

# Neural Architecture for Temporal Emotion Classification

Roland Schweiger, Pierre Bayerl, and Heiko Neumann

Universität Ulm, Neuroinformatik, Germany

{roland.schweiger,pierre.bayerl,heiko.neumann}@informatik.uni-ulm.de

**Abstract.** In this pilot study, a neural architecture for temporal emotion recognition from image sequences is proposed. The investigation aims at the development of key principles in an extendable experimental framework to study human emotions. Features representing temporal facial variations were extracted within a bounding box around the face that is segregated into regions. Within each region, the optical flow is tracked over time. The dense flow field in a region is subsequently integrated whose principal components were estimated as a representative velocity of face motion. For each emotion a Fuzzy ARTMAP neural network was trained by incremental learning to classify the feature vectors resulting from the motion processing stage. Single category nodes corresponding to the expected feature representation code the respective emotion classes. The architecture was tested on the Cohn-Kanade facial expression database.

## 1 Introduction

The automated analysis of human behavior by means of computational vision techniques is a research topic that gained increased attention. Several approaches were proposed. For example, Mase [1] utilized the Facial Action Coding System (FACS) to describe expressions based on the extracted muscle motions. Bascle et al. [2] tracked facial deformations by means of face templates generated from B-spline curves. Key-frames were selected to represent basic face expressions. Most similar to our own approach, Essa and Pentland [3] extracted spatio-temporal energy of facial deformations from image sequences that define dense templates of expected motions. Observed expressions of a human face were classified according to the most similar average motion pattern using a Bayesian classifier.

Unlike previous approaches, we propose a neural network architecture that aims at a framework for emotion recognition based on integrated velocities (amount and direction of motion) in different sectors of a human face. We introduce a simple framework for fast incremental neural network learning to classify different emotions. The architecture is extendable to serve as a tool of experimental investigation. For example, the architecture is flexible to allow the incorporation of features that represent temporal coordination of emotions. In this pilot study, we utilize a supervised principle of incremental allocation of categories to represent different emotions. We evaluate the proposed network using a database of image sequences from facial expressions [4] and demonstrate the discriminative power of the network.

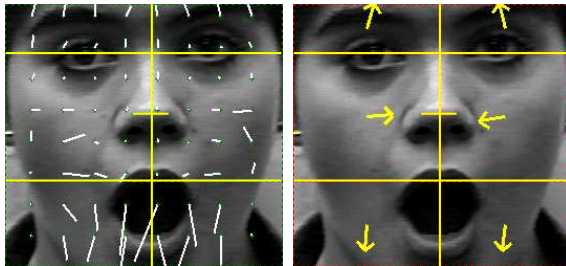
## 2 Framework for Temporal Emotion Classification

### 2.1 Extracting Facial Features

In order to pursue the analysis of facial expressions, we calculate optical flow features using a mechanism of local 2D motion integration proposed by Lucas and Kanade [5]. In order to reduce the dimension of the resulting optical flow data, the first frame in a sequence is labeled manually by drawing a bounding box around the face ranging from the top of the eyebrows down to the bottom of the chin. The bounding box is then subdivided into tiles of possibly varying size. In the first approach, we segregate the bounding box into two halves by a central (almost) vertically oriented symmetry line through the nose region. Horizontal separator lines that were manually adjusted to pass through the pupils and through the mouth further segregate the two halves. This finally leads to six facial regions  $R_k$  (see Fig. 1, left). The second approach divides the bounding box into rectangles of equal size,  $R'_k$ , irrespective of facial components and their orientation.

In order to minimize compute time, the Lucas-Kanade algorithm is first applied on an equidistant grid  $G_0 = \{\mathbf{g}_0^0, \mathbf{g}_1^0, \dots, \mathbf{g}_n^0\}$  in the first image pair of a sequence. This grid is warped by using the calculated flow vectors. The gray levels of new resulting pixel set  $G_1 = \{\mathbf{g}_0^1, \mathbf{g}_1^1, \dots, \mathbf{g}_n^1\}$  is taken as input data for flow estimation in the next image pair. The difference  $\mathbf{u}_j = \mathbf{g}_j^{N-2} - \mathbf{g}_j^0$  represents the optical flow for corresponding grid points of the image pair  $(I_j, I_{j+1})$ . The optical flow in a whole sequence is finally described by  $\mathbf{G} = (G_0, \dots, G_{N-2})$ .

To represent the optical flow of facial deformations, a feature vector  $\mathbf{w}_k$  is calculated in each region  $R_k$ . Flow estimates must be integrated over a sequence of images of variable length. In order to extract flow features invariant to sequence length, a vector of tracked motion is calculated for each grid-point by  $\mathbf{u}_j = \mathbf{g}_j^{N-2} - \mathbf{g}_j^0$ . For all vectors within a face region  $R_k$  (or  $R'_k$ ) we apply a principal component decomposition (PCA) for data reduction. Finally we project  $\sum_{j \in R_k} \mathbf{u}_j$  onto the first principal component which leads to individual feature vectors  $\mathbf{w}_k$  and the feature set  $\mathbf{F}_1 = (\mathbf{w}_1, \dots, \mathbf{w}_6)$  (or,  $\mathbf{F}_2$  for the equi-size tessellation) (see Fig. 1 right).



**Fig. 1. Left:** Tracked flow vectors on an equidistant grid. **Right:** Spatial integration of tracked flow vectors leading to the six feature vectors in feature set  $\mathbf{F}_1$ .

### 2.2 Learning Emotions with Fuzzy ARTMAP

Feature vectors coding different emotions are classified using a Fuzzy ARTMAP neural network architecture [6]. Fuzzy ARTMAP is an architecture for supervised learning

composed of two Fuzzy ART networks that are linked via a map field. In order to train this network, the feature vector in complement coding is presented to the first ART module while the desired output is presented to the second ART module. Learning is performed utilizing a form of hypothesis testing. When the network receives a feature vector, it deduces the best-matching category by evaluating a distance measure against all memory category nodes. Using the second input the network either confirms or rejects the hypothesis, in which case the search process is repeated for a new category node. If the search process failed, new category nodes are dynamically allocated to encode the input.

To test our framework, we used parts of the Cohn-Kanade Facial Expression Database (70 persons, [4]). For each person up to 10 sequences were available, containing 3 up to 25 images. Each sequence represents one of the six basis emotions (*surprise*, *happiness*, *sadness*, *fear*, *anger*, and *disgust*). The data, however, was rather inhomogeneous in that it contained only few sequences for the emotions *anger* or *disgust*. Therefore, results were unreliable and not stable in all cases. For each of the six emotions we trained one Fuzzy ARTMAP in fast learning mode [6]. To get suitable test cases for network performance evaluation, the leave-one-out cross-validation technique [7] was used. Also, a simple perceptron was trained in order to investigate linear separability of the feature vector.

**Table 1.** Error rates for test cases with  $\mathbf{F}_1$  ( $\mathbf{F}_2$ )

emotion	error rate in %	N- seq.	false positive	false negative
happin.	11.9 (11.4)	65	16 (16)	9 (8)
sadness	13.8 (6.1)	35	9 (4)	20 (9)
surprise	3.3 (3.3)	54	5 (6)	2 (1)
anger	13.3 (13.8)	41	18 (19)	10 (10)
fear	7.2 (6.6)	8	7 (6)	8 (8)
disgust	6.1 (5.7)	7	9 (7)	4 (5)

**Table 2.**  $\mathbf{F}_1$ -Confusion matrix

	happin.	sadness	surprise	anger	fear	disgust
happiness	57	0	2	6	4	3
sadness	3	26	4	8	2	0
surprise	2	0	53	0	0	4
anger	4	3	0	31	1	2
fear	5	1	0	2	0	0
disgust	5	0	0	2	0	3

### 3 Results

Unlike the simple perceptron, all of the six trained neural networks were able to represent and recall the training data set without error. This indicates that the emotions were not linear separable. Table 1 demonstrates the error-rates for the test-cases using both feature sets<sup>1</sup>. All data was obtained with a fixed set of network parameters. Although feature set  $\mathbf{F}_1$  was derived from manually adjusted and labeled regions, feature set  $\mathbf{F}_2$  obtains similar results using data from a simple grid of fixed sample width.<sup>2</sup> Table 2 shows the confusion matrix for the six learned emotions (achieved for the low-dimensional  $\mathbf{F}_1$  feature set). The results demonstrate quantitatively the rate of performance along with the between-class confusions. Whether the network behavior reliably

<sup>1</sup> Since we used six emotion classes derived from the original FACS labels, instead of classifying the FACS labels directly (for which our approach has a too low spatial resolution), we did not investigate a quantitative comparison with other approaches.

<sup>2</sup> A closer look at the results demonstrates that *sadness* is even better encoded in  $\mathbf{F}_2$ . We conclude that we may need to increase the number of features in critical regions.

reproduces human confusion needs further investigation of the network performance and processes for feature extraction. As already pointed out above, the results obtained for *fear* and *disgust*, respectively, were unreliable due to the limited training data available.

## 4 Summary and Further Work

In this pilot study, we presented a framework for emotion classification based on supervised neural network learning using Fuzzy ARTMAP. Our approach utilizes quantized optical flow measures that gain robustness through temporal and spatial integration. The feature representation that encodes the velocities of gross facial regions is built incrementally by allocating category nodes of the ART network. The computational architecture provides a testbed of further experimental investigation of processing and analysis of facial emotions.

The system performance can be further increased if more detailed motion features are sampled in regions of higher spatial detail, e.g., around the eyes. This could be achieved by automatic decomposition of regions into smaller parts if the variance of movements in the considered region exceeds threshold. Researchers have argued that the temporal dynamics of the expression, rather than averaged spatial deformations, is important in expression recognition (e.g., [8]). The network can be extended by augmenting spatial features by time-code differentiated features. Instead of using a localized category representation, a distributed self-organizing feature map may increase the robustness of the network. It may further allow to investigate the topographic representation of emotions and to study similarities between emotions based on distance measures in the map.

## References

1. Mase, K.: Human reader: A vision-based man-machine interface. In Cipolla, R., Pentland, A., eds.: *Computer Vision for Human-Machine Interaction*. Cambridge Univ. Press (1998) 53–81
2. Bascle, B., Blake, A., Morris, J.: Towards automated, real-time, facial animation. In Cipolla, R., Pentland, A., eds.: *Computer Vision for Human-Machine Interaction*. Cambridge Univ. Press (1998) 123–133
3. Essa, I.A., Pentland, A.P.: Facial expression recognition using a dynamic model and motion energy. In: *ICCV*. (1995) 360–367
4. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: *Proc 5th IEEE Int. Conf. on Automatic Face & Gesture Recogn.*, France (2000)
5. Lucas, B.D., Kanade, R.: An iterative image registration technique with an application to stereo vision. In: *Proc. 7th Int. J. Conf. on AI*. (1981) 674–679
6. Carpenter, G.A., Grossberg, S., Markuzon, M., Reynolds, J.H., Rosen, D.B.: Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks* **3** (1991) 698–713
7. Weiss, S.M., Kulikowski, C.A.: *Computer Systems that learn, Classification and Prediction methods from Statistics, Neural Nets, Machine Learning and Expert Systems*. Morgan Kaufmann Publishers, San Mateo (1991)
8. Bassili, J.: Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology* **4** (1978) 373–379