Multimodal Affect Classification using Deep Neural Networks

Friedhelm Schwenker

Ulm University
Neural Information Processesing
D-89069 Ulm
friedhelm.schwenker@uni-ulm.de

Abstract: Research activities in human-computer interaction increasingly addressed the aspect of integrating emotional intelligence into the overall system, and therefore the recognition of human emotions becomes important in such applications. Human emotions are expressed through various kinds of modalities such as voice, facial expressions, hand/body gestures or bio-physiological patterns, and therefore the classification of human emotions should be considered as a multimodal pattern recognition and machine learning problem. In this paper we propose artificial neural networks for the task of information fusion in multimodal affect classification.

1 Introduction

The success of today's digital computers is mainly based on their exceptional performance in numerical computation and symbol processing tasks. On the other hand rule-based systems fail in many real world applications, for instance in solving complex perceptional tasks such as machine vision, speech recognition, or computer-aided medical diagnosis, and although humans are outperformed by digital computers in terms of speed and accuracy to large extent, humans easily can solve these perceptional tasks. All these applications have in common that the relevant information of the underlying process is hidden and too complex to specify a computer program explicitly, e.g. the recognition of a spoken utterance, or the detection of a person's face in a complex visual scene, or the classification of the person's affective state.

Machine learning systems offer an alternative approach to tackle such problems. Here the idea is to extract the relevant (but hidden) knowledge from a sample of data. This approach is called data driven or data based, in contrast to the theory driven or software driven approach [1]. The learning approach has been successfully applied to a variety of real world applications, e.g. typical pattern recognition tasks such as image and video classification, speech recognition, speaker identification, affect recognition [2, 1],

In affect recognition in human-computer interaction (HCI) scenarios one has to deal with many different types of data, typically the data comes in a continuous stream, e.g. video streams, audio streams or sequences of biopotentials, making the analysis and interpretation of this data too difficult for a rule based approach Data must be pre-processed and the most relevant features must be extracted either by using per-defined feature extraction algorithms or feature learning approaches. This typically gives a sequence of feature vectors that are used as the input to the classifier models. For affect recognition, as for many other pattern recognition tasks the temporal structure of this time series is essential, and must be modeled using hidden Markov models (HMM) or recurrent neural networks [3].

Figure 1 – Screen shot of the ATLAS annotation and analysis tool. The most important windows are Line-Tracks, Video Windows, Main Control Menu, Label Details, Tabular Label View and Legend Window.

Collecting raw data of the entire task is just the first step of any machine learning or pattern recognition task. For the design of the classifiers, the quality of the data annotation with ground truth labels is crucial. Annotation of data is expensive and error-prone, in particular if the data is multimodal and shows a complex temporal structure [4]. Figure 1 shows a screen-shot of such a piece of HCI data. Because of the complexity of the data powerful specialized software tools are needed to process such data streams. For this, we developed the ATLAS system that is tailored to the analysis and annotation of such data [5]. Besides the display of the raw data streams, the ATLAS tool is able to visualize streams of extracted features, classifier results (crisp and probabilities) and label information in synchronized form.

Although labeling the data might be extremely time consuming and difficult, labeled data is needed to design and evaluate the classifiers. Labeling data from HCI is complex because the categories of interest may not be well defined in all cases, for instance the classification of the human's emotional state is extremely complex because the ground truth of the data is difficult to find. Another source of complexity is the multimodal nature of the data where different modalities potentially need different annotation schemes, and therefore labeling of HCI data is usually conducted by a team of independent raters.

2 Multimodal emotion recognition

Research activities in the field of modern HCI increasingly address the aspect of integrating some type of emotional intelligence, and therefore the recognition of human emotions becomes important. Human emotions are expressed through various kinds of modalities such as voice, facial expressions, gestures or biophysiological patterns, and therefore the classification of human emotions is related to modeling and recognizing multimodal time series affective computing [6, 7, 8, 9, 10].

One of the most important and pioneering researcher, who has dealt with the understanding of emotions, was Charles Darwin [11]. By examining the interaction between facial muscles and the associated emotions, he introduced the first rules for the recognition of human emotions. Ekman developed the Facial Action Coding System (FACS) [12] to encode the emotions by muscle activity patterns and to distinguish between the six basic emotions: *joy*, *anger*, *surprise*, *disgust*, *sadness*, and *fear*.

The aim of our work in [4] was to take into account the temporal structure in facial expressions by using Hidden Markov models (HMM) for the recognition, and because of their ability to model sequences in a probabilistic way, HMM are used in speech recognition [3] and more recently also for emotion recognition [13]. In the approach we followed in [4], four regions of interest were selected within the images, the entire face region, mouth region, and regions of right and left eye. Per region three different features were computed, principal component analysis, orientation histograms and optical flow estimation, were extracted. The resulting twelve paired combinations of feature and region were used to evaluate the HMM. Numerous experiments to optimally adjust the HMMs were conducted. The optimal number of states and the optimal number of the normal distributions of the Gaussian Mixture Models (GMM), which were attached to the states, were determined empirically. Additionally, two different model architectures were evaluated. To improve the over-all performance further, two approaches to combine the results of the ensemble classifiers were applied. A surprising result of our numerical experiments on the Cohn-Kanade data set was that the human perception capability was similar to the performance of the automatic classifiers. In an experiment on speech based emotion recognition conducted on the Berlin-Emotional-Database a similar result was found [14].

Emotion classification from audio data is also challenging task. Here characteristic features from the audio signal are extracted, such as fundamental frequency, linear predictive coding (LPC), mel-frequency cepstral coefficients (MFCC), modulation spectrum and given as a sequence of feature vectors to a classifier [14]. In [15] we investigated fuzzy-input fuzzy output SVM to classify voice quality samples from a vowel corpus. In the numerical evaluation the fuzzy Support Vector Machines (SVM) show low error rates in comparison to standard classifiers, such as binary SVM and naive Bayes, using cross validation and leave-one-speaker-out testing. This promising result is supporting the idea of dealing with fuzzy labels during learning as well as fuzzy outputs of classifiers in a MCS architecture.

Additionally to speech, face or gesture, human emotions obviously consist of internal physiological processes, and therefore measuring physiological parameters, such as skin conductivity, heart rate, respiration, muscle activity (for instance of particular facial muscles utilizing electromyography (EMG)), or brain activity (for instance by utilizing electroencephalography (EEG)) is an important step to study emotional states . Therefore, in the following the focus is on recognition of emotion using these different modalities in naturalistic HCI scenarios, for instance in Wizard-of-Oz settings [16, 4, 17, 18]. In naturalistic HCI scenarios the emotional utterances are mainly *neutral*, hence only a few (and often of low intensity) emotional patterns can be observed in such data. This leads to weak classifier decisions in single modalities and therefore fusing or combining of the classifier outputs from different modalities is a promising approach to improve the system's overall performance. In [16] we investigated the capabilities of MultipleClassifier Systems (MCS) (see, [19, 20, 21, 22, 23, 24]) to classify emotional states from a data set collected in a Wizard-of-Oz scenario. The numerical evaluation of audio-based classifiers and classifiers trained on biopotentials shows that multiple classifier systems using fixed and trainable fusion mappings applied to multimodal emotional data can outperform the unimodal classifiers. This is an expected result, but even in unimodal applications the overall recognition performance often increases by using MCS architectures trained on different feature views. For instance, in our work on EEG classification where we especially study the detection of P300 events [25] it was found that MCS architectures outperform monolithic classifiers (here Random forest are used to achieve baseline results), also in our work on audio-visual emotion recognition MCS architectures have been successfully applied [26, 27, 28].

We investigated MCS methods on common benchmark data sets, for instance data sets provided within the *Audio-Visual Emotion Recognition Challenge* [29], and could demonstrate excellent performances [26, 28], in 2014 our proposed system won the AVEC 2014 challenge. In [27] we investigated audio-visual detection (online and offline setting) of laughter in natural

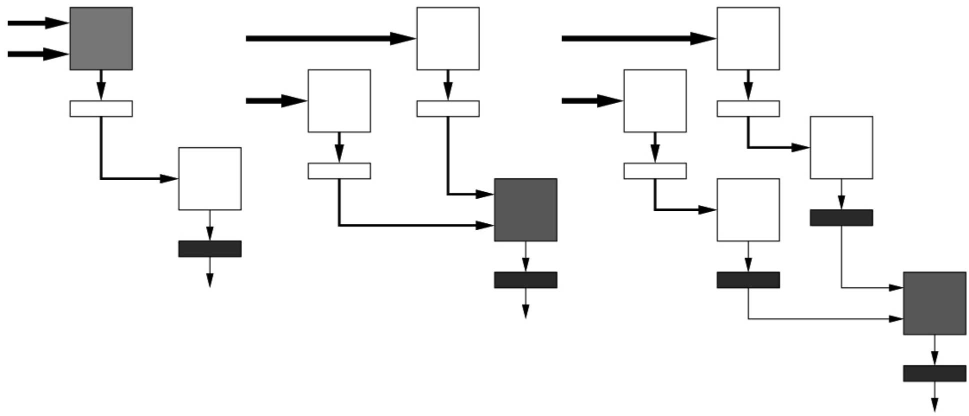

Figure 2 – Three different fusion schemes realized by layered deep neural network architectures: Early fusion (left panel), two different late fusion schemes (middle and right panel).

conversation. In this work the idea was that a human centered user interfaces should be able to infer the user's state and/or the state of the communication between the user and the computer. For this, the system must be able to recognize (and interpret) the user's verbal and nonverbal signals. In our study we investigated laughters, because laughs are believed to be important of lively conversation. In the experimental set up the conversations are not restricted by any constraints and data recording is done by centrally placed microphone and 360° camera system. This audio-visual data set was used for the evaluation for different classifiers (recurrent Echo-State-Networks, SVM with Gaussian Mixture Model (GMM) supervectors [30, 31], and HMM). The best performances (93.7% accuracy) was achieved by fusing audio and video modalities using HMM architectures, the best unimodal classification results on speech yield around 90% accuracy.

3 Conclusion

Multimodal recognition of user affects is a challenging issue which could be faced by the combination of various modalities. Besides the combination of suitable modalities and features, effort should be applied to a stringent development and handling of recognition architectures. Achieved results are based on publicly available benchmark data sets, remarkable results were obtained in the AVEC benchmark challenges. Besides the classification issue, the preprocessing of data is an important issue. For this, ATLAS is briefly discussed that allows a synchronous processing of various different data streams and offers special active learning techniques to assist the entire labelling process.

References

[1] SHAWE-TAYLOR, J. and N. CRISTIANINI: *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

[2] JAIN, A. K., J. MAO, and K. M. MOHIUDDIN: *Artificial neural networks: A tutorial*. *Computer*, 29(3), pp. 31–44, 1996. doi:10.1109/2.485891. URL `http://dx.doi.org/10.1109/2.485891`.

[3] RABINER, L. and B.-H. JUANG: *Fundamentals of Speech Recognition*. Prentice Hall PTR, 1993.

[4] SCHWENKER, F., S. SCHERER, M. SCHMIDT, M. SCHELS, and M. GLODEK: *Multiple classifier systems for the recognition of human emotions*. In N. E. GAYAR, J. KITTLER, and F. ROLI (eds.), *Proceedings of the 9th International Workshop on Multiple Classifier Systems (MCS'10)*, LNCS 5997, pp. 315–324. Springer, 2010. URL `http://www.springerlink.com/content/p41t12t345075758/`.

[5] MEUDT, S., L. BIGALKE, and F. SCHWENKER: *Atlas-annotation tool using partially supervised learning and multi-view co-learning in human-computer-interaction scenarios*. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, pp. 1309–1312. IEEE, 2012.

[6] SCHWENKER, F., S. SCHERER, M. SCHMIDT, M. SCHELS, and M. GLODEK: *Multiple classifier systems for the recognition of human emotions*. In N. EL GAYAR, J. KITTLER, and F. ROLI (eds.), *Proceedings of the 9th International Workshop on Multiple Classifier Systems (MCS'10)*, LNCS 5997, pp. 315–324. Springer, 2010.

[7] SCHELS, M., P. SCHILLINGER, and F. SCHWENKER: *Training of multiple classifier systems utilizing partially labeled sequential data sets*. In *ESANN 2011, 19th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 27-29, 2011, Proceedings*. 2011.

[8] GLODEK, M., S. TSCHECHNE, G. LAYHER, M. SCHELS, T. BROSCH, S. SCHERER, M. KÄCHELE, M. SCHMIDT, H. NEUMANN, G. PALM, and F. SCHWENKER: *Multiple classifier systems for the classification of audio-visual emotional states*. In S. K. D'MELLO, A. C. GRAESSER, B. W. SCHULLER, and J. MARTIN (eds.), *Affective Computing and Intelligent Interaction - Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9-12, 2011, Proceedings, Part II*, vol. 6975 of *Lecture Notes in Computer Science*, pp. 359–368. Springer, 2011. doi:10.1007/978-3-642-24571-8_47.

[9] GLODEK, M., F. HONOLD, T. GEIER, G. KRELL, F. NOTHDURFT, S. REUTER, F. SCHÜSSEL, T. HÖRNLE, K. C. J. DIETMAYER, W. MINKER, S. BIUNDO, M. WEBER, G. PALM, and F. SCHWENKER: *Fusion paradigms in cognitive technical systems for human-computer interaction*. *Neurocomputing*, 161, pp. 17–37, 2015. doi:10.1016/j.neucom.2015.01.076.

[10] MEUDT, S., D. ZHARKOV, M. KÄCHELE, and F. SCHWENKER: *Multi classifier systems and forward backward feature selection algorithms to classify emotional coloured speech*. In J. EPPS, F. CHEN, S. OVIATT, K. MASE, A. SEARS, K. JOKINEN, and B. W. SCHULLER (eds.), *2013 International Conference on Multimodal Interaction, ICMI '13, Sydney, NSW, Australia, December 9-13, 2013*, pp. 551–556. ACM, 2013. doi:10.1145/2522848.2531743.

[11] DARWIN, C.: *The Expression of the Emotions in Man and Animals*. Oxford University Press Inc, New York, 1 edn., 1872.

[12] EKMAN, P. and W. V. FRIESEN: *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[13] YEASIN, M., B. BULLOT, and R. SHARMA: *From facial expression to level of interest: A spatio-temporal approach*. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2, pp. 922–927, 2004.

[14] ESPARZA, J., S. SCHERER, A. BRECHMANN, and F. SCHWENKER: *Automatic emotion classification vs. human perception: Comparing machine performance to the human benchmark*. In *Proceedings of the 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA'12)*, pp. 1286–1291. IEEE, 2012.

[15] SCHERER, S., J. KANE, C. GOBL, and F. SCHWENKER: *Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification*. *Journal on Computer Speech and Language*, 27(1), pp. 263–287, 2013.

[16] WALTER, S., S. SCHERER, M. SCHELS, M. GLODEK, D. HRABAL, M. SCHMIDT, R. BÖCK, K. LIMBRECHT, H. C. TRAUE, and F. SCHWENKER: *Multimodal emotion classification in naturalistic user behavior*. In J. A. JACKO (ed.), *Proceedings of the 14th International Conference on Human Computer Interaction (HCI'11)*, LNCS 6763, pp. 603–611. Springer, 2011. URL http://www.springerlink.com/content/606237v0u5225w50/.

[17] SCHERER, S., M. GLODEK, M. SCHELS, M. SCHMIDT, G. LAYHER, F. SCHWENKER, H. NEUMANN, and G. PALM: *A generic framework for the inference of user states in human computer interaction: How patterns of low level communicational cues support complex affective states*. *Journal on Multimodal User Interfaces: Special Issue on Conceptual Frameworks for Multimodal Social Signal Processing*, pp. 117–141, 2012.

[18] TRAUE, H., F. OHL, A. BRECHMANN, F. SCHWENKER, H. KESSLER, K. LIMBRECHT, H. HOFFMAN, S. SCHERER, M. KOZYBA, A. SCHECK, and S. WALTER: *Framework for emotions and dispositions in man-computer interaction*. In M. ROJC and N. CAMPBELL (eds.), *Coverbal Synchrony in Human-Machine Interaction*, Artificial Neural Networks in Pattern Recognition, pp. 99–140. CRC Press, 1 edn., 2013.

[19] KITTLER, J. and F. ROLI (eds.): *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000, Proceedings*, vol. 1857 of *Lecture Notes in Computer Science*. Springer, 2000. doi:10.1007/3-540-45014-9.

[20] SCHWENKER, F., F. ROLI, and J. KITTLER (eds.): *Multiple Classifier Systems - 12th International Workshop, MCS 2015, Günzburg, Germany, June 29 - July 1, 2015, Proceedings*, vol. 9132 of *Lecture Notes in Computer Science*. Springer, 2015. doi:10.1007/978-3-319-20248-8.

[21] DIETRICH, C., F. SCHWENKER, and G. PALM: *Classification of time series utilizing temporal and decision fusion*. In J. KITTLER and F. ROLI (eds.), *Multiple Classifier Systems, Second International Workshop, MCS 2001 Cambridge, UK, July 2-4, 2001, Proceedings*, vol. 2096 of *Lecture Notes in Computer Science*, pp. 378–387. Springer, 2001. doi:10.1007/3-540-48219-9_38.

[22] DIETRICH, C., F. SCHWENKER, and G. PALM: *Multiple classifier systems for the recognition of orthoptera songs*. In B. MICHAELIS and G. KRELL (eds.), *Pattern Recognition, 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003, Proceedings*, vol. 2781 of *Lecture Notes in Computer Science*, pp. 474–481. Springer, 2003. doi:10.1007/978-3-540-45243-0_61.

[23] THIEL, C., F. SCHWENKER, and G. PALM: *Using dempster-shafer theory in MCF systems to reject samples*. In N. C. OZA, R. POLIKAR, J. KITTLER, and F. ROLI (eds.), *Multiple Classifier Systems, 6th International Workshop, MCS 2005, Seaside, CA, USA,*

June 13-15, 2005, Proceedings, vol. 3541 of *Lecture Notes in Computer Science*, pp. 118–127. Springer, 2005. doi:10.1007/11494683_12.

[24] ABDEL HADY, M. F. and F. SCHWENKER: *Decision templates based rbf network for tree-structured multiple classifier fusion*. In J. A. BENEDIKTSSON, J. KITTLER, and F. ROLI (eds.), *Proceedings of the 8th International Workshop on Multiple Classifier Systems (MCS'09)*, LNCS 5519, pp. 92–101. Springer, Reykjavik, Iceland, 2009.

[25] SCHELS, M., S. SCHERER, M. GLODEK, H. A. KESTLER, G. PALM, and F. SCHWENKER: *On the discovery of events in EEG data utilizing information fusion*. *Computational Statistics*, 28(1), pp. 5–18, 2013. doi:10.1007/s00180-011-0292-y.

[26] GLODEK, M., S. TSCHECHNE, G. LAYHER, M. SCHELS, T. BROSCH, S. SCHERER, M. KÄCHELE, M. SCHMIDT, H. NEUMANN, G. PALM, and F. SCHWENKER: *Multiple classifier systems for the classification of audio-visual emotional states*. In S. K. D'MELLO, A. C. GRAESSER, B. SCHULLER, and J.-C. MARTIN (eds.), *Affective Computing and Intelligent Interaction - Fourth International Conference, ACII 2011, Part II*, vol. 6975 of *Lecture Notes in Computer Science*, pp. 359–368. Springer, 2011.

[27] SCHERER, S., M. GLODEK, F. SCHWENKER, N. CAMPBELL, and G. PALM: *Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data*. *ACM Transactions on Interactive Intelligent Systems: Special Issue on Affective Interaction in Natural Environments*, 2(1), pp. 4:1–4:31, 2012. URL `http://dl.acm.org/citation.cfm?id=2133370`.

[28] KÄCHELE, M., M. SCHELS, and F. SCHWENKER: *Inferring depression and affect from application dependent meta knowledge*. In [29], pp. 41–48. doi:10.1145/2661806.2661813. URL `http://doi.acm.org/10.1145/2661806.2661813`.

[29] VALSTAR, M., B. W. SCHULLER, J. KRAJEWSKI, R. COWIE, and M. PANTIC (eds.): *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14, Orlando, Florida, USA, November 7, 2014*. ACM, 2014. doi:10.1145/2661806. URL `http://doi.acm.org/10.1145/2661806`.

[30] SCHWENKER, F., S. SCHERER, Y. MAGDI, and G. PALM: *The gmm-svm supervector approach for the recognition of the emotional status from speech*. In C. ALIPPI, M. POLYCARPOU, C. PANAYIOTOU, and G. ELLINAS (eds.), *Proc. of the 19th International Conference on Artificial Neural Networks (ICANN'09) - Part I*, LNCS 5768, pp. 894–903. Springer, 2009.

[31] SCHELS, M. and F. SCHWENKER: *A multiple classifier system approach for facial expressions in image sequences utilizing gmm supervectors*. In A. ERCIL (ed.), *Proc. of the 20th International Conference on Pattern Recognition (ICPR'10)*, pp. 4251–4254. IEEE, 2010. doi:10.1109/ICPR.2010.1033.