ICASSP 2010, Dallas TX

DECISION LEVEL COMBINATION OF MULTIPLE MODALITIES FOR RECOGNITION AND ANALYSIS OF EMOTIONAL EXPRESSION

Angeliki Metallinou, Sungbok Lee and Shrikanth Narayanan

Department of Electrical Engineering, University of Southern California,
Los Angeles, CA 90089-2560

metallin@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

ABSTRACT

Emotion is expressed and perceived through multiple modalities. In this work, we model face, voice and head movement cues for emotion recognition and we fuse classifiers using a Bayesian framework. The facial classifier is the best performing followed by the voice and head classifiers and the multiple modalities seem to carry complementary information, especially for happiness. Decision fusion significantly increases the average total unweighted accuracy, from 55% to about 62%. Overall, we achieve average accuracy on the order of 65-75% for emotional states and 30-40% for neutral state using a large multi-speaker, multimodal database. Performance analysis for the case of anger and neutrality suggests a positive correlation between the number of classifiers that performed well and the perceptual salience of the expressed emotion.

Index Terms— Multimodal Emotion Recognition, Hidden Markov Model, Bayesian Information Fusion, Perceptual Salience

1. INTRODUCTION

Emotional expression is a multimodal process. The affective state may be transmitted by one or more of various channels such as face, voice, speech content, body movement and posture. Moreover, emotion is perceived by combining those channels that may carry complementary, supplementary or even conflicting information. In [1] it is stated that the semantic content of a message contributes only 7% of the overall impression while the vocal and facial modalities contribute 38% and 55% respectively. Therefore, an emotion recognition system that takes into account multiple modalities may be able to achieve robust emotion recognition performance, even under noisy conditions or when subtle emotions are expressed. The widespread use of cameras and microphones facilitates the use of audio-visual information for emotion recognition applications.

The state of the art and challenges faced in developing an affect-sensitive multimodal human computer interface have been discussed in [2]. Multimodal research on emotion recognition has focused mostly in combining the face and voice modalities [3] [4]. Researchers have also used face images with markers to minimize the noise introduced by automatic facial feature detection [5],[6]. In [3] a multi-stream HMM is used for audio-visual emotion recognition. In terms of decision fusion, [7] provides an overview of existing classifier fusion methods, while in [8] authors propose a Bayesian framework for fusion.

In this study, we use a multimodal database and model the multiple channels with variable levels of detail. During emotional speech, the voice and the lower face are modulated by both speech and emotion. We take into account the speech-related modulations by explicitly modeling articulation. For the upper face and the head movements we use coarser modeling. In order to combine the individual classifiers we apply a Bayesian framework for decision-level fusion. We enrich the Bayesian fusion approach, with in-domain information so as to select the informative modalities and improve neutrality detection. Fusion at the decision level enables us to intrepret the behavior of different classifiers, in terms of recognition performance, and to gain insights about the role of multiple modalities during emotional expression. Finally, we examine the relation between the number of classifiers that perform well and the perceptual salience of an utterance in the valence-activation domain.

Results indicate that face is the best performing modality followed by voice and head motion. Face and voice seem to carry complementary information, especially for happiness. The orientation of the head seems to convey important emotional information, especially for sadness. Fusion of those classifiers significantly improves the overall performance from 55% to about 62%. Neutrality is the hardest class to recognize, however the proposed extensions to the Bayes fusion framework result in significant improvement in neutrality classification. Overall, we achieve average accuracies of the order of 65-75% for emotional states, 30-40% for neutral state and about 62% for total unweighted performance, using a large multi-speaker and multimodal database.

2. METHODOLOGY

2.1. Database

In this study, we use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database. This database contains approximately 12 hours of audiovisual data from five mixed gender pairs of actors, male and female [9]. IEMOCAP contains detailed face and head information obtained from motion capture as well as video, audio and transcripts of each session. Two techniques of actor training were used; scripts and improvisation of hypothetical scenarios. The goal was to elicit emotional displays that resemble natural emotional expression. Dyadic sessions of approximately 5 minute length were recorded and were later manually segmented into utterances. Each utterance was annotated by at least 3 annotators into categorical (anger, happiness, neutrality, etc) as well as dimensional tags (valence, activation, dominance). The annotations are a result of the overall impression of an utterance, since annotators considered audio, video, speech content and the interaction context. Therefore,

This work was supported in part by funds from NSF and the Army

the multimodal expression of emotion was taken into account both during the collection and the annotation of the database.

We examine all 10 available speakers and use multimodal information; the face and head motion capture data, the voice waveform and the phoneme-level transcripts. We examine classes of anger, happiness, excitation, neutrality and sadness. We have merged the classes of happiness and excitation into a single class which we will refer to as happiness. All utterances examined have been tagged by at least three annotators across which there was majority consensus regarding the emotional tag.

2.2. Modeling Individual Modalities

2.2.1. Face Modality

The IEMOCAP database contains detailed facial marker information, as illustrated in figure 1. Face markers are normalized for head rotation and translation and the nose marker is defined as the local coordinate center of each frame. We use information from 46 facial markers, the (x,y,z) coordinates. In a companion paper submitted to ICASSP10, we examine in detail the facial information stream and here we apply these results [10].

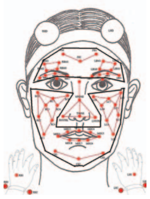

Fig. 1. Face and head marker positions and separation of face into upper and lower facial regions

Speaker face normalization smooths individual facial characteristics, that are unrelated to emotion. Our speaker normalization approach consists of finding a mapping from the individual average face to the general average face. This is achieved by shifting the mean value of each marker coordinate of each speaker to the mean value of that marker coordinate across all speakers. We use Principal Feature Analysis (PFA), to extract the face features [11]. This method performs Principal Component Analysis (PCA) as a first step and selects features (here marker coordinates) so as to minimize the correlations between them. We average the neighboring facial markers from 46 to 28, perform PFA, and finally normalize the selected face coordinates. We select 30 features, because the PCA transformation explains more than 95% of the total variability, and we append the first derivatives, resulting in a 60-dim representation[10].

In order to reduce the speech-related variability, we explicitly model articulatory movements. We use the concept of a viseme which is the lip shape during the voicing of a phoneme. Here, we use a phoneme to viseme mapping resulting in 14 visemes from the 42 English phonemes [10],[12]. Data corresponding to each viseme are grouped using the available phoneme-level transcriptions. We train a Hidden Markov Model (HMM) for each emotional viseme. During recognition, we condition on the knowledge of the current viseme and recognize the underlying emotion. For model training, we use the HTK Toolbox [13].

Alternatively, we split the face into an upper and a lower region, as shown in Figure 1 and performed feature extraction separately. The intuition is that upper and lower face movements have low correlation because of different underlying muscles and communicative functionality and that the lower face is much more affected by articulation. We extract 20 PFA features from the lower face and 15 features from the upper face. The number of features is selected so that at least 95% of the total data variance is explained. After appending the first derivatives, we obtain a 40-dim and a 30-dim representation for the lower and upper face respectively. The lower face is modeled by HMMs trained for each emotional viseme. The upper face is modeled by Gaussian Mixture Models (GMMs) with no viseme information, one for each of the examined emotions.

2.2.2. Voice Modality

We extract Mel Filterbank Coefficients (MFB), since they have been shown to perform better than Mel-Frequency Cepstral Coefficients (MFCCs) for emotion recognition tasks [14]. We extract the first 13 MFB coefficients as well as the pitch and energy values. We also append the first and second derivatives, resulting in a 45-dim representation. Speaker normalization is important since it has been shown that prosodic features such as pitch and energy are speaker and gender dependent [15]. We normalize pitch and energy values, using a similar approach used on the face features; we shift the the mean value of each feature of each speaker to the mean value of that feature across all speakers. Using the grouping proposed by [14], we group the data into 7 broad phonetic categories and we train an HMM for each emotional phonetic category. During recognition, we condition on the knowledge of the phoneme and recognize the underlying emotion.

2.2.3. Head Modality

The head features consist of the head translation in the (x,y,z) directions as well as the head angles (yaw, pitch and roll). Translations are derived from the nose marker and head angles are computed from all the markers using a technique based on Singular Value Decomposition (SVD), as described in [9]. We also compute the first and second derivatives of these features, resulting in an 18-dim representation. We train a GMM for each emotion.

In order to gain intuition about the inportance of these features for discriminating between emotional classes, we compute the Fisher Criterion values for each of the head features. Fisher criterion maximizes the between class variability and minimizes the within class variability [16]. According to this analysis the most discriminating features are the head angles, especially pitch and yaw. This agrees with our intuition that tilting or lowering of the head often conveys affect.

2.3. Bayesian Framework for Multimodal Fusion

Using the decisions of each individual classifier, we obtain a histogram with the amount of time that an utterance is classified as angry, happy, neutral or sad. We can use this histogram to approximate the probability that an utterance belongs to each of the emotional classes [17].

Examining and fusing the modalities at the decision level enables us to gain intuition about how multimodal cues interplay during emotional expression and to incorporate in-domain information in the decision process. Bayesian statistics provide a systematic way to combine empirical evidence with prior beliefs and to fuse multiple cues. It is argued that human behavior is close to that predicted by the bayesian decision theory [18].

The Bayesian fusion framework that we apply is proposed in [8]. It uses the conditional error distributions of each classifier to approximate uncertainty about that classifier's decision. The combined decision is the weighted sum of the individual decisions. Given a problem with K classes and C different classifiers, $\lambda_i, i = 1, \dots, C$, we like to infer the true class label ω, given the observation x. Assuming that for each classifier λ_i we have a predicted class label $\widetilde{\omega_k}$, $k = 1, \dots, K$, then the true class label can be derived as follows:

$$P(\omega|x) \approx \sum_{i=1}^{C} \sum_{k=1}^{K} P(\omega|\widetilde{\omega_k}, \lambda_i) P(\widetilde{\omega_k}|\lambda_i, x) P(\lambda_i|x)$$

Probabilities $P(\omega|\widetilde{\omega_k}, \lambda_i)$ and $P(\lambda_i|x)$ are used to weight the combined decision and can be approximated from the confusion matrix of classifier λ_i. The probability $P(\widetilde{\omega_k}|\lambda_i, x)$ of each predicted emotion, given a classifier and an utterance x is approximated using the histogram of the utterance.

During affective communication, some of the modalities decribed may not be emotionally activated, thus considering them introduces noise instead of boosting the performance. The number and type of activated modalities may change dynamically between utterances. In practice, for each test utterance we select only the activated modalities; those where the probability of the most likely emotion exceeds a threshold (the histogram ressembles a delta function) [17].

Neutrality detection is problematic because the neutral class is ill-defined and contains diverse expressions. One could define neutrality as absense of any emotion. Accordingly, instead of trying to recognize neutrality by assessing whether an observation matches a neutral probability distribution, we could detect neutrality when the observation does not match any of the emotional probability distributions. The intuition is that the emotional models are better trained and more reliable because emotional expressions may be better defined and less diverse than neutrality. Here, we detect neutrality when none of the modalities examined is emotionally activated.

3. EMOTION RECOGNITION EXPERIMENTS

3.1. Experimental Setup

We organize our emotion recognition experiments using 10-fold leave-one-speaker-out cross validation. The mean and standard deviation of the number of test utterances across the folds is 59 ± 28 angry, 79 ± 25 happy, 56 ± 22 neutral and 62 ± 23 sad utterances. The presented recognition results are the speaker-independent averages over the 10 folds. For the individual modality classifiers, we try various number of mixtures, ranging from 4 to 32, and we report the best performance per utterance. Majority rule is used to obtain the utterances-level decision. For the classifier fusion, we compute the weights on each train set and apply them to the corresponding test set.

3.2. Results and Discussion

In Table 1 we present the emotion classification percentages per utterances for each of the individual modalities. We report the mean of the 10-fold cross validation. For the total unweighted accuracy (%UW) we also report the standard deviation. We use the following notation; f=face, uf=upper face, lf=lower face, v=voice and h=head. The parentheses next to each model name indicate the number of gaussian mixtures that were used.

We notice that the face modality has the best performance in terms of total unweighted accuracy, followed by the voice and the head. The lower face performs significantly better than the upper face and is almost as good as the total face classifier. This suggests that the lower face (chin, mouth, cheeks) alone conveys most of the emotional information. The face and voice modality seem to carry complementary information for the case of happiness. This might indicate that when the face portrays obvious expressions of happiness (e.g through a smile) the overal impression of happiness can be transmitted from the face alone, therefore the voice could be less emotionally modulated. The head classifier has good performance for the sad and the neutral class. It seems intuitive that the head angles (lowering, tilting) are correlated with sadness.

Table 1. *Single Modality and Bayes Fusion Classification Percentages per utterance for 10-fold leave-one speaker out cross validations.*

single	%ANG	%HAP	%NEU	%SAD	%UW
f-HMM(16)	57.52	76.98	34.79	53.68	55.74 ± 5.26
uf-GMM(8)	51.18	74.35	25.84	36.43	46.95 ± 6.66
lf-HMM(8)	53.35	75.62	36.84	50.93	54.19 ± 4.76
v-HMM(4)	69.68	21.01	35.23	76.84	50.69 ± 5.14
h-GMM(16)	39.72	27.04	46.55	53.60	41.72 ± 7.21
Bayes Fusion	%ANG	%HAP	%NEU	%SAD	%UW
f+v	67.28	68.68	27.61	78.71	60.57 ± 4.26
uf+lf+v	74.70	74.01	16.33	84.03	62.27 ± 3.41
uf+lf+v*	67.78	72.24	32.24	72.33	61.15 ± 3.62
f+v+h	63.93	63.59	33.26	81.94	60.68 ± 5.22
f+v+h*	63.87	64.37	41.58	77.35	61.79 ± 3.96
uf+lf+v+h	69.58	72.18	20.86	84.34	61.74 ± 3.17
uf+lf+v+h*	68.77	72.52	31.48	76.92	**62.42 ± 3.16**

In Table 1 we present the average emotion classification percentages per utterances across the 10 folds, obtained from Bayes fusion. In terms of notation, f+v denotes combination of face and voice classifiers and similarly for the rest modalities. Symbol * denotes that we also applied modality selection and neutrality detection.

Fusing face and voice improves overall performance from about 55% to 62%. Separately modeling upper and lower face improves recognition of emotional states but impairs neutrality detection. An explanation could be that none of the single modalities achieves high neutrality recognition, thus detecting neutrality is an unreliable decision according to Bayes fusion. Including the head classifier, which achieves relatively good neutrality performance, slightly improves overall performance and especially recognition of neutrality. Our approach of selecting the activated modalities and detecting neutrality as absense of any emotion, significantly improves accuracy of neutral state while overall performance remains about the same. Our best performance is on the order of 65-75% for emotional states, 30-40% for neutral state and about 62% in terms of total unweighted performance.

An analysis of our best performing approach (uf+lf+v+h*) indicates that the number and type of modalities that are selected varies greatly across utterances. On average, 32% of the time we select four modalities, 34% three, 24% two modalities. These trends are consistent across emotions and are indicative of the dynamic and diverse nature of emotional expression. The percentage of neutral utterances where none of the classifiers appears emotionally activated is 16% while this percentage is 7%, 9% and 5% for anger, happiness and sadness respectively. Also out of all utterances where no classifier appears emotionally activated, 38% are neutral. This explains the good performance of our neutrality detection and indicates that recognizing neutral as a lack of emotion may be a viable approach.

4. MULTIMODAL EMOTION IN THE PERCEPTUAL DOMAIN

We analyze the behavior of our best performing classifiers, that is the face and the voice classifier, by computing how many times the decisions of these classifiers agree with the global utterance tag. The average percentages across the 10 folds are presented in Table 2.

Table 2. *Percentages of each modality agreement with the global utterance tag.*

agree with tag	%ANG	%HAP	%NEU	%SAD	%TOTAL
f or v	80.78	81.75	53.64	88.87	77.05
f and v	45.75	15.97	12.26	41.80	28.26
f xor v	35.03	65.78	41.39	46.95	48.79

The first line of this table can be interpreted as the upper bound of the performance that can be achieved by fusing our face and voice classifiers. The percentages per emotion follow the same trends as the classification percentages of the previous section. The count is low for neutrality explaining in part the low classification accuracy of the neutral class. Inability to recognize the emotion using certain modalities could be attributed to inadequate features and modeling. It may also be the case that the emotion is transmitted by other modalities, rather than the ones that we examine. Looking at the lines 2-3 of Table 2, we notice that most of the time less than 2 of the face and voice modalities recognize the emotion. This indicates that classifier fusion is very important so as to amplify the weaknesses of individual classifiers and resolve cases when not all modalities are emotionally modulated.

We examine the relation between the perceptual salience of an utterance and the the number of classifiers that correctly recognize the emotion of the utterances. Perceptual salience is described by the valence (positive v.s negative) and activation (calm v.s excited) ratings. Those properties are rated on scales 1-5 and are averaged across at least 3 annotators. Low valence (V) denotes negative and low activation (A) denotes calm. For example, we would expect angry utterances to correspond generally to high activation and low valence and neutral utterances to have medium values for both attributes.

We find that from all angry utterances that are tagged as highly salient ($V<2$ and $A>4$) 82% are recognized as angry by both face and voice, 18% by one and 0% by no classifiers. These percentages change to (2,1,0) classifiers = (51,37,12)% for utterances in the perceptual area of medium salience ($2\leq V<3$ and $3<A\leq 4$), and to (13,43,44)% for utterances in the perceptual area of low salience ($V\geq 3, A\leq 3$). We notice a clear tendency that the more salient utterances are recognized by more classifiers. This agrees with our intuition that when the angry emotion is perceived as salient from the receiver, it is likely that the transmitter strongly modulates both face and voice expression and therefore it is easier to automatically recognize anger from both modalities. Similarly, for neutrality, the percentage of utterances that are recognized by both classifiers increases for utterances towards the center of the perceptual domain (V and A close to 3). Such tendencies are not supported by our data for happiness and sadness.

5. CONCLUSION AND FUTURE WORK

In this work, we have examined multimodal expressions of emotion for the purpose of recognition and analysis. We have modeled face, voice and head cues with variable levels of detail and used them individually or after applying late fusion for emotion recognition. Classifier fusion significantly outperforms individual classifiers since it may resolve cases when not all examined modalities are emotionally activated or some modalities are noisy. Overall, we achieve average performance on the order of 65-75% for emotional states, 30-40% for neutral state and about 62% in terms of total unweighted performance.

Possible future directions include taking into account additional modalities such as content of speech, i.e. detection of emotionally salient words as well as context of the interaction, i.e past emotional state. Moreover, as indicated by the difference between the upper bounds of performance in the previous section and the actual classification performance, there is further room for improvement as far as multimodal fusion is concerned. Finally, analyzing the dynamic relations between the multiple complementary or supplementary modalities during emotional expression is an interesting and challenging problem.

6. REFERENCES

[1] A. Mehrabian, “Communication without words,” *Psychology today*, vol. 2, pp. 53–56, 1968.

[2] M. Pantic and L.J.M. Rothkrantz, “Toward an affect-sensitive multimodal human-computer interaction,” *Proceedings of the IEEE*, vol. 91, pp. 1370–1390, 2003.

[3] Z. Zeng, J. Tu, M. Liu, T.S. Huang, B. Pianfetti, D. Roth, and S. Levinson, “Audio-visual affect recognition,” *IEEE Transactions on Multimedia*, vol. 9, pp. 424–428, 2007.

[4] L.S. Chen and T.S. Huang, “Emotional expressions in audiovisual human computer interaction,” in *IEEE ICME*, 2000.

[5] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, “Analysis of emotion recognition using facial expressions, speech and multimodal information,” in *ICMI*, State College, PA, October 2004, pp. 205–211, ACM Press.

[6] A. Metallinou, S. Lee, and S. Narayanan, “Audio-visual emotion recognition using gaussian mixture models for face and voice,” in *In proc of the IEEE ISM*, 2008.

[7] D. Ruta and B. Gabrys, “An overview of classifier fusion methods,” 2000.

[8] T. Serre Y. Ivanov and J. Bouvrie., “Error weighted classifier combination for multi-modal human identification.,” Tech. Rep., MIT, Cambridge, MA, 2005.

[9] C. Busso, M. Bulut, C-C Lee, A.Kazemzadeh, E. Mower, S. Kim, J. Chang, S.Lee, and S.Narayanan, “IEMOCAP: interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[10] A. Metallinou, C.Busso, S. Lee, and S. Narayanan, “Visual emotion recognition using compact facial representations and viseme information,” Submitted to ICASSP10, 2009.

[11] I. Cohen, Q. T. Xiang, S. Zhou, X. Sean, Z. Thomas, and T. S. Huang, “Feature selection using principal feature analysis,” 2002.

[12] S. Lee and D. Yook, “Audio-to-visual conversion using hidden markov models,” in *Proc. 7th Pacific Rim Int. Conf. on Artificial Intelligence*, 2002.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Cambridge Research Laboratory, Cambridge, England, 2006.

[14] C. Busso, S. Lee, and S.S. Narayanan, “Using neutral speech models for emotional speech analysis,” in *Interspeech 2007*, 2007.

[15] J.R. Deller, J.H.L. Hansen, and J.G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, 2000.

[16] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley and Sons, 2001.

[17] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, “Interpreting ambiguous emotional expressions,” in *In proc of ACII Special Session*, 2009.

[18] K. P. Kording and D. M. Wolpert, “Bayesian decision theory in sensorimotor control,” *TRENDS in Cognitive Sciences*, vol. 10, pp. 319–326, 2006.