese female subjects (6 different emotion displays), see Figure 4.

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). The facial expression images feature slight variations with regard to face scale, face rotations and face translations, which would be difficult to handle by e.g. a simple MLP-based classifier. We used 140 images to train our neural networks and 70 images for testing.

Network Setups	**Recognition Task**	**Corr. Recognition (Exp./Id./Both)**
(1) Single CNN	Exp.	**68.5%**
(2) Single CNN	Id.	98.6%
(3) Single CNN	Exp.+Id.	81.4%/98.6%/80.0%
(4) Fusion 1 (Mlp)	Exp.+Id.	**87.1%**/98.6%/**87.1%**
(5) Fusion 2a (AutoAss)	Exp.+Id.	57.1%/40.0%/82.9%
(6) Fusion 2b (AutoAss)	Exp.+Id.	65.7%/97.1%/40.0%

Table 1: Facial expression and identity recognition results using different network setups. The best facial expression recognition result was obtained with setup 4 by fusing the outputs of a dedicated facial expression and a face identity recognition convolutional neural network.

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 procedures in the CNNs). Instead we trained our convolutional neural networks until a small error was obtained on the training images (which occurred after 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 and identity 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 and 5 correspond to the network architecture depicted in Figure 3. We obtained a significant improvement of the facial expression recognition results, when the convolutional neural networks were not applied on their own, but in combination with a MLP fusion network. The auto-associative network in setup 5 showed good results for the combined recognition of facial expression and face identity. Analyzing the output vector for either facial expressions or face identity resulted in low recognition results. This is due to how this fusion network was trained (featuring the same values at the input and at the output of the network) and also due to the fact that we obtained the classification performance by measuring the Euclidean distance with regard to corresponding training input vectors. However, classification using a threshold (as done in all network setups apart from number 5) leads to completely different recognition results, see network setup 6. Unfortunately, we cannot compare our facial expression recognition results with the ones Lyons and Akamatsu [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. Figure 5 illustrates different feature extraction kernels obtained for the tasks of face recognition and identity recognition as well as their application on a sample face image.

4 Conclusions

Multi-layered convolutional neural networks are of interest for the task of face analysis. We currently focused on facial expression and face identity recognition. However, other face analysis tasks such as age, gender and ethnicity determination may be achieved by using similar approaches. We have demonstrated that it is possible to improve facial expression recognition results by about 20% when using the synergy that stems from processing the output of the facial expression and face recognition networks in a fusion network, where the extracted information is combined. This leads to personalized facial expression recognition and presumably enhanced face identity recognition, the latter especially with high intensity facial expressions. The use of convolutional neural networks lowers the requirement for accurate face pose and scale normalization procedures. As the employed database is rather small, further research has to prove that convolutional neural networks are capable of scaling up with regard to the number of subjects as well as with regard to their capacity of analyzing a greater number of facial expressions.

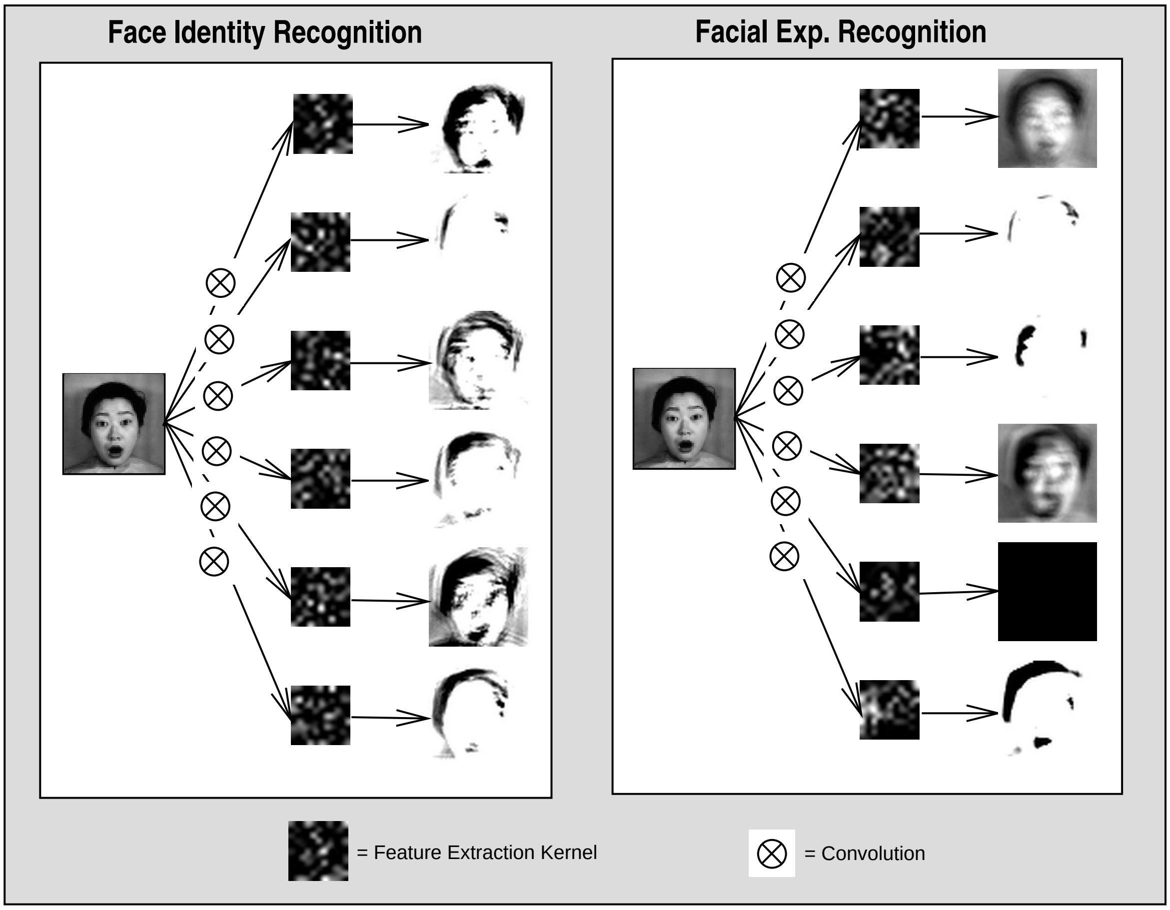

Figure 5: Data-driven feature extraction: Shown are the first layers of two 2-layered CNNs with task specific feature extraction kernels (of size 11×11) for facial expression recognition on the right-hand side and face identity recognition on the left-hand side.

5 Acknowledgements

This work was carried out in the framework of IM(2) under the grant No. 21-54000.98 issued by Swiss National Science Foundation.

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.