Classifier Fusion Applied to Facial Expression Recognition: An Experimental Comparison

Martin Schels, Christian Thiel, Friedhelm Schwenker, and Günther Palm

Abstract. In this paper classifier fusion approaches are investigated through numerical evaluation. For this purpose a multi classifier architecture for the recognition of human facial expressions in image sequences has been constructed on characteristic facial regions and three different feature types (principal components, orientation histograms of static images and temporal features based on optical flow). Classifier fusion is applied to the individual channels established by feature principle and facial region, which are addressed to by individual classifiers. The available combinations of classifier outputs are examined and it is investigated how combining classifiers can lead to more appropriate results. The stability of fusion regarding varying classifier combinations is studied and the fused classifier output is compared to the human view on the data.

1 Introduction and Related Work

Combining classifiers via classifier fusion is a well established technique in pattern recognition. But in spite of theoretical studies it is not yet known how to best construct a combination of classifiers in general, thus there is still a need of empirical evaluation for every particular application.

In [12] static and trainable classifier fusion approaches were evaluated regarding to a multi-modal person verification application. For this, six individual classifiers were constructed on video and audio data, and experiments were conducted in order to investigate the impact on accuracy of a rising number of classifiers contributing to a fused classifier and the effects of fusing classifiers with imbalanced performances. It was affirmed that combining complementary classifiers is more effective than weighting classifiers only by their individual performance. An improvement

Martin Schels · Christian Thiel · Friedhelm Schwenker · Günther Palm
Institute of Neural Information Processing, University of Ulm, 89069 Ulm, Germany
e-mail: {martin.schels,christian.thiel,friedhelm.schwenker, guenther.palm}@uni-ulm.de

H. Ritter et al. (Eds.): Human Centered Robot Systems, COSMOS 6, pp. 121–129.
springerlink.com

of the classification accuracy was achieved using fusion, but there was no relation found between an increasing number of experts and a decrease of the error rate. Furthermore, it was found that static fusion approaches do perform well when the contributing classifiers perform approximately even, and trainable fusion was able to incorporate bigger differences of performance. But this behavior is found to be strongly tied to the quantity and quality of training data.

In [7] several so-called fixed fusion rules were evaluated, applied to a hand written digit recognition task. Four different individual classifiers were trained, each using a different classifier principle. In this study it is was also emphasized, that the single classifiers' outputs should produce independent errors to be suitable for classifier fusion. This could be achieved in this study by using different feature views on the data or utilizing various classifier principles, which make different assumptions about the underlying data. Thus the recognition performance could be improved with respect to the best single classifier in several cases, especially for this application the sum rule performed well, which was more deeply examined.

In [8] another experimental evaluation of classifier fusion can be found. The main scope of this work was the evaluation of Decision Templates, but plenty other approaches were also addressed. The experiments were conducted on the Satimage and the Phoneme data set, which are part of the ELENA collection, and six respectively ten individual classifiers were constructed. There was only little improvement of the combined classifier over the best single classifier and the authors propose that increasing the independence of the different training sets could lead to a greater benefit from the classifier fusion. But an interesting result of their study is that simple static fusion rules do well, but not with every data. And even though they report Decision Templates to be superior in their experiments, they state that there is no fixed approach, which is to be superior at any application.

The detection and recognition of the user's emotional state is an important aspect for the design of more natural interfaces in human computer interaction applications. In face-to-face interaction of humans hand gestures and facial expressions convey nonverbal communication cues and therefore the automatic recognition of such expressions can also play an important role in human computer interaction. In this paper we focus the recognition of facial expressions. There is a rich body of literature available on the detection of emotions in human facial expressions, see [4] for a survey.

The main objective of our study is to investigate and to design ensembles of classifiers and to study the problem of classifier fusion in this application. Several fusion approaches are numerically evaluated and the effects of combining different classifiers with a varying performances are investigated.

2 Classifiers and Classifier Fusion

In the following we give a brief introduction to the SVM and RBF-network classifiers we use, followed by an explanation of the different approaches for classifier fusion. Basically, Support Vector Machines (SVM) are binary classifiers that can be extended to multi-class SVMs (see [14] for an introduction on SVM). In our study SVMs implementing the one-against-one decomposition, which is extended to fuzzy-input fuzzy-output classification are utilized [17].

RBF-Networks are multilayer networks [3, 10], which have distance computing neurons with a radial basis transfer function in the hidden layer. The parameters of RBF networks were trained by gradient descent using the RPROP technique [11].

Combining multiple classifiers can produce a more accurate output than a single one. As mentioned in the introduction an important constraint for an improvement of the recognition rate is that the ensemble of classifiers should produce independent errors. This can be achieved by using different feature views of a dataset or by using different classifier architectures or by utilizing different subsamples of the training data set.

Classifier fusion approaches, that implement a fixed rule without any training procedure are called static. In our experiments minimum, maximum, mean and product rule fusion were evaluated. These fusion rules apply the minimum, maximum, mean or product operator within the particular classes.

Decision Templates and pseudoinverse solution are examples for trainable fusion approaches [15]. Both mappings are using the following form, implementing different transformation matrices V^i, which is incorporating the confusion matrix YC_i^T of the individual classifiers: $z = \sum_{i=1}^{N} V^i C^i(x)^T$. Here z denotes the fused output and $C^i(x)$ is the output of the i-th classifier for data-vector x. The variable Y denotes the desired classifier output. For Decision Templates [15] the matrix is specified as $V^i = (YY^T)^{-1}(YC_i^T)$. The multiplication of $(YY^T)^{-1}$ with the confusion matrix normalizes it with the number of samples of a class. This formulation of Decision Templates is equivalent calculating to the means of the classifiers for each class. Pseudoinverse [15] solution calculates a least-squares linear mapping from the output of the classifiers to the desired output. The matrix V_i is calculated as $V^i = Y \lim_{\alpha \to 0_+} C_i^T (C_i C_i^T + \alpha I)^{-1}$. The limit ensures that the covariance matrix is always invertible.

We also evaluated a brute force fusion approach: Another SVM is constructed as fusion layer, which employs inputs the concatenated outputs of the individual classifiers as input. Do note, that this mapping is not restricted to a linear mapping as the trainable fusion approaches described above.

3 Data Collection

The Cohn-Kanade dataset is a collection of image sequences with emotional content [6], which is available for research purposes. It contains image sequences, which were recorded in a resolution of 640×480 (sometimes 490) pixels with a temporal resolution of 33 frames per second. Every sequence is played by an amateur actor who is filmed from a frontal view. The sequences always start with a neutral facial

Table 1 Confusion matrix of the human test persons against the majority of all 15 votes (left). The right column shows the share of the facial expressions in the data set (hardened class labels).

maj.\test pers.	hap.	ang.	sur.	disg.	sad.	fear	no. samples
hap.	**0.99**	0	0	0	0	0.01	105
ang.	0	**0.8**	0	0.12	0.07	0.01	49
sur.	0.01	0	**0.78**	0	0.01	0.19	91
disg.	0.01	0.15	0.01	**0.67**	0.01	0.15	81
sad.	0	0.08	0.02	0.02	**0.88**	0.01	81
fear	0.01	0.01	0.14	0.27	0.01	**0.56**	25

expression and end with the full blown emotion which is one of the six categories “fear”, “happiness”, “sadness”, “disgust”, “surprise” or “anger”.

To acquire a suitable label the sequences were presented to 15 human labelers (13 male and two female). The sequences were presented as a video. After the playback of a video the last image remained on the screen and the test person was asked to select a label. Thus, a fuzzy label for every sequence was created as the mean of the 15 different opinions. The result of the labeling procedure is given in Tab. 1, showing the confusion matrix of the test persons according to the majority of all persons. It is revealed that the data collection is highly imbalanced only 25 samples expression “fear” occur in the data set and in addition, this expression could not be identified by the test persons.

In all automatic facial expression recognition systems first some relevant features are extracted from the facial image and these feature vectors then utilized to train some type of classifier to recognize the facial expression. One problem is here how to categorize the emotions: one way is to model emotions through a finite set of emotional classes such as anger, joy, sadness, etc, another way is to model emotions by a continuous scales, such as valence (the pleasantness of the emotion) and arousal (the level of activity) of an expression [9]. In this paper we use a discrete representation in six emotions. Finding the most relevant features the definitely the most important step in designing a recognition systems. In our approach prominent facial regions such as the eyes, including the eyebrows, the mouth and for comparison the full facial region have been considered. For these four regions orientation histograms, principal components, optical flow features have been computed. Principal components (eigenfaces approach) are very well know in face recognition [18], and orientation histograms were successfully applied for the recognition of hand gestures [5] and faces [16], both on single images. In order to extract the facial motion in these regions optical flow[1] features from pairs of consecutive images have been computed, as suggested in [13].

[1] We were using a biologically inspired optical flow estimator, which was developed by the Vision and Perception Science Lab of the Institute of Neural Processing at the University of Ulm [2, 1] .

4 Experiments and Results

The classification experiments are all executed in the following set-up: For every distinct combination of feature and region a SVM or a RBF-network operating on a single frame, i.e. feature of an image or pair of successive images is trained. The classifiers are optimized using 8-fold cross validation and all the following results are also 8-fold cross validated. A list of classifiers and the individual performances for each of these channels can be found in Tab. 2, evidently our classifiers show rather imbalanced recognition rates concerned to the emotional categories. The frame-wise results are temporally integrated by taking the average over the decisions [2]. These results are then fused into a global result using one of the approaches presented in Sect. 2.

We evaluate the whole combinatory space, which is given by the 14 available classifiers, for every fusion approach to study its properties on different feature combinations. The best results of the fusion steps are noted in Tab. 3, which shows also an entry called "oracle" [8]. This is the amount of samples, that is correctly classified by any classifier. It is of course not one of the proposed approaches for classifier fusion, but it should give a hint to the potential of the ensemble. With a value of 98.3 % this measure is quite promising in our case. Comparing the fused results with the best single classifier shows that nearly every fusion approach leads to an improvement of the recognition rate. Only the usage of maximum fusion has a

Table 2 Overview of the individual classifiers ordered by recognition rate. These 14 single classifiers are the candidates for our classifier fusion experiments.

No.	Feature	Region	Classifier	Rec.-Rate (%)
1	Orientation histograms	mouth	SVM	74.0
2	Orientation histograms	mouth	RBF	70.3
3	Optical flow	face	RBF	67.1
4	Optical flow	mouth	RBF	67.1
5	PCA	mouth	RBF	65.9
6	PCA	face	RBF	62.7
7	Orientation histograms	face	SVM	54.3
8	Orientation histograms	face	RBF	52.3
9	Optical flow	right eye	RBF	51.1
10	Optical flow	left eye	RBF	45.8
11	Orientation histograms	right eye	RBF	42.3
12	Orientation histograms	left eye	RBF	41.4
13	PCA	left eye	RBF	37.9
14	PCA	right eye	RBF	35.1

[2] It was also considered to to use Hidden Markov Models for this step, which did not turn out to be feasible.

Table 3 Fusion results of experiments with static and trainable fusion. Also listed are the best single machine classifier (compare Tab. 2) and the average performance of the human labelers against the mean of all labelers and the "oracle" [8].

Fusion approach	Rec.-rate (%)
Product rule	83.1
Mean	81.7
Minimum	77.8
Maximum	73.8
Brute force (SVM)	83.6
Pseudoinverse	81.4
Decision-Templates	75.9
Best single classifier	74.0
Average human labeler	81.5
"Oracle"	98.3

negative effect. Applying brute force as fusion method results in a recognition rate of 83.6 %, which is the highest rate in this study. This even slightly outperforms the average human test person (81.5 %), as determined in the labeling experiments. Many static fusion approaches, namely mean fusion and product rule, result in high recognition rates. Only one of the classical trainable fusion approaches, which are mainly linear mappings, does only result an acceptable recognition rate: pseudoinverse fusion is able to reach up to 81.4 % correctly classified samples, while decision tables can only slightly improve over the best individual classifier.

To examine the performance on varying combinations, we are sorting the results by the percentage of correctly classified samples and classifier combination within the fusion approaches (see Fig. 1). It can be observed that the pseudoinverse approach outperforms the best individual classifier more often than probabilistic product and brute force, which are fusion approaches with higher recognition rates in Tab. 3. So the training of the pseudoinverse mapping does result in more robust results and it should be easier to pick a classifier combination for an application. This observation sustains for this case the claim, that trainable fusion could incorporate weaker classifiers better [12] because of the training procedure. On the other hand the brute force fusion approach does result in the highest over-all recognition rate, but proves to be even more unstable than the simple product rule.

In Tab. 4 (bottom) the confusion matrix of the champion architecture using the product rule for the classifiers 2, 4, 6, 9 and 10 is displayed. The fused classifier reveals a quite similar recognition performance as the human test persons (see Tab. 1), referring to the particular classes. All the classes except "surprise" and "fear" show an almost identical performance on the diagonal line of the confusion matrices in both cases. For the machine classifier the task seems to imply similar difficulties as for the test persons. It is obvious that our classifier performs only weak for the class "fear", similarities can be observed: In both cases this class is confused with classes "surprise" and "disgust" more often than others.

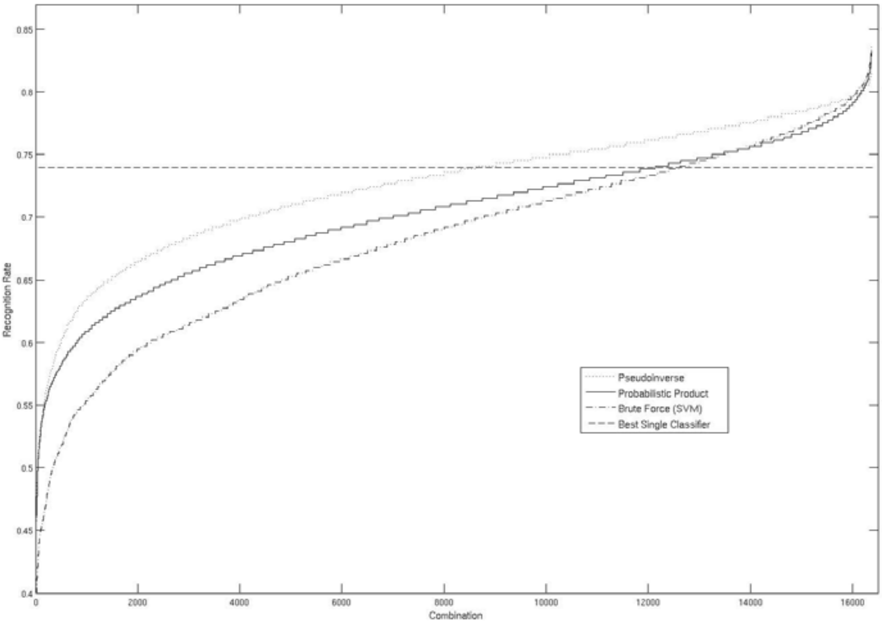

Fig. 1 Sorted recognition rates of three selected fusion approaches. Pseudoinverse fusion does not produce the highest recognition rates of the evaluated approaches, but this mapping does outperform the best individual classifier more often.

Table 4 Top: Confusion matrix of two classifiers (left: classifier 4; right: classifier 6), which are contributing to the best classifier architecture (bottom) using probabilistic product (built out of classifiers 2, 4, 6, 9 and 10). The entries of the matrices are fractions.

true\classif.	hap.	ang.	sur.	disg.	sad.	fear
happiness	**0.87**	0.1	0.1	0.8	0.3	0.1
anger	0.2	**0.59**	0.2	0.4	0.29	0.0
surprise	0.1	0.0	**0.89**	0.4	0.2	0.3
disgust	0.25	0.20	0.7	**0.32**	0.10	0.6
sadness	0.7	0.20	0.1	0.5	**0.67**	0.0
fear	0.28	0.4	0.28	0.20	0.8	**0.12**

true\classif.	hap.	ang.	sur.	disg.	sad.	fear
happiness	**0.79**	0.2	0.4	0.8	0.4	0.4
anger	0.12	**0.53**	0.0	0.12	0.22	0.0
surprise	0.9	0.1	**0.82**	0.3	0.3	0.1
disgust	0.20	0.19	0.2	**0.37**	0.16	0.6
sadness	0.10	0.15	0.5	0.5	**0.64**	0.1
fear	0.12	0.0	0.20	0.28	0.16	**0.24**

true\classif.	hap.	ang.	sur.	disg.	sad.	fear
happiness	**0.95**	0.01	0	0.04	0	0
anger	0	**0.80**	0	0.10	0.10	0
surprise	0.01	0	**0.97**	0	0.01	0.01
disgust	0.17	0.12	0.01	**0.63**	0.01	0.05
sadness	0.01	0.05	0.04	0.01	**0.89**	0
fear	0.08	0	0.20	0.32	0.04	**0.36**

Table 4 shows also two confusion matrices of classifiers 4 and 6, which are part of the ensemble forming the third confusion matrix. Now we can exemplarily investigate the possible impact of the fusion procedure: Classifier 4 is found to be superior to classifier 6 in class “happiness” and “surprise”, but inferior in classes “disgust” and especially “fear”. In the final fused classifier these results are merged and in all cases there is a further improvement in all classes.

5 Summary and Conclusion

In this paper we studied various multiple classifier fusion approaches applied to the classification of human facial expressions. To this end we trained individual classifiers for characteristic facial segments (left and right eye, mouth and full face) and three feature types. Trainable and static fusion approaches were examined concerning the incorporation of classifiers with different performances. The linear transformation fusion with pseudoinverse shows at this application greater stability with respect to the combination of individual classifiers and does also result in high recognition rates. Brute force and the static fusion rules did provide less stability even though some of these techniques do deliver a high top recognition rate.

To motivate classifier fusion, the behavior of two individual classifiers was exemplarily inspected by analyzing the confusion matrices. The classifiers supplemented each other by producing different errors and thus the fused classifier is able to reach a better performance. By comparing the confusion matrix of a fused ensemble to the confusion matrix of the human test persons a remarkable analogies were observed.

Acknowledgements. The authors would like to thank Prof. Dr. H. Neumann, Dr. P. Bayerl and S. Ringbauer from the Vision and Perception Science Lab of the Institute of Neural Processing at the University of Ulm for generous assistance in extracting the optical flow features.

This paper is based on work done within the project SCHW623/4-3, and the “Information Fusion” subproject of the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems”, both funded by the German Research Foundation (DFG). The work of Martin Schels is supported by a scholarship of the Carl-Zeiss Foundation.

References

1. Bayerl, P., Neumann, H.: Disambiguating visual motion through contextual feedback modulation. Neural Comput. 16, 2041–2066 (2004)
2. Bayerl, P., Neumann, H.: A fast biologically inspired algorithm for recurrent motion estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 246–260 (2007)
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)

4. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18, 32–80 (2001)
5. Freeman, W.T., Roth, M.: Orientation Histograms for Hand Gesture Recognition. In: International Workshop on Automatic Face and Gesture Recognition, pp. 296–301 (1994)
6. Kanade, T., Cohn, J., Tian, Y.L.: Comprehensive database for facial expression analysis. In: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2000), pp. 46–53 (2000)
7. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20, 226–239 (1998)
8. Kuncheva, L., Bezdek, J.C., Duin, R.P.W.: Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition 34, 299–314 (2001)
9. Lang, P.J.: The emotion probe. studies of motivation and attention. The American Psychologist 50, 372–385 (1995)
10. Poggio, T., Girosi, F.: A theory of networks for approximation and learning. Laboratory, Massachusetts Institute of Technology 1140 (1989)
11. Riedmiller, M., Braun, H.: RPROP – description and implementation details. Technical report, Universität Karlsruhe (1994)
12. Roli, F., Kittler, J., Fumera, G., Muntoni, D.: An experimental comparison of classifier fusion rules for multimodal personal identity verification systems. In: Roli, F., Kittler, J. (eds.) MCS 2002. LNCS, vol. 2364, pp. 325–336. Springer, Heidelberg (2002)
13. Rosenblum, M., Yacoob, Y., Davis, L.: Human expression recognition from motion using a radial basis function network architecture. IEEE Transactions on Neural Networks 7, 1121–1138 (1996)
14. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning). The MIT Press, Cambridge (2001)
15. Schwenker, F., Dietrich, C., Thiel, C., Palm, G.: Learning of decision fusion mappings for pattern recognition. International Journal on Artificial Intelligence and Machine Learning (AIML) 6, 17–21 (2006)
16. Schwenker, F., Sachs, A., Palm, G., Kestler, H.A.: Orientation Histograms for Face Recognition. In: Schwenker, F., Marinai, S. (eds.) ANNPR 2006. LNCS (LNAI), vol. 4087, pp. 253–259. Springer, Heidelberg (2006)
17. Thiel, C., Giacco, F., Schwenker, F., Palm, G.: Comparison of Neural Classification Algorithms Applied to Land Cover Mapping. In: Proceedings of the 18th Italian Workshop on Neural Networks, WIRN 2008. IOS Press, Amsterdam (2008)
18. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)