ENSEMBLE APPROACHES TO PARAMETRIC DECISION FUSION FOR BIMODAL EMOTION RECOGNITION

Jonghwa Kim and Florian Lingenfelser
Institute of Computer Science, University of Augsburg, Universitätsstr. 6a, 86159 Augsburg, Germany

Keywords: Emotion recognition, Biosignal, Speech, Decision fusion, Multisensory data fusion, Pattern recognition, Affective computing, Human-computer interface.

Abstract: In this paper, we present a novel multi-ensemble technique for decision fusion of bimodal information. Exploiting the dichotomic property of 2D emotion model, various ensembles are built from given bimodal dataset containing multichannel physiological measures and speech. Through synergistic combination of the ensembles we investigated parametric schemes of decision-level fusion. Up to 18% of improved recognition accuracies are achieved compared to the results from unimodal classification.

1 INTRODUCTION

Recently a burgeoning interest in automatic emotion recognition from different modalities such as speech, facial expression, and physiological signals has been prompted in affective human-computer interface (Zeng et al., 2009; Kim and André, 2008). Also multimodal approaches by exploiting synergistic combination of multiple modalities are reported for improved recognition accuracy (Chen et al., 1998; Bailenson et al., 2008; Kim and André, 2006). Particularly for such approach we first need to design suitable fusion method for multichannel sensory data.

Commonly the fusion can be performed at least at three levels; data, feature, and decision level. If observations are of same type, the data-level fusion might be probably the most appropriate way where we simply combine raw multisensory data. Feature-level fusion should be efficient for *multichannel* data that are measured from similar sensor type, synchronized time, and unique signal dimension. Usually a single classifier (expert) is employed for the combined feature vectors to make decision. For *multimodal* sensory data containing discriminative data from different modalities, decision-level fusion might be the most convincing way. In the decision fusion, multiple experts that use different classifiers trained by same data or same type of classifier trained by different data are generated to derive favorable final decision. Many methods for this type of fusion have been reported in various names, such as multiple classifier systems, mixture of experts, or ensemble systems (Polikar, 2006).

However, most of machine learning algorithms are generalized method based on statistics or linear regression of given data and most suitable for binary classification problems. Therefore they might not be able to capture characteristics of input variables in order to efficiently solve multiclass problems. Same problem holds true for ensemble approaches because those results are strongly depending on characteristics of input variables.

In this paper, we present a novel ensemble approach to parametric decision fusion for automatic emotion recognition from two modalities, biosignals and speech. The results are compared with the recognition accuracies that we presented in our previous work (Kim and André, 2006) using feature-level fusion and generalized decision-level fusion.

2 BIMODAL DATASET

In this work we used the same dataset and feature vectors presented in our previous work (Kim and André, 2006) to which we refer for details of the dataset and feature vectors. The dataset contains the four emotions that represent each quadrant of the 2D emotion model (Kim and André, 2008), i.e., HP (high arousal / positive valence), HN (high arousal / negative valence), LN (low arousal / negative valence), and LP (low arousal / positive valence). Dur-

Kim J. and Lingenfelser F. (2010).
ENSEMBLE APPROACHES TO PARAMETRIC DECISION FUSION FOR BIMODAL EMOTION RECOGNITION.
In *Proceedings of the Third International Conference on Bio-inspired Systems and Signal Processing*, pages 460-463
DOI: 10.5220/0002753204600463

ing the quiz experiment five channels of biosignals are recorded, blood volume pulse (BVP), Respiration rate (RSP), skin conductivity (SC), electromyogram (EMG), and body temperature (TEMP). For the ensemble approach in this paper, each of the biosignals forms a single physiological channel and all channels are summed up to generate a complete BIO channel. Also speech data recorded during the experiment are segmented according to measured time periods of biosignals and stored as SPE channel.

Feature sets consist of 77 features from the five channel BIO data by analyzing in time, frequency, and statistic domain and 61 MFCC (Mel-frequency cepstral coefficients) features including common statistic values from SPE data.

3 BUILDING ENSEMBLES

3.1 Basic Bimodal Ensemble

After feature selection through the sequential backward search algorithm (Jain and Zongker, 1997), the feature sets (BIO and SPE) are separately classified for the four emotion classes. We used the pLDA[1](pseudoinverse linear discriminant analysis (Kim and André, 2008)). Table 1 shows all results of unimodal classification. The classifiers trained by each modality represent individual experts that can be used to build ensembles for decision fusion. The basic idea of decision level fusion is to reduce the total error rate of classification by strategically combining the members of the ensemble and their errors. Therefore the performance of the single classifiers needs to be diverse from one another, i.e., neither must these classifiers provide perfect performance on some given problem, nor do their outputs need to resemble each other.

3.2 Cascading Specialists Approach

Using generalized decision-level fusion methods such as majority voting and Borda count that repetitively apply weighted decisions causes in general problem with extremely unbalanced overall performance because of overemphasizing some classes by weighting. To overcome this problem, we developed a novel algorithm, we called as cascading specialists (CS) method that chooses experts for single classes and brings them in a special sequence. Figure 1 illustrates this approach.

[1]In this work, we used this single classifier for all channels and ensemble decisions

Table 1: Basic multichannel ensemble (available channels and individual classification results).

Subject A					
Channel	*HP*	*HN*	*LN*	*LP*	*avg*
BIO	86.36	70.83	61.90	74.07	**73.29**
SPE	77.27	58.33	76.19	66.67	69.62
Subject B					
Channel	*HP*	*HN*	*LN*	*LP*	*avg*
BIO	55.56	62.50	67.65	44.83	57.64
SPE	72.22	62.50	79.41	79.31	**73.36**
Subject C					
Channel	*HP*	*HN*	*LN*	*LP*	*avg*
BIO	52.17	65.52	66.00	61.90	61.40
SPE	60.87	72.41	70.00	78.57	**70.46**
Subject Independent					
Channel	*HP*	*HN*	*LN*	*LP*	*avg*
BIO	44.26	43.04	51.43	59.18	49.48
SPE	32.79	58.23	71.43	54.08	**54.13**

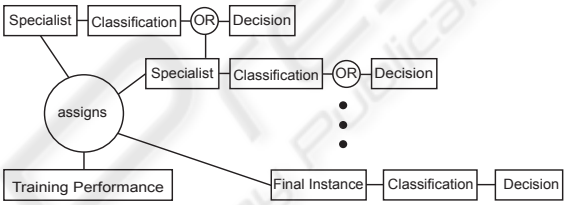

Figure 1: Cascading Specialists.

First, the experts are selected by finding the classifier with best true positive rating for every class of the classification problem during training phase. Then the classes are rank ordered, beginning with the worst classified class across all classifiers and ending with the best one. After the preparation, the algorithm works as follows: the first class in the sequence is chosen and the corresponding expert is asked to classify the sample. If the output matches the currently observed class the output is chosen as ensemble decision. If not, the sample is passed on to the next weaker class and corresponding expert whilst repeating the strategy. It is often observed that none of the experts classifies its connected class and the sample remains unclassified at the end of the sequence. Then the classifier with the best overall performance on the training data is selected as final instance and is asked to label the sample as ensemble decision.

This approach promises more uniformly distributed classification results and a more accurate overall performance than most ensemble methods that rely on experts because weakly recognized classes are treated with priority and the belonging samples are more unlikely to end up falsely classified as a more dominating class later on.

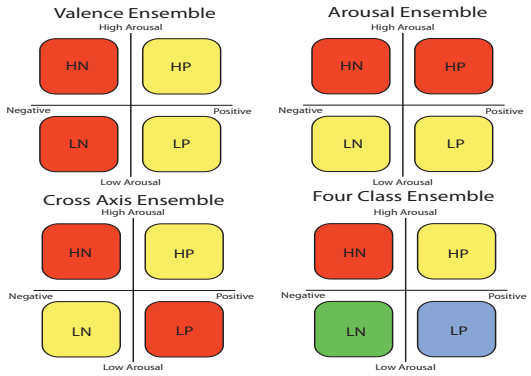

Figure 2: Considered emotion-specific dichotomous ensembles and four class ensemble.

Table 2: Classification performance of the dichotomous ensembles.

Subject A			**Subject B**		
Arousal			**Arousal**		
high	*low*	*avg*	*high*	*low*	*avg*
89.13	91.67	**90.40**	85.71	96.83	**91.27**
Valence			**Valence**		
positive	*negative*	*avg*	*positive*	*negative*	*avg*
91.84	97.67	**94.81**	82.98	93.10	**88.04**
Cross Axis			**Cross Axis**		
HP+LN	*HN+LP*	*avg*	*HP+LN*	*HN+LP*	*avg*
95.35	90.20	**92.78**	82.69	88.68	**85.69**
Subject C			**Subject Independent**		
Arousal			**Arousal**		
high	*low*	*avg*	*high*	*low*	*avg*
86.54	93.48	**90.01**	63.57	86.21	**74.89**
Valence			**Valence**		
positive	*negative*	*avg*	*positive*	*negative*	*avg*
92.31	63.29	**77.80**	70.44	76.09	**73.27**
Cross Axis			**Cross Axis**		
HP+LN	*HN+LP*	*avg*	*HP+LN*	*HN+LP*	*avg*
76.71	85.92	**81.32**	71.08	71.19	**71.14**

3.3 Dichotomous Ensembles

Using the CS algorithm, we considered three dichotomous ensembles (arousal, valence, and cross axis) and four class ensemble, based on the axes of the 2D emotion model (see Figure 2). Table 2 and 3 show the ensembles and their classification performance.

4 PARAMETRIC DECISION FUSION

Each of the all ensembles (Table 2 and 3) generates its decision by using the CS algorithm and the provided votes are all given a numerical value of one and then take part in a stepwise combination process positively leading to a final decision. Classification is guaranteed, as the final step inevitably leads to a result (if preceding steps could not establish it due to voting ties).

Table 3: Classification performance of four class ensemble.

Subject A					**Subject B**				
HP	*HN*	*LN*	*LP*	*avg*	*HP*	*HN*	*LN*	*LP*	*avg*
81.82	70.83	61.90	85.19	**74.94**	72.22	66.67	82.35	72.41	**73.41**
Subject C					**Subject Independent**				
HP	*HN*	*LN*	*LP*	*avg*	*HP*	*HN*	*LN*	*LP*	*avg*
60.87	65.52	70.00	78.57	**68.74**	45.90	53.16	56.19	59.18	**53.61**

Step 1: Combination of Arousal, Valence and Cross Axis. This step exactly matches the static combination method presented in dichotomous approach with cross axis. Each ensemble distributes its votes among the two quadrants that fit the recognised alignments in the 2D emotion model. This step results in one of two possible outcomes:

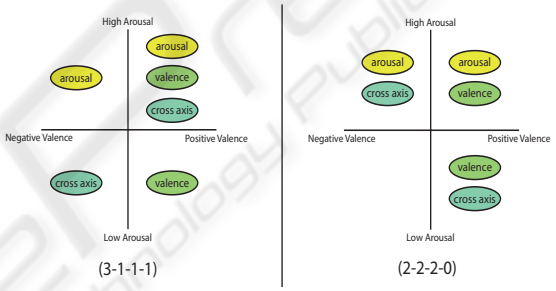

Figure 3: Possible vote distributions in the step 1.

(3-1-1-1) If the ensembles agree on one emotion-quadrant, it receives three votes and can already be chosen as final decision.

(2-2-2-0) If the ensembles do not manage to agree on one emotion-quadrant, a voting tie occurs. No final decision can be chosen, instead the draw has to be dissolved and the algorithm moves on to the next step.

Step 2: Resolving of Draws through Direct Tendencies. In order to resolve the draw, the direct classification ensemble designates exactly one vote to the class it predicts. Two situations can arise through this supplemental vote:

(3-2-2-0) If the ensemble chooses an emotion-quadrant that already holds two votes, the tie is resolved and the corresponding emotion is determined to be the final decision.

(2-2-2-1) If the ensemble chooses the emotion-quadrant that has not received any votes yet, the tie is not resolved and the last possible step has to be executed.

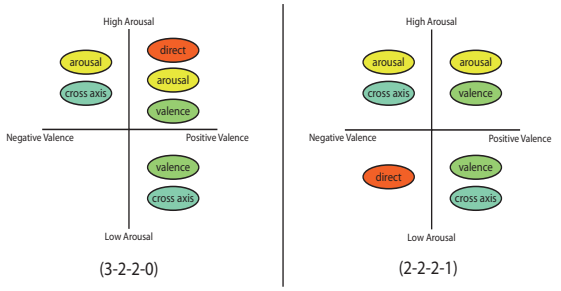

Figure 4: Possible vote distributions in the step 2.

Table 4: Results of the parametric ensemble fusion. */** are the results achieved in the work (Kim and André, 2006).

Subject A					
	HP	*HN*	*LN*	*LP*	*avg*
Feature Fusion*	91.00	92.00	100.00	85.00	**92.00**
Decision Fusion**	64.00	54.00	76.00	67.00	65.00
Ensemble Fusion	100.00	83.33	85.71	96.30	91.34
Subject B					
	HP	*HN*	*LN*	*LP*	*avg*
Feature Fusion*	71.00	56.00	94.00	79.00	75.00
Decision Fusion**	59.00	68.00	82.00	69.00	70.00
Ensemble Fusion	83.33	79.17	91.18	82.76	**84.11**
Subject C					
	HP	*HN*	*LN*	*LP*	*avg*
Feature Fusion*	50.00	67.00	84.00	74.00	69.00
Decision Fusion**	32.00	77.00	74.00	64.00	62.00
Ensemble Fusion	69.57	72.41	74.00	92.86	**77.21**
Subject Independent					
	HP	*HN*	*LN*	*LP*	*avg*
Feature Fusion*	46.00	57.00	63.00	56.00	55.00
Decision Fusion**	34.00	50.00	70.00	54.00	52.00
Ensemble Fusion	49.18	55.70	70.48	66.33	**60.42**

Step 3: Decision through Arousal and Valence Combination. If actually no decision could be established by the previous steps, the emotion-class that was originally determined by arousal and valence ensembles is ultimately chosen as final decision. In practice this case rarely occurs, but it is definitely needed to guarantee that no sample passes the decision process unclassified.

5 RESULTS

All recognition results obtained by using our parametric ensemble fusion are summarized in Table 4 and compared by referencing the results of feature-level fusion (merging the features) and generalized decision-level fusion (majority voting) achieved in the previous work (Kim and André, 2006).

6 CONCLUSIONS

In this paper we proposed a novel decision-level fusion method based on emotion-specific multi-ensemble approach. The objective of this work was to provide guideline towards parametric decision fusion in order to overcome the limitation of the generalized fusion methods that are not able to exploit specific characteristics of a given dataset. Compared to the generalized feature-level and decision-level fusion methods used in the earlier work, the proposed method achieved about 8% improvement of recognition accuracy for both subject-dependent and subject-independent classification.

ACKNOWLEDGEMENTS

The work described in this paper is partially funded by the EU under research grant IST-34800- CALLAS and ICT-216270-METABO.

REFERENCES

Bailenson, J., Pontikakis, E., Mauss, I., Gross, J., Jabon, M., Hutcherson, C., Nass, C., and John, C. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. *Int'l Journal of Human-Computer Studies*, 66(5):303–317.

Chen, L. S., Huang, T. S., Miyasato, T., and Nakatsu, R. (1998). Multimodal human emotion/expression recognition. In *Proc. 3rd Int. Conf. on Automatic Face and Gesture Recognition, IEEE Computer Soc.*, pages 366–371.

Jain, A. and Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. and Machine Intell.*, 19:153–163.

Kim, J. and André, E. (2006). Emotion recognition using physiological and speech signal in short-term observation. In *LNCS-Perception and Interactive Technologies*, pages 53–64. Springer-Verlag Berlin Heidelberg.

Kim, J. and André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Trans. Pattern Anal. and Machine Intell.*, 30(12):2067–2083.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45.

Zeng, Z., Pantic, M., Roisman, G., and Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(1):39–58.