21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Multi-Modal Fusion Based on Classifiers Using Reject Options and Markov Fusion Networks

Michael Glodek, Martin Schels, Günther Palm, and Friedhelm Schwenker
Institute of Neural Information Processing
Ulm University, Germany
firstname.lastname@uni-ulm.de

Abstract

Classifying continuous signals from multiple channels poses several challenges: different sample rates from different types of channels have to be incorporated. Furthermore, when leaping from the laboratory to the real world, it is mandatory to deal with failing sensors and also uncertain or even incorrect classifications. We propose a new Multi Classifier System (MCS) based on the application of classifier making use of an reject option and a Markov Fusion Network (MFN) which is evaluated in an off-line and on-line manner. The architecture is tested using the publicly available AVEC corpus, that collects affectively labeled episodes of human computer interaction. The MCS achieved a significant improvement compared to the results obtained on the single modalities.

1. Introduction

In order to reliably classify using multiple data streams a classifier architecture must meet various requirements: different sample rates from different types of channels have to be combined, sensors or classification results may not be available at every distinct time step and, furthermore, the classification results for the single channels can be uncertain or sometimes even incorrect. On the other hand, a combination of the different modalities bears the chance to compensate failures of sensors and enhance the over-all classification This paper proposes a Multi Classifier System (MCS) that handles uncertain and unreliable classifications from multiple sources. The architecture is evaluated in a realistic human computer interaction (HCI) scenario which is related to emotion recognition. Emotion recognition in real-world scenarios is a vivid and challenging field of research [11, 13]. Affective states have to be recognized from multiple sources (e.g. audio, video or even bio-physiological channels) such that it is available for applications within an acceptable lag of time.

The remainder of this paper is organized as follows: Section 2 describes the methods. In Section 3 the experiments and results are described. In the last section, we conclude and give an outlook to future work.

2. Proposed Architecture

The suggested architecture is based on two main concepts, namely classifiers using reject option and the Markov Fusion Network (MFN), which will be presented in this section.

2.1. Classification with Uncertainty Measures

The proposed architecture is based on a set of independent classifiers for each modality which are capable of returning a fuzzy class membership along with a corresponding confidence of the decision, both ranging in the interval of $[0, 1]$. The final classification is obtained using model averaging, while the confidence is obtained based on the standard deviation of the decisions (i.e. the more diverse the memberships of the classifiers, the less confident the averaged decision).

Statistical learning theory in general aims at minimizing the misclassification rate or the expected loss. However, in practical application classifiers may reject samples without rendering any decision, e.g. in case of uncertain classifiers. This approach is called "classification with reject option" [5, 2]. Typically, the classification errors arise in the region of the decision boundary in which all posterior-probabilities are low. Standard approaches are based on thresholds determined by heuristics on probabilistic classifier outputs or on

the agreement of classifiers as implement in the given MCS [8].

2.2. Markov Fusion Network

As a result of the reject option the fusion has to deal with a sparse vector of decisions coming from different sources. The MFN has been designed to combine these decisions from multiple sources with temporal dependencies and has strong relations to the application of Markov Random Fields in image processing [2, 6]. The value y_t is defined as the estimated decision obtained by the combination of streams of different sources. The streams are given by $\mathbf{x}_m$ where $m \in M$ is the index of a source. In Figure 1, a graphical model of a MFN combining two sources is shown. According to the graphical model the estimates y_t are connected in a chain. Whenever a classifier of a stream provides a decision x_{tm}, a link to the estimate y_t is added to the graph.

The MFN is defined by two potentials Ψ and Φ. The potentials Ψ_m of modality $m = 1, \ldots, M$ encourage the outcomes y_t to be equal to the decisions x_{tm} and is given by

$$\Psi = \sum_{m}^{M} \Psi_m = \sum_{m}^{M} \sum_{t \in \mathcal{L}_m} k_m (x_{tm} - y_t)^2,$$

where $\mathcal{L}_m$ is the set of time steps in which a decision for m is available, and, k_m is a parameter defining the strength of its influence. The second potential Φ enforces lateral similarity:

$$\Phi = \sum_{t=1}^{T} \sum_{i \in N(t)} w_{t+i} (y_t - y_{t-(1-2i)})^2,$$

where $\mathbf{w} \in \mathbb{R}^{T-1}$ weights the cost of a difference between two adjacent nodes. The set $N(t)$ contains 0 and 1 in case y_t has a neighbor node y_{t-1} or y_{t+1}, respectively. The parameter $\mathbf{w}$ can be set using domain knowledge at design time, e.g. to weaken similarity in case an extraordinary event is induced by the computer.

The joint distribution of the estimated vector $\mathbf{y}$ given the decisions of the modalities is defined by

$$p(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_M) = \frac{1}{Z} \exp\big(-\frac{1}{2}(\Psi + \Phi)\big).$$

Since we are not interested in the probability itself, but in minimizing the mode of the log-posterior probability, the variable $\mathbf{y}$ can be determined using gradient descent. The MFN is capable of combining the decisions from multiple sources of a complete recording at once. However, we additionally implemented an on-line classification by shifting a window over the data stream such that the combined estimate can be provided close to real-time.

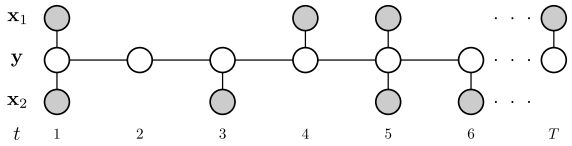

Figure 1. Graphical model of the Markov Fusion Network. The estimates y_t are influenced by the the available decisions x_{tm} of the source m and $t \in \mathcal{L}_m$ and the adjacent estimates y_{t-1} and y_{t+1}.

3. Experiments

To evaluate the proposed MCS, the AVEC corpus 2011 has been chosen as the reference data set[1]. The Audio/Visual Emotion Challenge (AVEC) 2011 was first introduced in the context of the ACII 2011 workshop and come with audio-visual recordings of episodes of affective HCI [12, 10]. The data was recorded in a HCI scenario in which the subjects were instructed to interact with an affectively colored artificial agent. The data was collected in over-all 63 recordings from 13 different subjects and labeled in four affective dimensions: *arousal*, *expectancy*, *power* and *valence*. The ranking of the submitted results of AVEC 2011 was performed only on the dimension of arousal since the other dimensions yielded poor results.

The audio classification was performed on three bags of features:

- *Fundamental frequency*, the *energy* and *linear predictive coding* (LPC)
- *Mel frequency cepstral coefficient* (MFCC)
- *Relative Spectral Transform - Perceptual Linear Prediction* (RASTA-PLP) [7].

To obtain a fixed-length feature vector based on an arbitrary long sequence of features, a transformation using HMM was implemented, as suggested in [1]. The classification was conducted using five random forests [4] and the final output is made by averaging. The corresponding standard deviation is used to calculate the confidence measure.

Classification from the video channel was performed on features obtained from the computer expression recognition toolbox (CERT) [9], which is designed to recognize facial properties (such as action units or basic emotions). The output of the modules "Basic Emotions 4.4.3", "FACS 4.4", "Unilaterals" and "Smile Detector" were concatenated to form a 36-dimensional fea-

[1] Details can be found at `sspnet.eu/avec2011/` (28/3/2012).

ture vector per frame. The classification and the confidence value is obtained analogously to the audio classification using five naive Bayes classifiers with bagging [3]. It is worth noting that the detection of the subjects face failed in about 8% of the frames.

The audio and video classification is combined using the MFN in an off-line and on-line mode. Off-line reefers to the MFN using the complete recording at once, while on-line means to shifting a window over the data stream. Within each window the MFN is iterated till convergences and the estimated result of a certain position in the window is returned as the final decision.

3.1. Results

To obtain valid results, we set up a subject independent four-fold cross-validation based on the training and development set of the AVEC 2011 corpus. Each fold is limited to a set of persons, such that no training on test subjects is performed. The results of the word-wise classification and the recognition based on the facial expression with no rejection, 10% rejection and 90% rejection are given in Table 1 in form of the accuracy, the F_1-measure[2] and $\overline{F}_1$ which is based on the complement of the target class (all values in average and standard deviation in brackets based on the test set). The video channel benefits most of the rejection (i.e. arousal, expectancy and valance), while for audio the rejection does only show significant improvement in the arousal dimension.

Table 2 shows the results of the MFN using off-line and on-line processing. For all experiments the w_t value have been set to 64 within a turn and 32 between the turns. Three configurations of the parameter $\mathbf{k}$ are optimized using the development set of the cross-validation: $(0.5\ 0.5)$, $(0.3\ 0.5)$ and $(0.5\ 0.3)$, where the first value weights the audio, and the second the video channel. The arousal dimension shows the best improvement compared to the single classifier results. Since, the rejection of audio results is not required and the parameter $\mathbf{k}$ is set to $(0.5\ 0.3)$ a strong influence by the audio stream is present. The recognition of the expectancy dimension makes most use of the rejection. Although, the F_1 measure of the video channel performs better, the average of F_1 and $\overline{F}_1$ is better for the fused result. The parameter $\mathbf{k}$ is set to $(0.3\ 0.5)$ and, therefore, put an emphasis to the video channel. The power dimension shows only minor improvement of the accuracy. Although the audio rejection rate is set high, $\mathbf{k}$ is set to $(0.5\ 0.3)$. The dimension of the valance shows a clear improvement compared to the complete video channel without rejection. The decrease of the $\overline{F}_1$ is due to the audio recognition characteristics. Dispite most of the audio result have been rejected, again the parameter $\mathbf{k}$ is set to $(0.5\ 0.3)$. The off-line and on-line results are almost identical which gives evidence to draw the conclusion that the proposed approach is applicable in a real-world scenarios without loose of performance. Figure 2 illustrates the estimation of the combined decision over time using the on-line procedure. The orange and light blue dots represent the decisions of the video

[2]Defined by the harmonic mean of the recall and the precision.

Table 1. Word-wise and frame-wise results of the audio and video modality.

(a) No rejection

Audio	Arousal	Expectancy	Power	Valance
↑Acc.	61.8 (3.6)	58.9 (6.3)	57.5 (9.4)	57.5 (7.9)
↑F_1	65.8 (3.8)	16.4 (7.1)	69.6 (9.3)	70.1 (6.8)
↑$\overline{F}_1$	56.7 (3.4)	72.6 (5.2)	24.7 (6.6)	24.9 (8.4)
Video	Arousal	Expectancy	Power	Valance
↑Acc.	57.0 (4.3)	54.7 (4.0)	55.7 (2.8)	59.9 (7.4)
↑F_1	60.9 (5.1)	49.6 (9.4)	57.4 (11.3)	67.1 (11.5)
↑$\overline{F}_1$	51.3 (9.3)	56.6 (10.7)	48.7 (12.2)	43.5 (7.1)

(b) 10% rejection rate

Audio	Arousal	Expectancy	Power	Valance
↑Acc.	62.0 (3.5)	58.7 (6.3)	57.5 (9.4)	57.5 (7.6)
↑F_1	66.1 (3.8)	15.8 (7.3)	69.6 (9.3)	70.1 (6.6)
↑$\overline{F}_1$	56.9 (3.3)	72.5 (5.2)	24.5 (6.2)	24.4 (8.8)
Video	Arousal	Expectancy	Power	Valance
↑Acc.	57.7 (4.3)	55.0 (3.9)	55.7 (2.8)	60.4 (7.8)
↑F_1	61.5 (5.2)	50.3 (9.5)	57.2 (11.3)	67.6 (11.9)
↑$\overline{F}_1$	51.8 (9.3)	56.6 (10.6)	49.1 (12.2)	43.5 (7.5)

(c) 90% rejection rate

Audio	Arousal	Expectancy	Power	Valance
↑Acc.	62.7 (3.1)	59.1 (6.6)	59.1 (9.9)	57.5 (7.6)
↑F_1	67.3 (2.8)	15.9 (8.8)	71.1 (9.4)	70.1 (6.6)
↑$\overline{F}_1$	56.4 (4.3)	72.7 (5.6)	24.2 (5.3)	24.4 (8.8)
Video	Arousal	Expectancy	Power	Valance
↑Acc.	59.5 (4.3)	56.9 (1.6)	54.4 (4.1)	67.8 (10.4)
↑F_1	65.0 (5.1)	56.0 (13.5)	54.9 (10.3)	75.6 (11.8)
↑$\overline{F}_1$	50.6 (9.3)	52.5 (14.0)	50.7 (10.0)	41.0 (12.7)

Table 2. Off-line and on-line MFN results.

Offline	Arousal	Expectancy	Power	Valance
↑Acc.	65.7 (4.0)	59.5 (3.9)	57.3 (4.8)	61.2 (8.0)
↑F_1	69.9 (2.9)	48.2 (11.3)	60.3 (11.8)	69.6 (10.5)
↑$\overline{F}_1$	59.7 (8.3)	65.2 (7.8)	47.0 (14.6)	41.6 (4.7)
Online	Arousal	Expectancy	Power	Valance
↑Acc.	65.8 (2.8)	59.6 (4.4)	57.2 (4.5)	61.0 (7.8)
↑F_1	69.2 (2.5)	48.9 (10.1)	60.2 (11.5)	69.2 (10.4)
↑$\overline{F}_1$	61.0 (6.8)	65.3 (8.1)	47.4 (14.2)	42.3 (4.8)
A. rej.	0%	90%	90%	90%
V. rej.	90%	90%	10%	10%

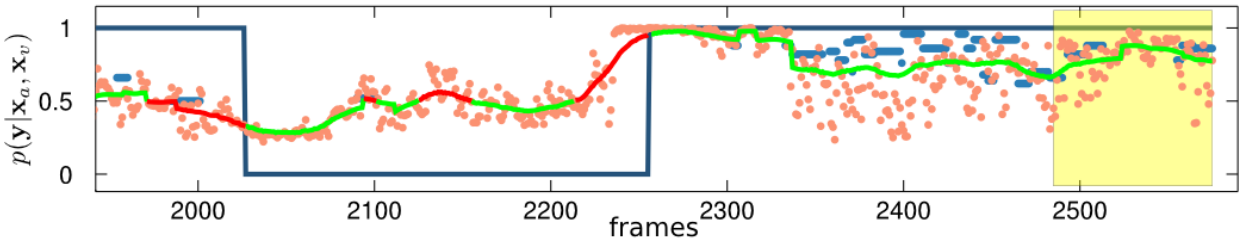

Figure 2. Estimated decisions (continuous green and red curve) of audio (light blue dots) and video decision (orange dots) using the on-line MFN. More details are to be found in the text.

and audio channel. The dark blue curve is the binary ground-truth given by the AVEC dataset. The continuous green and red curves correspond to the estimated combined result (green correct/red wrong). The window size is set to five seconds and the estimated result of the fourth second is returned. The example shows how the MFN compensates the missing audio decisions in moments of silence. In the case of a conversational turn, the estimated decision is allowed to have a stronger change with respect to the neighboring decisions.

4. Conclusions

We presented a novel multi-modal fusion architecture based on classifiers with rejection option. According to the design, the remaining decisions are integrated using a Markov Fusion Network (MFN). The approach has been successfully tested utilizing the AVEC 2011 corpus [12, 10]. Sensor failure and rejected decisions are reconstructed by the MFN. Furthermore, the MFN was studied in an on-line manner by shifting a window over the test sequence. In future work, we want to realize a system to detect similar emotions operating in real-time and making use of the proposed approach.

Acknowledgment

This paper is based on work done within the Transregional Collaborative Research Centre SFB/TRR 62 Companion-Technology for Cognitive Technical Systems funded by the German Research Foundation (DFG).

References

[1] M. Bicego, V. Murino, and M. Figueiredo. Similarity-based clustering of sequences using hidden Markov models. In *Proceedings of the International conference on Machine Learning and Data Mining (MLDM)*, volume 2734 of *Lecture Notes in Computer Science (LNCS)*, pages 95–104. Springer, 2003.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41 – 46, 1970.

[6] J. Diebel and S. Thrun. An application of Markov random fields to range sensing. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, volume 18, pages 291–298. MIT Press, 2006.

[7] X. Huang, A. Acero, H. Hon, et al. *Spoken language processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 2001.

[8] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2004.

[9] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (CERT). In *Proceedings of the International Conference on Automatic Face & Gesture Recognition and Workshops*, pages 298–305. IEEE, 2011.

[10] G. McKeown, M. Valstar, R. Cowie, and M. Pantic. The SEMAINE corpus of emotionally coloured character interactions. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, pages 1079–1084. IEEE, 2010.

[11] S. Scherer, 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*, 2(1):4:1–4:31, 2012.

[12] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. AVEC 2011 — the first international audio visual emotion challenges. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, volume 6975 of *Lecture Notes in Computer Science (LNCS)*, pages 415–424, 2011. Part II.

[13] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signal processing: State-of-the-art and future perspectives of an emerging domain. In *Proceedings of the International Conference on Multimedia (MM)*, pages 1061–1070. ACM, 2008.