Audio-Visual Emotion Recognition using Gaussian Mixture Models for Face and Voice

Angeliki Metallinou, Sungbok Lee and Shrikanth Narayanan
School of Electrical Engineering
University of Southern California
Los Angeles, CA 90089-2560
metallin@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

Emotion expression associated with human communication is known to be a multimodal process. In this work, we investigate the way that emotional information is conveyed by facial and vocal modalities, and how these modalities can be effectively combined to achieve improved emotion recognition accuracy. In particular, the behaviors of different facial regions are studied in detail. We analyze an emotion database recorded from ten speakers (five female, five male), which contains speech and facial marker data. Each individual modality is modeled by Gaussian Mixture Models (GMMs). Multiple modalities are combined using two different methods: a Bayesian classifier weighting scheme and support vector machines that use post classification accuracies as features. Individual modality recognition performances indicate that anger and sadness have comparable accuracies for facial and vocal modalities, while happiness seems to be more accurately transmitted by facial expressions than voice. The neutral state has the lowest performance, possibly due to the vague definition of neutrality. Cheek regions achieve better emotion recognition accuracy compared to other facial regions. Moreover, classifier combination leads to significantly higher performance, which confirms that training detailed single modality classifiers and combining them at a later stage is an effective approach.

1. Introduction

The expression of emotions and the recognition of a person's affective state are abilities indispensable for natural human interaction and social integration. The study of emotions has attracted interest of researchers from very diverse areas, ranging from psychology [24],[21] to the applied sciences. The term affective computing was introduced by Picard [20] and the questions related to the recognition and processing of multimodal affective information have been widely investigated and discussed [17]. Human communication is multimodal with specific channels operating and interacting dynamically across time. In fact, Mehrabian has stated that the semantic contents of a message contribute only 7% of the overall impression while the vocal part and the facial expression contribute 38% and 55% respectively [15]. Human expression is communicated through various channels, producing complementary and redundant information that is used to resolve problems when one of the modalities is not properly transmitted (e.g., speech in noisy environment)[18]. Therefore, a joint analysis of facial expression and speech is hypothesized to achieve better performance and be more robust than the use of single modalities.

Facial expression has been widely studied in terms of how the human face is perceived [22] and how facial expressions can be quantified. The most widely used method for measuring facial behaviors is the Facial Action Coding System (FACS) which was introduced by Ekman and Friesen [7]. FACS decomposes all possible facial expressions into Action Units(AUs)[8]. Automatic detection of AUs from video input is described in [1],[6],[12], where authors apply machine learning techniques for data driven facial expression classification. A rule based system for automatic AU recognition is described in [19]. Other approaches for facial expression modeling and recognition include the use of Active Appearance Models, which are nonlinear parametric models that can be fitted into an input video image [14]. Researchers have also used face images with markers to minimize the noise introduced by automatic facial feature detection [3],[4]. Specifically, in [3] a single subject study is presented where facial and vocal information are used for emotion classification.

In the current work, we investigate to what extent facial and vocal modalities can be used separately and in com-

978-0-7695-3454-1/08 $25.00 © 2008 IEEE
DOI 10.1109/ISM.2008.40

bination for emotion recognition. This study is based on the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) which was recorded from multiple speakers of both genders. Although it was recorded from actors, the elicitation techniques that were used are expected to produce more genuine realizations of real life emotions than past acted databases [2], [5]. The IEMOCAP database contains detailed facial marker information from ten speakers. Although, the use of face markers is not a realistic option for real time systems, this study provides insight about which facial regions convey more expressive information. Such facial regions should be tracked with greater precision by an automatic facial feature extraction system.

In this study, the face is separated into six regions and Gaussian Mixture Models (GMMs) are trained for each examined emotion both for the facial and vocal modalities. Hidden Markov Models (HMMs) and Gaussian Mixture Models(GMMs) have been widely used for modeling audio and visual information streams [13], [16]. As a first step, we examine the discriminative power of each information frame for emotion recognition; therefore GMM modeling is appropriate for the present study. Emotion recognition performance is analyzed and compared across modalities for various emotions. Classifier combination is addressed and different combination methods are compared. Bayesian methods provide general and mathematically rigorous techniques for combining streams of information and it is argued that human behavior for cue combination is close to what is predicted by Bayesian Decision Theory [11]. The Bayesian approach for classifier combination presented in [10] and [9] is utilized in this study. This method is compared to a more ad-hoc method for facial and vocal combination, which utilizes a support vector classifier (SVC).

The contribution of this study is that it extends the analysis presented in [3] to multiple speakers of both genders. We also study in detail the behaviors of different facial regions and of the vocal modality during emotion expression. Individual classifier results indicate that anger and sadness have comparable performance for facial and vocal modalities. Happiness is found to be more accurately transmitted through facial expressions than through voice. Cheek regions achieve better emotion recognition accuracies compared to other facial regions. Combined classifier results indicate that the combination of multiple modalities can lead to acceptable recognition accuracies for anger, happiness and sadness, in the range of 65%-80%. However, the expression of the neutral emotional state is highly variable and therefore difficult to recognize. Furthermore, training detailed single-modality classifiers and combining them at a later stage seems to be a more effective approach than training one general classifier.

2 Methodology

2.1 The IEMOCAP database

For the analysis of this paper, we used the Interactive Emotional Dyadic Motion Capture database (IEMOCAP), which was recently collected at USC [2]. This database contains approximately 12 hours of audiovisual data from ten different actors, of both genders. In this database, the actors were asked to perform selected emotional scripts and to improvise following specific hypothetical scenarios. The actors were recorded in dyadic sessions in order to facilitate a more natural interaction and expression of the targeted emotion. The data were emotionally evaluated by three evaluators, who tagged each sentence with the most appropriate emotional tag (e.g., happiness, sadness), according to the overall audiovisual impression of the sentence. Majority voting was used to decide the final emotional label of the sentence. The current analysis examines sentences which are classified in four emotional states; angry, happy, neutral and sad. In classification experiments we use the facial and vocal information that is available for each sentence.

2.2 Face Gaussian Mixture Models for emotion recognition

Markers were placed on the faces of the actors to collect the spatial information of these markers for each video frame. The positions of the markers can be seen in figure 1, represented by points on the face. After capturing the marker data, markers were normalized so that head rotation is canceled out and so that the nose marker becomes the local coordinate center of each frame [2].

In the current analysis we examine and compare the effectiveness of different facial regions in discriminating between different emotional states. For this purpose, facial markers are separated into six blocks, each of which defines a different facial region; forehead (FH), right eyebrow (RE), left eyebrow (LE), right cheek (RC), left cheek (LC) and chin (CH). These facial regions can be seen in Figure 1 and they are approximately consistent with the analysis of [3] for single speaker data. Because of their limited movement the markers belonging to the nose region are not included in this analysis. Moreover, a GMM for the total face (TOTAL) is trained using marker data from all facial regions. Neighboring markers are averaged in order to reduce the total number of markers. The choice of which markers to average is ad-hoc, however, the markers that are averaged belong to the same facial muscles and their movements are correlated.

A Gaussian Mixture Model (GMMs) is trained for each of the emotional states that we examine; angry (ANG), happy (HAP), neutral (NEU) and sad (SAD). All available

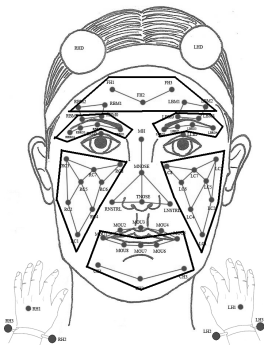


Figure 1. Face markers and face regions: Six distinct groups are defined corresponding to forehead, left and right eyebrows, left and right cheeks and chin.

sentences belonging to the above mentioned emotions are selected from the IEMOCAP database. These sentences come from ten different speakers and belong either to an improvisation session or a scripted session. The sentences for each emotion are randomly split into two equal sets, one of which was used for model training, the other for testing.

The marker point coordinates are used as features for the training of the Gaussian mixture models. The frame rate of the markers is 8.3 ms. We preprocess the marker positions by averaging them over a window of six frames, with an overlap between consecutive windows of three frames(40 fps). The feature vector for each facial region consist of the 3-D coordinates of the markers belonging to that region plus their first and second derivatives. For the computation of the frame-level derivatives we use the following formula [23].

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}$$

where d_t is a delta coefficient at time t. This is computed in terms of the corresponding static coefficients $c_{t-\theta}$, $c_{t+\theta}$ at times $t-\theta$, $t+\theta$, where θ is the time step. In our work, the static coefficients c_t are the 3-D marker coordinates.

Concerning the GMM training process, diagonal covariance matrices were chosen so as to reduce the number of parameters that we train. Various number of mixtures were experimentally examined. For the experiments presented here, we choose GMMs with 64 mixtures because it was found empirically that this number of mixtures achieves good performance for the training dataset.

2.3 Voice GMM models for emotion recognition

The Mel Frequency Cepstral Coefficients (MFCCs) are used for vocal analysis. The feature vector consists of 12 MFCCs and energy, and their first and second derivatives; a 39-dimensional feature vector. The window length for the MFCC extraction is 50ms and the overlap is 25ms, to match the window of the facial data extraction (40fps).

Similarly to the face analysis, we train a GMM for each of the examined emotions. Because most of the available sentences contain portions of silence, during which only low background noise can be heard, we train an extra GMM to model background noise. This model is trained using the word alignment which is available for each sentence (forced alignment) [2]. Therefore, we train 5 models using the voice data of the train set. We also select diagonal covariance matrices and choose models with 32 gaussian mixtures.

2.4 Combination of the individual classifier outputs

The GMMs for each separate modality give us a limited picture of the overall emotional impression of a sentence. Our goal is to effectively combine the information of individual modalities in order to infer the overall emotional content of the sentence. In this work, we explore two different classifier combination techniques. The first is a Bayesian approach for multiple cue combination. The second is an ad-hoc method which utilizes support vector machines that use post classification accuracies as features.

The Bayesian approach is described in [10]. It uses the conditional error distribution. This distribution can be approximated from the confusion matrix of each separate classifier. The confusion matrices are used to weight the output of each classifier in order to make a combined decision.

Given a problem with K classes and C different classifiers, we want to classify an observation into one of the available classes by combining the decisions of the individual classifiers. Each observation x can be written as $x = \{x_1, x_2, \ldots, x_C\}$, where x_i is the subset of the features of x used by classifier λ_i, $i = 1, \ldots, C$. For each observation x, based on the C classifier outputs, we want to infer a class label ω belonging to one of the K classes. Assuming that subsets x_i are disjoint, we deduce that for all classifiers λ_i: $P(\omega|x, \lambda_i) = P(\omega|x_i, \lambda_i)$, therefore:

$$P(\omega|x) = \sum_{i=1}^{C} P(\omega, \lambda_i|x) = \sum_{i=1}^{C} P(\omega|\lambda_i, x)P(\lambda_i|x) \quad (1)$$

Assuming that for each classifier λ_i we have a predicted class label $\widetilde{\omega_k}$, then the true class label can be derived as follows:

$$P(\omega|\lambda_i, x) = \sum_{k=1}^{K} P(\omega, \widetilde{\omega_k}|\lambda_i, x)$$
$$= \sum_{k=1}^{K} P(\omega|\widetilde{\omega_k}, \lambda_i, x) P(\widetilde{\omega_k}|\lambda_i, x) \quad (2)$$

where $P(\widetilde{\omega_k}|\lambda_i, x)$ is the prediction of the classifier λ_i. The authors in [10] make the assumption that $P(\omega|\widetilde{\omega_k}, \lambda_i, x) \approx P(\omega|\widetilde{\omega_k}, \lambda_i)$, because the latter probability can be simply obtained from the confusion matrix of the corresponding classifier. With this assumption, after combining Equations 1 and 2 we arrive at the equation:

$$P(\omega|x) = \sum_{i=1}^{C} \sum_{k=1}^{K} P(\omega|\widetilde{\omega_k}, \lambda_i) P(\widetilde{\omega_k}|\lambda_i, x) P(\lambda_i|x) \quad (3)$$

$P(\omega|\widetilde{\omega_k}, \lambda_i)$ is the probability of label ω given that classifier λ_i has decided the label ω_k. This can be approximated from the confusion matrix of classifier λ_i. Furthermore, $P(\lambda_i|x)$ is a weight assigned to each classifier representing the confidence of the decision of the classifier in question. This weight can also be computed from the confusion matrix of classifier λ_i.

In our classification problem, we have K=4 classes (angry, happy, neutral and sad). There are six facial region classifiers and one voice classifier, that is C = 7 classifiers. Each classifier consists of four GMMs, one for each emotion, except for the vocal classifier. There are five GMMs for the vocal classifier due to the extra noise GMM that is trained on the noise (non-speech) regions of the sentence. For each observation (frame) we compute $P(x|\widetilde{\omega_k}, \lambda_i)$; the probability that observation x is generated by GMM of emotion ω_k from classifier λ_i. Then the probability $P(\widetilde{\omega_k}|\lambda_i, x)$ from equation 3 can be computed using Bayes rule as follows:

$$P(\widetilde{\omega_k}|\lambda_i, x) = \frac{P(x|\widetilde{\omega_k}, \lambda_i) P(\widetilde{\omega_k}, \lambda_i)}{P(x, \lambda_i)}$$
$$= \frac{P(x|\widetilde{\omega_k}, \lambda_i) P(\widetilde{\omega_k}|\lambda_i) P(\lambda_i)}{P(x|\lambda_i) P(\lambda_i)}$$
$$= \frac{P(x|\widetilde{\omega_k}, \lambda_i) P(\widetilde{\omega_k}|\lambda_i)}{P(x|\lambda_i)} \quad (4)$$

where $P(x|\widetilde{\omega_k}, \lambda_i)$ is known and $P(x|\lambda_i)$ is common to all classes and can therefore be omitted. $P(\widetilde{\omega_k}|\lambda_i)$ is the probability that a classifier λ_i assigns a label ω_k. This probability can be approximated from the confusion matrix of λ_i.

Using this framework, we can compute the probability $P(\omega|x)$ for each frame x combining the decisions of each individual classifier. We classify each frame to the emotion class with the highest probability. Finally, each sentence is classified to an emotion class, according to the majority vote of the labels of the frames.

In the second approach, we add one more classification step in order to combine the various model outputs and decide the emotional content of the input sentence. For each input sentence we have information from six facial regions and from one vocal modality. For each modality we have four possible decisions; namely angry, happy, neutral and sad. Therefore for each modality the GMMs produce the percentage of the sentence frames belonging to each of the possible emotions. The classification percentages convey each classifier's confidence for the assigned tags for each sentence. Therefore, these percentages themselves may be used as features for emotion classification. We use the classification percentages of each available modality to construct a feature vector. This feature vector is the input of a final classifier, which decides on the overall emotional tag of the sentence, based on the local classification decisions of the various available modalities.

In practice, we found that the support vector classifier (SVC) with radial basis kernel achieves good results. SVC is a good choice because it can be trained with high-dimensional feature vectors and relatively few data samples without serious overtraining problems. For the SVC experiments, we split the available test data into two equal sets, one of which is used to train the support vector classifier and the other to test it. We perform ten cross validation experiments on the available data and report the classification confusion matrix, as well as the mean error rate of the classifier.

3 Results

Table 1 lists and explains all the abbreviations that are used in the experimental section.

Table 1. List and explanation of abbreviations.

Abbreviations	
ANG	angry sentence
HAP	happy sentence
NEU	neutral sentence
SAD	sad sentence
SVC	support vector classifier
FH	forehead GMM model
CH	chin GMM model
RC	right cheek GMM model
LC	left cheek GMM model
RE	right eyebrow GMM model
LE	left eyebrow GMM model
TOTAL	total face GMM model

3.1 Individual Classifier Results

We collect all the available sentences belonging to the four emotional states that we examine; angry, happy, neutral and sad. During all of these sentences the actor is speaking, at least for some part of the sentence. In the IEMOCAP database in total, there are 620 sentences which are tagged as angry, 314 as happy, 604 as neutral, and 652 as sad. They vary in length from few tenths of a second to 3-4 minutes. This corpus is randomly split into two disjoint equal sets to be used as train and test set.

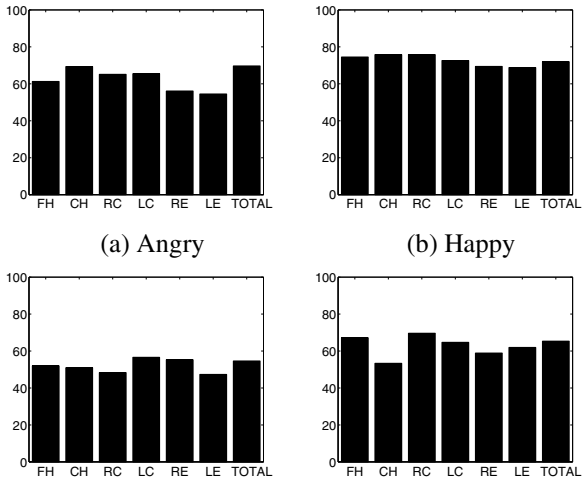
The classification accuracy for each face region GMM as well as for the total face GMM is presented in Table 2. As mentioned in the Section 2, all GMMs have diagonal covariance matrices and 64 mixtures. Accuracy is computed per sentence and majority rule is used for the classification of each sentence. In Figure 2 we present the classification accuracy of different face models for each emotion across speakers.

Table 2. Emotion classification accuracy for face models for speaker data.

MODEL	ANG	HAP	NEU	SAD
FH	61.29	74.52	51.99	67.18
CH	69.35	75.8	50.99	53.37
RC	65.16	75.8	48.34	69.63
LC	65.48	72.61	56.62	64.72
RE	56.13	69.43	55.3	58.9
LE	54.52	68.79	47.35	61.96
TOTAL	69.68	71.97	54.64	65.34

Anger and happiness have better recognition accuracies compared to emotional states with lower levels of activation, such as sadness and neutrality. For each specific emotion we observe that the cheek regions (RC and LC), the forehead (FH), and the chin (CH), have better overall performance compared to the eyebrow regions (RE and LE). Moreover, individual face models perform comparably to the total face model (TOTAL).

The classification accuracy for the voice GMM model can be seen in Table 3. Accuracy is computed per sentence using only the frames that belong to speech regions, according to the word alignment. Majority rule is used for the classification of each sentence. Note, however, in this case we have one extra noise model, since a given frame could be classified as noise. The noise model was included to model parts of a sentence where the speaker makes a pause. This partially explains the lower classification accuracy that is achieved by the voice classifier, compared to the face classifiers.



(a) Angry (b) Happy

(c) Neutral (b) Sad

Figure 2. Emotion classification accuracy of different face models for each emotion for speaker data

Table 3. Emotion classification accuracy for voice model for speaker data.

MODEL	ANG	HAP	NEU	SAD
VOICE	76.45	19.75	50.0	71.17

3.2 Combined Classifier Results

We present the experimental results for the two different approaches of classifier combination presented in section 2.

For the data available in the test set, we combine the classifiers using the Bayesian approach. The probability $P(\omega|x)$, where ω is the emotion tag, is computed for each frame x and for all possible emotions; angry, happy, neutral and sad. The emotion with the highest probability is selected. Therefore, each frame is classified to an emotion class using information from all available classifiers. The total sentence is classified using the majority voting. The weights for the combination are computed from the confusion matrices of the individual classifiers for train dataset. The results per sentence can be seen in table 4.

This method gives classification accuracies in the range of 65%-80% for emotional states, using a general database(speaker and gender independent). However, results for the neutral state seem to be problematic.

A support vector classifier (SVC) is used to combine the separate face and voice model decisions. The classification percentages for each emotion and for each individual

Table 4. Combination of 6 face regions and voice for speaker data: Confusion matrix for Bayesian approach of classifier combination

	ANG	HAP	NEU	SAD
ANG	64.60	5.57	17.70	12.13
HAP	7.10	79.35	9.03	4.52
NEU	11.15	10.81	51.69	26.35
SAD	8.95	4.63	3.70	82.72

model are concatenated into a feature vector of 28 dimensions (four emotions times six face models and four emotions for the voice model). The remaining sentences which do not belong to the GMM train set are randomly split into two equal sets, one of which is used to train the SVC and the other to test it. Therefore, 50% of the available data is used to train the GMMs, 25% is used to train the SVC and 25% to test the SVC. This cross validation experiment is repeated 10 times and the mean error rate and the standard deviation of the error rate are presented in table 5. Also table 5 shows the confusion matrix.

Table 5. Combination of 6 face regions and voice for speaker data: Confusion matrix for SVC classifier and the corresponding classifier performance.

SVC	ANG	HAP	NEU	SAD
ANG	82.5	3.88	10.69	2.94
HAP	5.11	73.19	13.09	8.60
NEU	10.17	6.42	63.61	19.80
SAD	4.83	2.80	9.85	82.52

mean error rate	standard deviation
24.02	1.62

The results obtained by this method are comparable to the results of the Bayesian classifier. Anger and sadness have the highest accuracy while neutrality has the lowest accuracy.

4 Discussion

The results presented in Section 3, enable us to explore the roles of the different facial regions and the vocal channel in the expression of emotions.

Individual performance of the facial regions is presented in Table 2 across emotions. Anger and happiness have better recognition accuracies compared to emotional states with lower levels of activation, such as sadness and neutrality. This result is consistent for all face model accuracies. The neutral state has the lowest classification accuracies, possibly because the definition of neutrality is vague and may include very diverse facial expressions. Moreover, if we think of the different face regions as multiple information channels used to transmit emotion information, it is possible that only some of them are used to give an emotional impression while the others remain neutral. This observation may account for the low accuracies of the neutral state.

For each specific emotion we observe that the cheek regions, the forehead and the chin, have better overall performance compared to the eyebrow regions. We note that if marker data are not available, robust feature extraction from cheek regions is a difficult task. On the other hand, eyebrow regions which are more suitable for automatic feature extraction generally have lower performance. This result agrees with the analysis presented in [3] for single speaker data. Moreover, individual face models perform comparably to the total face model (TOTAL) which indicates that even individual face regions have significant discriminative power for emotion recognition. We should note, however, that the total face model is trained using averaged marker data in order to reduce the dimensionality of the feature vector. This may have lead to some loss of information and may have impaired the classifier performance.

For the voice classifier anger and sadness are the best performing emotions, while neutrality has classification accuracy on the order of 50% (Table 3). The accuracy achieved for happiness is low and could be attributed, to some extent, to the actors' expression choices, as many of the sentences that are tagged as happy have a smiling face but a neutral voice. Therefore the voice model classifies these sentences as neutral. The happy state includes diverse expressions of happiness such as joy and contentment. Contentment is usually expressed using a happy face and calm voice. This observation along with the fact that happiness has the highest recognition accuracy when we use face modalities, might indicate that the happy emotion is accurately transmitted by the face; the voice channel may not have significant discriminative power for this emotion. Finally, another reason for the low recognition of happiness using the vocal model may be that MFCC features are not appropriate for this problem, and other contextual features may be required.

As a general comment, we note that the experimental data contain the expression of emotions that are designed to be close to a natural and realistic emotional expression and not a caricature. The emotions that fall under a general emotional tag may be very diverse and are in context with the script or the improvisation scenario that is being acted. Moreover, not all modalities are equally active during a sentence that is attributed to an emotional category, and some

modalities may transmit conflicting information.

The emotional sentences used in our analysis were selected according to their tags, which were attributed based on the global audiovisual evaluation of a sentence by human evaluators. We make an assumption that the global tag is representative of the emotions expressed in the facial and vocal modalities individually. However, during the natural expression of an emotion, this assumption may not always hold. Humans may use conflicting audio and visual impressions in order to communicate the targeted emotion. One example is sarcasm, where the speaker's face may look happy while his voice sounds angry. In this case, most people will recognize the overall emotion as anger even if the audio and visual modalities are conflicting. However, a model trained only on facial data will most probably detect happiness instead of anger, thus failing to recognize the overall emotion. Such conditions may limit the performance of single modality classifiers and should be handled by appropriate classification and information fusion strategies.

The Bayesian classifier combination technique achieves classification accuracies on the order of 80% for happiness and sadness and of 65% for anger (Table 4). However, the performance for neutral state is low. One of the reasons for the low accuracy of neutral state is the weight computation technique. The confusion matrices for all single modalities indicate that the neutral state is very often misclassified. Therefore, the weight attributed to the decision that a frame is neutral is relatively low, so this decision is not reliable. Consequently, the overall classifier is less likely to classify a given frame to the neutral state. This explains why the neutral sentences are often misclassified with all other emotions. On the other hand, the performance for happiness is not affected by the low performance of the voice modality because the weight computation approach gives very low weight (low reliability) to the voice classifier for recognizing the happy emotional state.

The results for the support vector machine fusion technique follow a similar trend (Table 5). Anger and sadness have accuracies of approximately 80%. The accuracy for happiness is slightly lower, possibly because of the bad performance of the voice modality. Neutral state has the lowest accuracy (about 60%).

Both fusion approaches have comparable performance. However, the Bayesian method is more mathematically grounded and combines the classifiers in a way that is easier to explain and is more intuitive. The happiness case exemplifies how the Bayesian approach can effectively combine multiple channels even when some of them (here the voice channel) are less effective. On the other hand, when all channels have relatively low performance at recognizing a specific class, like in the neutral case, the overall classifier is very likely to misclassify an observation belonging to that class. In this case, the simpler support vector classification method achieves better recognition accuracy.

The combined accuracies are significantly higher than the single modality performances and the performance of the total face GMM (TOTAL). We achieve classification accuracy on the order of 65-80% for emotional states, on a large speaker and gender independent database. This indicates that training detailed single-modality classifiers and then combining them might be a more appealing approach than training one general classifier. Multiple modality combination can mitigate each individual classifier's weaknesses and can make overall classification more robust. However the effective combination of multimodal cues in order to achieve good recognition accuracies for all available classes is a challenging problem.

5 Conclusion and Future Directions

The goal of the present study was to explore the role of the different facial regions, and the vocal channel, in the expression of emotions during human interactions. The study provided a systematic investigation of the discriminative power of the unimodal and multimodal classifiers for recognition of four emotional states; anger, happiness, sadness and the neutral state. We used a large database (IEMOCAP) with multiple speakers of both genders, where the aforementioned emotions are expressed in a interactive and natural manner. The use of direct facial marker data enabled us to overcome some of the present challenges in feature processing from video data, and focus on establishing feasibility bounds for classification using visual features for the challenging emotion recognition problem.

Individual modality recognition performances indicate that anger and sadness have comparable accuracies for facial and vocal modalities, while happiness is found to be more accurately transmitted through facial expressions than through voice. Recognition accuracies for neutral state were low, possibly because of the very diverse face and voice expressions which are contained in the neutral class. Cheek, forehead and chin regions generally have ample discriminative power for emotion recognition. For the combination of multiple modalities, a Bayesian approach and a support vector classifier approach were used. The combined accuracies were significantly higher than the accuracies obtained by individual modalities and by a general total face classifier.

There are potential directions for future work that can build on the foundations of this study. First, each regional classifier is trained separately, irrespective of its neighbors. However, it is evident that voice affects the face modalities, especially the lower face regions, and that face regions affect each other. The face and voice modalities interplay across time for the expression of an emotion. Future work

will include a more detailed modeling of the spatial and temporal relations of the face and voice modalities. Moreover, the genuine expression of human emotions is difficult not only to model but also to label and the simple categorical tags used may not be appropriate to describe the emotion expressed. In the future, we will examine whether tagging using attribute based evaluation such as valence, activation and dominance is more suitable for these type of experiments.

Acknowledgment

This work was funded in part by NSF and the Department of the Army. The authors would like to thank Carlos Busso for his useful comments and advice.

References

[1] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal Of Multimedia*, 1(6):22–35, September 2006.

[2] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, In press, 2008.

[3] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Sixth International Conference on Multimodal Interfaces ICMI 2004*, pages 205–211, State College, PA, October 2004. ACM Press.

[4] C. Busso and S. Narayanan. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2331–2347, November 2007.

[5] C. Busso and S. Narayanan. Recording audio-visual emotional databases from actors: a closer look. In *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008.

[6] J. Cohn, L. Reed, Z. Ambadar, J. Xiao, and T. Moriyama. Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *IEEE International Conference on Systems, Man and Cybernetics*, 2004.

[7] P. Ekman and W. Friesen. Facial action coding system (FACS): Manual. *Palo Alto: Consulting Psychologists Press*, 1978.

[8] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System (FACS)*. Research Nexus division of Network Information Research Corporation, 2002.

[9] Y. Ivanov, B. Heisele, and T. Serre. Using component features for face recognition. In *International Conference on Automatic Face and Gesture Recognition,Seoul, Korea,*, May 2004.

[10] Y. Ivanov, T. Serre, and J. Bouvrie. Error weighted classifier combination for multi-modal human identification. Technical Report CBCL Paper 258, Massachusetts Institute of Technology, Cambridge, MA, 2005.

[11] K. P. Kording and D. M. Wolpert. Bayesian decision theory in sensorimotor control. *TRENDS in Cognitive Sciences*, 10(7):319–326, July 2006.

[12] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. In *Conference on Computer Vision and Pattern Recognition Workshop*, 2004.

[13] J. Luettin, G. Potamianos, and C. Neti. Asynchronous stream modeling for large vocabulary audio-visual speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 2001.

[14] I. Matthews and S. Baker. Active appearance models revisited. Technical Report CMU-RI-TR-03-02, Robotics Institute, Carnegie Mellon University, April 2003.

[15] A. Mehrabian. Communication without words. *Psychology Today*, 2:53–56, 1968.

[16] M. Ostendorf, V. Digalakis, and O. Kimball. From HMM'S to segment models: A unified view of stochastic modelingfor speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4:360–378, September 1996.

[17] M. Pantic and L. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, September 2003.

[18] M. Pantic and L. J. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 22(12):1424–1445, December 2000.

[19] M. Pantic and L. J. M. Rothkrantz. An expert system for recognition of facial actions and their intensity. In *Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, 2000.

[20] R. W. Picard. Affective computing. Technical Report 321, MIT Media Laboratory Perceptual Computing Section, Cambridge, MA,USA, November 1995.

[21] J. Rottenberg and J. Gross. Emotion and emotion regulation: A map for psychotherapy researchers. *Clinical Psychology: Science and Practice*, 14(4):323–328, December 2007.

[22] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. In *Proceedings of the IEEE*, volume 94, November 2006.

[23] S.J.Young, J.Jansen, J.J.Odell, D.G.Ollason, and P.C.Woodland. *The HTK Book*. Cambridge University Engineering Department and Entrophic Cambridge Research Laboratory, Cambridge, UK., 1995.

[24] D. M. Sloan and A. M. Kring. Measuring changes in emotion during psychotherapy: Conceptual and methodological issues. *Clinical Psychology: Science and Practice*, 14(4):307–322, December 2007.