International Journal of Advanced Smart Convergence Vol.9 No.1 193-201 (2020)
http://dx.doi.org/10.7236/IJASC.2020.9.1.193

IJASC 20-1-23

Multimodal Parametric Fusion for Emotion Recognition

Jonghwa Kim

Professor, Dept. Intelligent System Engineering, Cheju Halla University, Jeju Island, Korea
kim@ieee.org

Abstract

The main objective of this study is to investigate the impact of additional modalities on the performance of emotion recognition using speech, facial expression and physiological measurements. In order to compare different approaches, we designed a feature-based recognition system as a benchmark which carries out linear supervised classification followed by the leave-one-out cross-validation. For the classification of four emotions, it turned out that bimodal fusion in our experiment improves recognition accuracy of unimodal approach, while the performance of trimodal fusion varies strongly depending on the individual. Furthermore, we experienced extremely high disparity between single class recognition rates, while we could not observe a best performing single modality in our experiment. Based on these observations, we developed a novel fusion method, called parametric decision fusion (PDF), which lies in building emotion-specific classifiers and exploits advantage of a parametrized decision process. By using the PDF scheme we achieved 16% improvement in accuracy of subject-dependent recognition and 10% for subject-independent recognition compared to the best unimodal results.

Keywords: *Multimodal Emotion Recognition; Decision Fusion; Ensemble; Parametric Decision Making*

1. Introduction

During the last decade, researchers have made great efforts to empower machines with emotional sensitivity, and affective man-machine interaction (AMMI) is becoming an indispensable component of today's emerging high-tech applications. One of the most important prerequisites for the success of AMMI is the reliability of the automatic emotion recognition system. Recently, numerous methods have been proposed to detect human emotions from various modalities including facial expression, gesture, speech, and physiological measurements [1]. The focal point for multimodal emotion recognition is to design efficient fusion methods, which pursue human-like decision process. Basically, the combination of multimodal information can be performed at least at three levels, i.e., data, feature, and decision level. When dealing with observations coming from the same or a very similar modality source, the data-level fusion that simply merges the multiple observations might be the most appropriate method that does not require any separate preprocessing. Feature-

Manuscript Received: February. 27, 2020 / Revised: March. 6, 2020 / Accepted: March. 10, 2020
*Corresponding Author: kim@ieee.org (Jonghwa KIM)
Tel: +82-64-741-1606, Fax: +82-64-741-1605
Professor, Department of Intelligent System Engineering, Cheju Halla University, Jeju 63092, Korea

level fusion is more efficient when the modalities are characteristically tightly-coupled, time synchronized, and mutually complementary in low-level. For multimodal sensory data obtained from heterogeneous (or loosely coupled) modalities, decision-level fusion might be the best choice. 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 a favorable final decision. Since it requires a modality-specific preprocessing and individual classification for each modality, failure and noise sensitivity is relatively low compared to the former methods.

The motivation of this work is twofold; (a) to investigate the impact of additional modalities on recognition accuracy by comparing the recognition performance of various systems with different settings, (b) to develop a novel fusion method using feature ensembles and parametric decision rule for multimodal emotion recognition. In this paper, we present emotion recognition results obtained by combining three modalities that are most frequently used in literature for unimodal emotion recognition, i.e. physiological signals, speech, and facial expression. Moreover, we propose a novel classification scheme for multimodal emotion recognition, which exploits advantages of feature- and decision-level fusion and binary (dichotomous) class tree classification.

2. Related Work

A vast amount of studies in the automatic emotion recognition has been reported during the last decade. Researchers have shown that emotion can be successfully recognized by detecting affective cues in typical expression channels of emotion such as speech signal, facial expression, gesture, and physiological changes [2, 3]. Besides unimodal approaches, many studies in multimodal affect recognition have also been introduced by exploiting complementary combination of different modalities, mostly by combining audiovisual information, e.g., speech and facial expression [4, 5]. In the work of Busso et al. [6] an emotion-specific comparison of feature-level and decision-level fusion has been reported by using an audiovisual database containing four emotions, sadness, anger, happiness, and neutral state, deliberately posed by an actress. They observed that feature-level fusion was most suitable for differentiating anger and neutral state while decision-level fusion performed better for happiness and sadness. They concluded that the best fusion method depends on the application.

In our previous work [7] on bimodal fusion of physiological signals and speech we proposed a hybrid-fusion method that utilizes decision of feature fusion for final decision fusion. Recently, Povolny et al. [8] achieved recognition accuracy of 71% for arousal and 60% for valence by exploiting fusion of audio, video, and physiological data.

3. Trimodal Dataset

3.1 Experimental Setting

To generate spontaneous multimodal emotion dataset, we developed a Wizard-Of-Oz quiz program which is similar to the German TV quiz show ``Who wants to be a millionaire?''. In the graphical interface, a virtual agent presents the quiz and communicates with the user. A human quiz master (wizard) has control of the agent and the actual course of the quiz, following a working script to evoke situations that lead to a certain emotional response. The interface does not offer the user any letters as abbreviations for the single options, but forces the user to answer always with a complete sentence, in order to get sufficient length of speech data. The virtual

agent is represented by a disembodied voice system using the AT&T Natural Voices speech synthesizer which transforms the typed text by the wizard to a natural voice. The wizard may freely type utterances, but also has access to a set of macros that contain pre-defined questions or comments which made it easier for the human wizard to follow the working script and to get reproducible situations. The working script of the wizard contains four strategic phases that serve to induce the four representative emotional states on the 2D (valence vs. arousal) emotion model. The entire session implies the four phases and takes about 45 minutes.

Phase 1 (LP): The users are offered a set of very easy questions every user is supposed to know to achieve equal conditions for all of them. This phase is characterized by a slight increase of the score and gentle appraisal of the agent and serves to induce an emotional state of low arousal and positive valence (LP) in the user.

Phase 2 (HP): In the second phase, the user is confronted with extremely difficult questions nobody is supposed to know. Whatever option the user chooses, the agent pretends the users answer is correct so that the user gets the feeling that one hits the right option just by chance. In order to evoke high arousal and positive valence (HP), this phase leads to a high gain of money.

Phase 3 (LN): During the third phase, the wizard tries to stress the user by presenting a mixed set of solvable and difficult questions. Yet, this should not cause a drastic loss of money. Furthermore, the agent often attempts to provide superfluous information related to the topics addressed in the questions so that the user will be boring. Thus, this phase should lead to low arousal and negative valence (LN).

Phase 4 (HN): Finally, the user gets frustrated by unsolvable questions. Whatever option the user chooses, the agent always pretends the answer is wrong, resulting in a high loss of money. Furthermore, we include simple questions for which we offer similar-sounding options. The user is supposed to choose the right option, but the situation makes the user believe that the speech recognizer is not working properly and deliberately select the wrong option. This phase is intended to evoke high arousal and negative valence (HN).

3.2 Collected Sensor Data

During the quiz sessions with three male German-speaking students in their twenties, the speech (48 KHz/16 Bit, mono), video (webcam, 640 x 480), and the 5-channel physiological signals are measured; electromyogram (EMG), skin conductivity (SC), blood volume pulse (BVP), temperature (Temp), and respiration (RSP). The sampling rates are 32 Hz for EMG, SC, RSP, and Temp, 256 Hz for BVP. Each of long class segments is annotated by two labelers and self-reports of subjects. We then trimmed each class segment into many small samples based on spoken phrases. As a result we obtained a total of 343 trimodal samples (subject A: 94, subject B: 105, subject C: 144) for classification process.

4. General Methodology and Result

In this section recognition results of uni- and multimodal approaches are presented, which motivated our novel decision fusion scheme described in the Section 5.

4.1 Multimodal Feature Calculation

Physiological features (BIO): we calculated a total of 77 features for each segment of 5-channel biosignals. For details about physiological features, we refer to our previous work [9].

Speech features (SPE): in frequency domain, we calculated three spectral features using the STFT; pitches using a window length of 40ms, energy spectrum, and formant object using a window length of 25ms.

Moreover 10 MFCCs (Mel-frequency cepstral coefficients) from each segment are calculated using a window length of 15ms. From pitch and energy spectrum, also the series of the minima and maxima, and of the distances, magnitudes and steepness between adjacent extrema were obtained. For the MFCCs, we first exponentiated the cepstral coefficients to obtain non-negative values and calculated the spectral entropy as in the case of the biosignal in order to capture the distribution of cepstral energy. As a result, we obtained a total of 61 features from each speech segment.

Facial features (VID): similar to our previous work [10] we identified 18 points of interest (POI) that are relevant to affective facial expression. To each fiducial point, we applied the Gabor filter and obtained 18 complex coefficients, i.e., a total of 324 features for each image.

4.2 Classification

Since the goal of this work relates to conceptualizing new efficient decision fusion scheme, rather than performance comparison of existing classifiers, we use single pLDA (pseudoinverse linear discriminant analysis [9], for all classification problems after feature selection using sequential backward search (SBS) in this work. For the multimodal (bimodal and trimodal) approaches we analyze the performance of two common fusion methods; feature level fusion (FF) which merges all multimodal feature sets into single feature vector and employs single classifier, and decision level fusion (DF) which classifies each modality separately and combines the multiple decisions for final decision by using majority voting for example.

4.3 Recognition Results

Unimodal Recognition: Table 1 presents the correct classification ratio CCR of subject-dependent (Subjects A, B, and C) and subject-independent (All) classification where the features of all subjects are merged. Particularly for the subject-independent case the normalization of merged feature vector is necessary to degrade possible individual difference of magnitude scales. We used mean-standard deviation (z-score) normalization.

Table 1. Unimodal classification results in CCR (%) of four emotions (HP, HN, LN, LP)

Modality	*Subject A*	*Subject B*	*Subject C*	*All**
BIO	**73.3**	57.6	61.4	49.5
SPE	69.6	**73.4**	**70.5**	**54.1**
VID	63.6	58.5	60.2	47.7

*: subject-independent classification

Multimodal Recognition: To answer the question, whether the common logic ``the more data the better precision" is valid for automatic emotion recognition, we considered all possible combinations of the given three modalities. As shown in Table 2, the logic seems not to necessarily be valid for our experiment. Overall, we achieved about 12% improvement of recognition accuracy with multimodal approaches compared to unimodal results.

Table 2. Multimodal classification results in average CCR (%) of four emotions. Best multimodal combination for each subject is in bold.

BIO + SPE				
Fusion	*Subject A*	*Subject B*	*Subject C*	*All*
Feature Fusion	81.5	74.8	73.4	58.0
Decision Fusion	75.2	70.4	70.5	53.6
SPE + VID				
Fusion	*Subject A*	*Subject B*	*Subject C*	*All*
Feature Fusion	84.1	75.8	**76.3**	60.6
Decision Fusion	82.6	71.1	68.4	49.5
VID + BIO				
Fusion	*Subject A*	*Subject B*	*Subject C*	*All*
Feature Fusion	**88.4**	75.6	73.3	57.0
Decision Fusion	84.3	74.1	71.6	55.9
BIO + SPE + VID				
Fusion	*Subject A*	*Subject B*	*Subject C*	*All*
Feature Fusion	84.7	**81.8**	76.1	**62.1**
Decision Fusion	81.5	76.9	71.9	61.1

5. Parametric Decision Fusion

By taking the best results (88.4%, 81.8%, 76.3%) of the subjects in Table 2, it shows an average accuracy of 82.1% for subject-dependent and 62.1% for subject-independent classification. During this first analysis with nonparametric strategy, we could observe following evidences; **a)** the best modality and the best combination of modalities for emotion recognition is not determinable but varies with subject, **b)** the disparity between recognition rates of single classes is remarkably high and impairs the average CCR ultimately, **c)** the decision-level fusion using a generalized decision making algorithm such as majority voting and Borda count can often be faced with problem of extremely unbalanced overall performance due to overemphasized classes by repetitive weighting.

Based on these observations, we developed a novel fusion method, we called ``parametric decision fusion (PDF)", which lies in building class-specific classifiers and parametrized decision process. Figure 1 shows the architecture of PDF system.

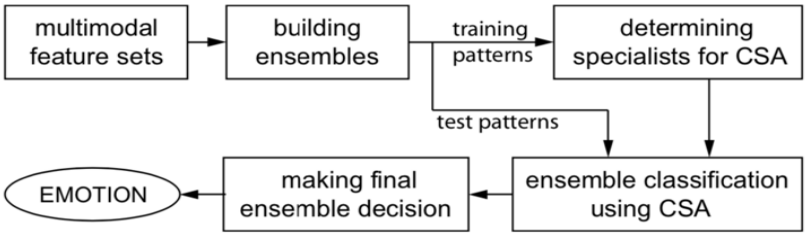

Figure 1. Architecture of PDF recognition system

5.1 Building Dichotomous Classifiers

The basic idea of PDF is to build multiple classifiers that are essentially independent for each other within a given classification problem, and to determine special classifiers for each of them. In multimodal approach, therefore, the number of classifiers increases with the number of modalities and their combinations. For the work in this paper, we consider three dichotomous classifiers, in addition to the four-class classifier (HP, HN, LN, LP), that are built by grouping two classes into one classifier member according to the two reference axes of the 2D emotion model. PDF aims to make maximum use of such emotion-specific restructuring of classes and exploit the advantage of binary classification, for which linear classifiers such as pLDA might be most suitable. Figure 2 illustrates the possible two-class combinations based on arousal, valence, and cross axis.

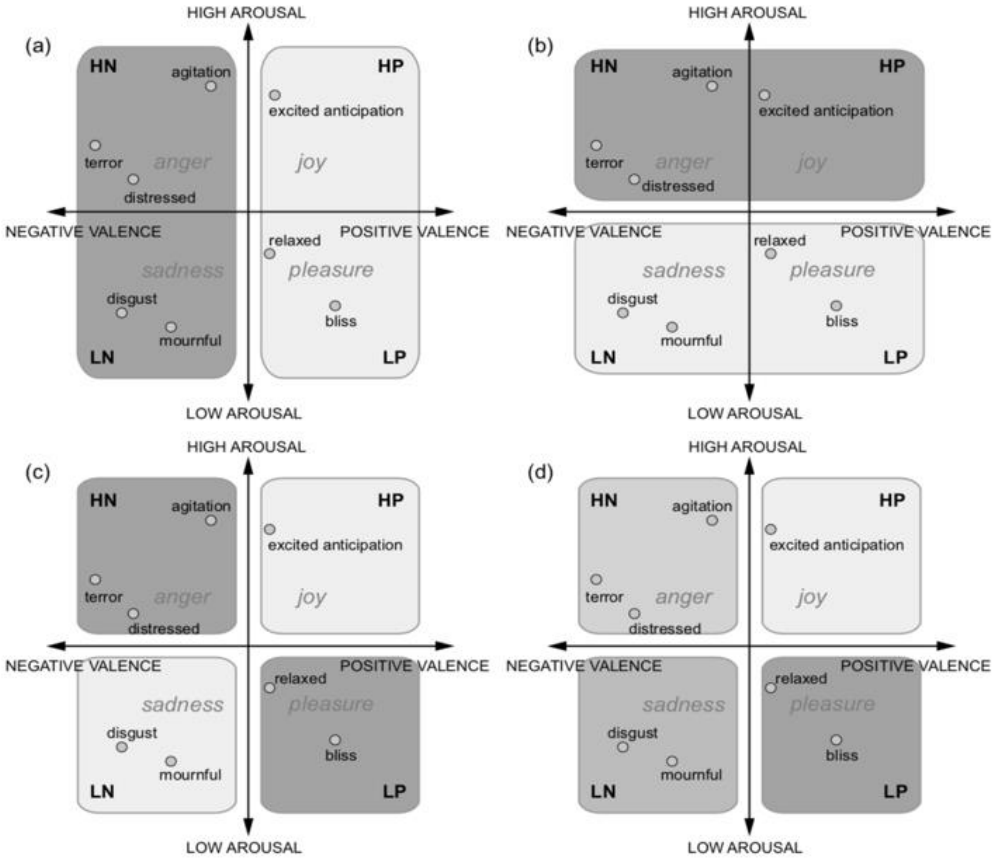

Figure 2. Suggested emotion-specific dichotomous ensembles

Each of these four classifiers produces their own decisions independently, i.e., three binary decisions and one four-class decision.

5.2 Cascaded Specialists Algorithm (CSA)

In most classification problems it is desired to get a well-balanced recognition rate for all classes, without high disparity between the classes. However, it is often overseen in literature to monitor single class performance that might be in practice more important than an average overall accuracy. Taking this into account we proposed cascaded specialists algorithm (CSA) in our previous work (we refer to [11] for more

details). In this work we further adapted the CSA to PDF.

5.3 Making Decision

We apply the CSA to all classifiers separately. Table 3 shows examples of specialists determined for three dichotomous classifiers and four-class classifier, e.g., the specialist for arousal classification of subject-independent case is a pLDA classifier trained by SPE feature vector, while there are four different specialists for the four-class classifier.

Table 3. Example of selected specialists for CSA

Specialists for dichotomous classifiers					
Subject A			**Subject Independent**		
Arousal (**BIO**)			Arousal (**SPE**)		
high	*low*	*avg*	*high*	*low*	*avg*
89.1	89.6	89.4	63.6	86.2	74.9
Valence (**BIO**)			Valence (**SPE**)		
positive	*negative*	*avg*	*positive*	*negative*	*avg*
91.8	100	95.9	70.4	76.1	73.3
Cross Axis (**BIO**)			Cross Axis (**VID**)		
HP+LN	*HN+LP*	*avg*	*HP+LN*	*HN+LP*	*avg*
95.4	90.2	92.8	72.9	83.6	78.3

Specialists for direct classifier of Subject A			
HP	*HN*	*LN*	*LP*
BIO (86.4)	**VID** (62.5)	**SPE** (76.2)	**BIO** (74.1)

CCR in %

Consequently, we obtain a total of four different decisions in terms of arousal, valence, cross axis, and direct classifier. To make a final decision among four emotions we use a voting method that is quite straightforward. We count the votes of four decisions (all votes are uniformly valued without weighting) onto the four quadrants of the 2D emotion model and then determine the quadrant (emotion), which obtained the most votes, as the final decision. Classification is guaranteed in most cases, except for the case of a draw, which rarely occurs in practice. In such case, we take the voting result of arousal and valence classifiers as a final decision.

5.4 Results

Table 4 summarizes recognition results of PDF in comparison with best uni- and multimodal recognition results from Table 1 and 2. The proposed PDF could improve the best overall recognition accuracy of the multimodal approach using feature-level fusion by 5.3% and 14.6% compared to the best uni-modal approach.

Moreover, PDF succeeded not only in improving recognition accuracy subject-independently, but also in rectifying the high disparity of single class accuracies observed in uni- and multimodal approaches as shown in the Table 4. For the significance evaluation of the improvement, we calculated paired t-tests and Cohen's effect sizes (d) [12],

$$d = \frac{x_1' - x_2'}{\sqrt{s_1^2 + s_2^2 / 2}} \quad (1)$$

where the x' and s denote the mean value and standard deviation, respectively. It turned out that the improvement of recognition accuracy related to the best multimodal feature fusion ($x' = 82.1\%, s = 6.05$) and PDF ($x' = 88.4\%, s = 6.33$) is significant ($p < 0.01$) with large effect size ($d < 1.01$). For the unimodal approach and PDF, the effect size increases even up to $d < 3.46$.

Table 4. Recognition results of PDF, compared with best uni- and multimodal (feature level fusion) results.

	HP	*HN*	*LN*	*LP*	*Average*
Subject A					
BIO	86.4	70.8	61.9	74.1	73.3
BIO+VID	90.9	87.5	90.5	92.6	88.4
PDF	100	87.5	95.2	96.3	**94.8**
Subject B					
SPE	72.2	62.5	79.4	79.3	73.4
BIO+SPE+VID	86.8	76.4	89.7	85.4	81.8
PDF	88.9	83.3	91.2	89.7	**88.3**
Subject C					
SPE	60.9	72.4	70.0	78.6	70.5
SPE+VID	62.5	86.1	82.7	87.4	76.3
PDF	74.5	82.3	84.6	87.0	**82.1**
Subject Independent					
SPE	30.1	55.8	69.5	53.1	54.1
BIO+SPE+VID	48.1	62.9	78.3	59.0	62.1
PDF	52.5	58.2	69.5	77.6	**64.4**

$\mathcal{CCR}$ in %

6. Conclusion

In this paper, we presented trimodal approach for automatic emotion recognition using a novel parametric decision fusion. From unimodal to trimodal we investigated the impact of additional modality on recognition accuracy and compared the classification performance between common feature- and decision-level fusion. It turned out that bimodal approach (regardless of which combination) always improves recognition accuracy of unimodal approach, while the performance of trimodal approach varies strongly depending on the individual. In line with previous works in literature, feature-level fusion outperformed by far common decision-level fusion. Based on the observations we proposed a novel fusion method, called parametric decision fusion (PDF), and showed its potential and effectiveness with the significantly improved recognition accuracies for subject-dependent and subject-independent case as well. As a future work remains evaluating PDF with extended number of subjects and modalities in order to improve it as a generalizable solution for multimodal fusion.

References

[1] Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. IEEE Signal Processing Mag., 18, pp. 32–80, DOI: 10.1109/79.911197, 2001

[2] Wu, C.H.; Lin, J.C.; Wei, W.L. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing, 3*, pp. 1–18, 2014

[3] Jang, E.; Park, B.; Kim, S.; Sohn, J. Emotion classification by Machine Learning Algorithm using Physiological Signals. in Proc. of Computer Science and Information Technology, Singapore, pp. 1–5, 2012

[4] Nicolaou, M.; Gunes, H.; Pantic, M. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence- Arousal Space. *IEEE Trans. on Affective Computing*, *2*, 92-105, DOI: 10.1109/T-AFFC.2011.9, 2011

[5] Zhang, S.; Zhang, S.; Huang, T.; Gao, W.; Tian, Q. Learning Affective Features with a Hybrid Deep Model for Audio-Visual Emotion Recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, pp. 3030 – 3043, DOI: 10.1109/TCSVT.2017.2719043, 2017

[6] Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.H.; Kazemzaden, A.; Lee, S.; Neumann, U.; Narayanan, S. Analysis of emotion recognition using facial expression, speech and multimodal information. ICMI'04, pp. 205–211, DOI: 10.1145/1027933.1027968, 2004

[7] Kim, J. Bimodal Emotion Recognition using Speech and Physiological Changes. In Robust Speech Recognition and Understanding; I-Tech Education and Publishing, chapter 15, pp. 265–280, DOI: 10.5772/4754, 2007

[8] Povolny, F.; Matejka, P.; Hradis, M.; Popkova, A.; Otrusina, L.; Smrz, P.; Wood, I.; Robin, C.; Lamel, L. Multimodal emotion recognition for AVEC 2016 challenge. in Proc. of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, pp. 75–82, DOI: 101145/2988257.2988268, 2016

[9] Kim, J.; André, E. Emotion recognition based on physiological changes in music listening. IEEE Trans. Pattern Anal. and Machine Intell., 30, pp. 2067–2083, DOI: 10.1109/TPAMI.2008.26, 2008

[10] Kim, J.; Jung, F. Emotional Facial Expression Recognition from Two Different Feature Domains. In Proc. of ICAART 2010, Intl. Conf. on Agents and Artificial Intelligence, pp. 631–634, 2010

[11] Kim, J.; Lingenfelser, F. Ensemble Approaches to Parametric Decision Fusion for Bimodal Emotion Recognition. Proc. of Biosignals 2010, Int. Conf. on Bio-inspired Systems and Signal Processing, pp. 460–463, DOI: 10.5220/0002753204600463, 2010

[12] Cohen, J. A Power Primer. *Psychological 'Bulletin*, *112*, pp. 155–159, DOI: 10.1037//0033-2909.112.1.155, 1992