Bimodal Information Analysis for Emotion Recognition

Malika Meghjani, Frank Ferrie and Gregory Dudek†
McGill University,
Department of Electrical and Computer Engineering and School of Computer Science †
{malika,ferrie,dudek}@cim.mcgill.ca

Abstract

We present a bimodal information analysis system for automatic emotion recognition. Our approach is based on the analysis of video sequences which combines facial expressions observed visually with acoustic features to automatically recognize five universal emotion classes: Anger, Disgust, Happiness, Sadness and Surprise. We address the challenges posed during the temporal analysis of the bimodal data and introduce a novel technique for combining the best features of instantaneous and temporal based visual recognition systems. We obtain robust appearance-based visual features which we classify instantaneously and aggregate it temporally to improve the recognition rates when compared to single-frame based instantaneous classification. The performance of the system is further boosted by using the complementary audio information for the bimodal emotion recognition. We combine the two modalities at both feature and score level to compare the respective joint emotion recognition rates. The emotions are instantaneously classified using a Support Vector Machine and sequentially aggregated based on their classification probabilities. This approach is validated on a posed audio-visual database and a natural interactive database. The experiments performed on these databases provide encouraging results with the best combined recognition rate being 82%.

1. Introduction

We propose an audio-visual information analysis system specifically for automatic bimodal emotion recognition. Some of the applications of automatic emotion recognition are in domains such as, tele-health care monitoring, tele-teaching assistance, gaming, automobile driver alertness monitoring, stress detection, lie detection and user personality type detection. A comprehensive list of these applications is reviewed in [1].

A bimodal emotion recognition system (combining audio and visual analysis systems) can broadly be classified in two categories based on the methods used for analysis and training of the visual component of the system [2]. These are instantaneous and temporal based recognition systems. The instantaneous systems are trained using appearance-based features from static images and the temporal systems are trained by tracking specific facial features over a sequence of frames. These two systems have their corresponding limitations. The temporal systems require accurate tracking of facial feature points along with a method for associating their temporal patterns with the respective emotion classes. Whereas, the instantaneous systems require the selection of key frames which can substantially summarize the expressed emotion over a period of time.

Therefore, based on the above observations, we combine the best aspects of the two systems by extracting and classifying the robust appearance based visual features at each instant, and temporally aggregating their classification probabilities to improve the recognition rates. The objective of our work is to prove the hypothesis that temporal inference and data fusion techniques, together, improve the recognition rates significantly when compared to instantaneous single mode analysis. In addition, we also discuss our semi-supervised method of selecting the useful frames from the visual sequences to ease the training process of the instantaneous visual emotion recognition system.

2. Related work

One of the first approaches of integrating audio-visual information for automatic emotion recognition was proposed by De Silva et al. [3]. They studied the human subjects' ability to identify the universal emotion classes in order to derive a weighting function for audio and visual modalities respectively. The important conclusions of their findings suggested the complementary nature of the audio and the visual data with emotions like 'Anger', 'Happiness' and 'Surprise' being better recognized based on the visual analysis, and emotions like 'Sadness' and 'Fear' being easily detected using the audio features. Based on these findings, the idea of automatic bimodal emotion recognition was later adapted by many

978-1-4244-5498-3/09/$25.00 ©2009 IEEE

researchers, of which some of the relevant contributions are acknowledged in this section.

In another work, De Silva et al. [4] implemented an independent audio-based emotion recognition system using pitch as an audio feature vector and performed the classification using a nearest neighbor algorithm. They also proposed a visual system which tracked the facial points based on optical flow information and classified it using a Hidden Markov Model (HMM). The two systems were combined using a rule based fusion technique which gave an average recognition rate of 72%. Zeng et al. [5] suggested an approach for integrating the facial texture with pitch and intensity features using Adaboost learning scheme to fuse the feature vectors from the two modalities which are then classified using a HMM. They reported an average recognition rate of 89% using a two subject, person dependent database consisting of only two emotion classes.

One of the challenging tasks of the visual tracking systems addressed above is to deal with changes in the shape of the mouth caused due to speech. In order to deal with this situation, Dragos et al. [6] proposed a data fusion technique where they rely only on the visual data in the silent phase of the video sequence and the fused audio-visual data during non-silent segments. The visual modality during non-silent segments only focused on the upper half of the facial region to eliminate the effects caused by changes in the shape of the mouth. In a similar work, Song et al. [7] proposed an approach for multimodal emotion recognition which was specifically focused on temporal analysis of three sets of features: 'audio only features', 'visual only features' (upper half of facial region) and 'visual speech features' (lower half of facial region) using a triple HMM, i.e. one HMM for each of the information modes. This model was proposed to deal with state asynchrony of the audio-visual features while maintaining the original correlation of these features over time. The person dependent recognition rate for their system was 85%.

Yongjin et al. [8] proposed a relatively inexpensive computational method for visual based emotion recognition which selected a single key frame from each audio-visual sequence to represent the emotion present in the entire sequence. The criterion for selecting the key frames from the audio-visual sequences was based on the heuristic that peak emotions are displayed at the maximum audio intensities. The visual features are extracted from these key frames using Gabor wavelets whose spatial dimensionality is reduced by considering only five statistical features (mean, median, min, max and standard deviation) measured over each of the Gabor filters. The audio component of the system extracts features related to pitch and intensity along with Mel-frequency Cepstral Co-efficient (MFCC) and formant frequencies. The audio-visual modes are combined at the feature level, and the features are selected using a stepwise method according to which each feature is selected or rejected based on the criteria that it maximizes the between-class Mahalanobis distance. These features were classified using a multi-class SVM with an average recognition rate of 82%.

3. Proposed approach

The outline of our proposed approach for automatic bimodal emotion recognition is given in Figure 1. We train our visual system based on static peak emotions present in the key representative frame of the audio-visual sequence. Each of the audio-visual sequences contain one of the five universal emotions: 'Anger', 'Disgust', 'Happiness', 'Sadness' and 'Surprise'. We select the key frames from the audio-visual sequences using a semi-supervised clustering technique. This method automatically divides the audio-visual frames in two major clusters; namely, the emotion frames and the non-emotion frames. The cluster containing the highest number of continuous frames is considered to be the emotion frame cluster, since the audio-visual sequences are pre-segmented into five emotion classes with each sequence containing the majority of the frames from the corresponding emotion class. We select the frame which is at the minimum distance from the centre of this cluster for training our visual system. In order to deal with the outliers, e.g. only one frame in the cluster, we apply a re-clustering technique where we remove the outlier frames and perform the clustering process iteratively till we obtain a minimum pre-defined number of continuous frames in each of the clusters.

The audio component of our system is trained based on global statistics of features obtained from the speech signal. The information obtained from the two modalities is combined using feature and score level fusion techniques. The details of each processing step are discussed in the following sections.

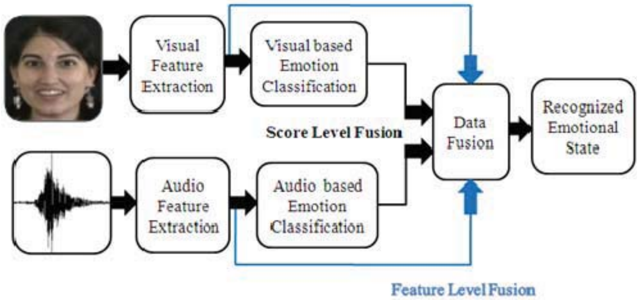
Visual Feature Extraction
Visual based Emotion Classification
Score Level Fusion
Data Fusion
Recognized Emotional State
Audio Feature Extraction
Audio based Emotion Classification
Feature Level Fusion

Figure 1: Bimodal Emotion Recognition System based on Score and Feature Level Fusion.

3.1. Acoustic feature analysis

The audio information analysis for emotion recognition generally comprises linguistic and paralinguistic measurements [2]. The linguistic measurement conforms to the rules of the language whereas paralinguistic measurement is the meta-data; i.e. related to how the words are spoken based on variations of pitch, intensity and spectral properties of the audio signal. In our implementation, we only extract the paralinguistic features for audio based emotion recognition since it can be generalized to any language database.

Table 1: List of Acoustic Features

Feat. No.	Feat. Description	Feat. No.	Feat. Description
1	Pitch relative maximum position	11	Intensity standard deviation
2	Pitch standard deviation	12	Intensity mean fall time
3	Pitch mean absolute slope	13	Intensity mean rise time
4	Pitch mean	14	Intensity mean
5	Pitch maximum	15	Signal mean
6	Pitch range	16	Speech rate
7	Pitch relative minimum position	17	Unvoiced mean duration
8	Pitch minimum	18	Intensity maximum
9	Voiced mean duration	19	Spectral energy below 650 Hz.
10	Spectral energy below 250 Hz.	20	Intensity relative maximum position

We derive a list of global statistics from the paralinguistic features (pitch, intensity and spectral properties) as presented in Table 1. In addition to these features, we also evaluate temporal features like the speech rate and the MFCCs which highlight the dynamic variations of the speech signal. We obtain five statistical features derived from the first 13 MFC coefficients (minimum, maximum, mean, median and standard deviation) which result in a total of 85 audio features for each audio-visual sequence. The advantage of using the global statistical features for audio based emotion recognition is that, it provides the same number of features for a variable length input speech signal. These features have also been known to outperform the continuous dynamic features for audio based emotion recognition [9].

The feature extraction process involves the pre-processing of the audio signal by removing the leading and trailing silent edges of the signal. The pitch and intensity contours of the audio signal are obtained using PRAAT [10], the speech analysis software. Once, the features are extracted they are normalized separately for each subject before performing the classification using a multi-class SVM.

3.2. Visual feature analysis

The emotion cues from the visual information channel are obtained by analyzing the facial expressions of the subjects in the scene. The facial expression recognition is performed by detecting upright frontal faces in the video frames using the Adaboost face detection algorithm [18] with modified implementation of the code available in OpenCV [11]. These detected facial regions are reshaped to equal sizes (32x32) and spatially sampled using a bank of 20 Gabor filters. The advantage of using the Gabor filters for feature extraction in the present context is that they preserve the local spatial relations between facial features and eliminate the need for explicitly tracking each facial point. The cost of this feature extraction method is the high dimensionality of the feature vector obtained at each instance of time (32x32x20). The following section discusses the method of spatially reducing this feature vector to improve the recognition rates and also speed up the required computations. On the other hand, in order to temporally reduce the dimensionality of the data we select a key visual frame from each sequence based on the clustering method described in the previous section. The complete visual analysis process is represented in Figure 2.

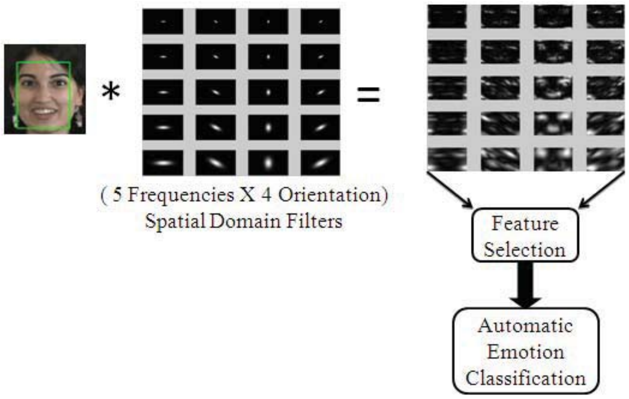
*
=
(5 Frequencies X 4 Orientation)
Spatial Domain Filters
Feature Selection
Automatic Emotion Classification

Figure 2: Visual analysis using bank of 20 Gabor filters (5 spatial frequency and 4 orientations).

3.3. Feature reduction

The high dimensional feature vectors obtained from the two modalities are reduced using a feature reduction method. We apply the Recursive Feature Elimination (RFE) method as implemented in [12] to obtain a minimum subset of the most discriminative feature set. The RFE method, iteratively removes the input features using a ranking criterion. The ranking criterion is based on the weights obtained from a classifier like SVM. The features with minimum weights are eliminated and the iterative process is continued until we obtain an optimal number of features which provide the best cross-validation results. The optimal number of features to be used for classification is decided based on the maximum accuracy of 5 folds cross-validation results obtained from the training set. The intervals used for testing the optimal number of features are chosen in powers of 2 as the number of features eliminated at each time step is half the total number of features. The best cross-validation accuracy obtained for the visual features is 93% for selecting 1024 visual features from a total of 20480 features and 67% for selecting 21 audio features out of 85 features.

3.4. Classification

The features selected from the above process are classified using a Support Vector Machine (SVM). The SVMs are binary classifiers which can be extended for classifying more than two classes using two techniques: (a) one-against-rest and (b) one-against-one classification. We use the one-against-one technique to classify five emotion classes. This method compares pair-wise classes which results in C_2^n combinations of the classifier. The final classification result is based on maximum-wins voting scheme.

The output of the SVM classification is given in terms of the decision values which indicate the distance of the test examples from the discriminating plane. These decision values can be converted into probability estimates using the method described in [13]. This estimate provides the probability distribution of the test data over all the emotion classes. The conversion of decision values to a probability distribution is obtained by using the following formulation:

$$p(q_i|x_i) = g(f(x_i), A, B) = \frac{1}{1 + \exp(Af(x_i) + B)} \quad (1)$$

where,

$q \rightarrow Emotion\ Class$,
$x \rightarrow Input\ Feature\ Vector$,
$f(x) \rightarrow Decision\ Function$,
$A, B \rightarrow Unknown\ Parameters\ to\ be\ evaluated\ using\ maximum\ likelihood\ between\ the\ training\ data\ and\ their\ decision\ values.$

The probability estimates thus obtained are used as weights for combining the audio and visual modes for the score level fusion.

3.5. Fusion technique

We combine the audio and visual data by using two fusion techniques: (a) feature level fusion and (b) score level fusion. In the feature level fusion we obtain a key representative visual frame from each of the test sequences based on the heuristic used in [8], i.e. we select maximum audio intensity frames which represent the peak emotions and extract the visual features from these frames only. The visual feature vector of the key frame is concatenated with the global audio feature vector to fuse the information at the feature level. This extremely large feature vector is reduced by using the RFE method, and classified using the SVM.

The second method for combining the two modalities is based on the scores or the probability estimates obtained after individually classifying the two modalities using the multi-class SVM. The score for the visual system is obtained by temporally aggregating the weighted scores from an interval of frames in the visual sequence. The weights and the interval of frames for the temporal aggregation are decided based on two criteria: (a) maximum audio intensity and (b) minimum entropy of the probability distribution. The outline of the score level fusion is presented in Figure 3. The results obtained for the two fusion techniques are discussed next.

```
% VISUAL ANALYSIS
Cluster (Emotion/Non-Emotion Frames): C1,C2

while ((C1 < pre-defined number)) && (C2 < pre-defined number))
    Remove outlier frames
    Re-cluster
end

for (k = 1:n_frames)
    Set counters for continuous number of frames in each cluster
    Bubble sort maximum number of continuous frames belonging to
    one cluster
end

Pick minimum distance frame from continuous longest cluster
Train visual SVM
Classify all frames in test sequence
Obtain visual scores for all the frames in the sequence
Temporally aggregate test frames using sum rule
    Select the interval for temporal aggregation using
    (a) Maximum audio intensity (or)
    (b) Minimum entropy of probability distribution

% AUDIO ANALYSIS
Train audio SVM
Classify global statistical features of test sequences
Obtain audio scores

% FUSION
Score level fusion using product rule

Maximum score class is considered as final decision for emotion
recognition
```

Figure 3: Algorithm for score level fusion

4. Experimental results

We evaluate our approach on two types of databases. The first database [14] we use is made up of posed audio-visual sequences in a lab environment with all subjects facing the camera. The second database is a subset of the 'Belfast Naturalistic Database' [15] which consists of natural conversations between participants and interviewer with unconstrained lighting and head movements. A sample of the two databases is shown in Figure 4.

The posed database consists of 9 subjects expressing 5 emotions in 5 different forms. We test our approach on each subject for all emotions which provides a total number of 45 audio-visual test sequences and 180 training sequences. We perform experiments using the feature and the score level fusion techniques. We further refine the system's performance by temporally aggregating the scores of the frames around the maximum audio intensity and minimum entropy frames. The results of these experiments are presented in Table 2.

We performed the above set of experiments based on the training set selected manually with peak emotions and compared them with the results obtained using the semi-supervised training data. The recognition rates obtained using the semi-supervised training data performed surprisingly better than the manually selected training set. A comparison of the recognition rates obtained using the two training sets are also summarized in Table 2. The average audio based emotion recognition rate for this database is 53%.

The second set of experiments was performed on the natural database. The number of examples per emotion class in this database is not uniformly distributed. Hence, we used a combination of image databases [16], [17] for training our visual system. These image databases have instances of subjects expressing the emotions with the peak intensity. We train the system using these peak

Figure 4(a): Posed audio-visual database selected from 'eNTERFACE 2005' database.

Figure 4(b): Spontaneous audio-visual sequences selected from 'Belfast Naturalistic Database'.

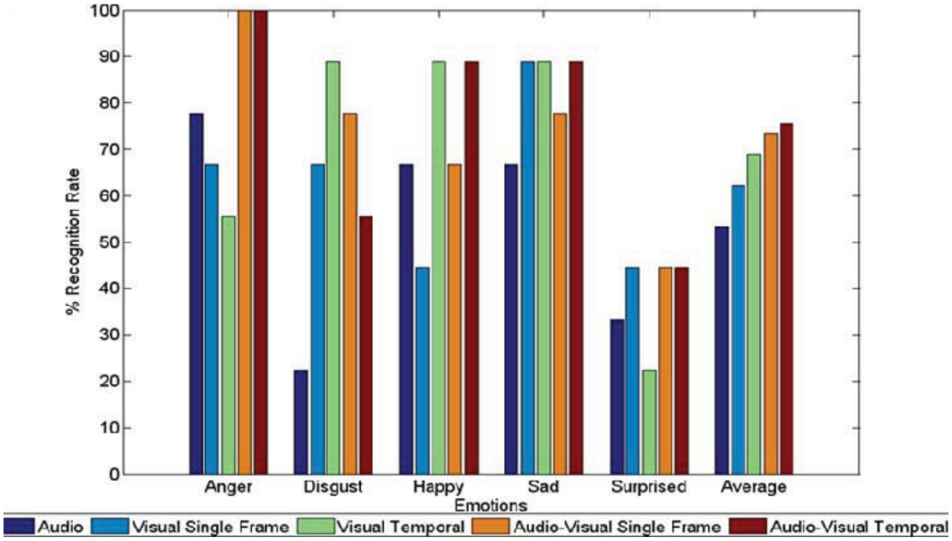

Figure 5: Manual Training based Recognition Rates

Table 2: Recognition rates (%) for posed audio-visual database

Fusion Technique	Instantaneous Maximum Audio		Temporal Maximum Audio		Instantaneous Minimum Entropy		Temporal Minimum Entropy	
Training Process	Manual	Semi-Auto	Manual	Semi-Auto	Manual	Semi-Auto	Manual	Semi-Auto
Visual	62	73	69	78	67	78	76	82
Audio-Visual (Feature Level)	67	80	-	-	-	-	-	-
Audio-Visual (Score Level)	73	82	76	82	67	78	78	82

intensity fames for identifying two emotions: 'Happy' and 'Sad' in the natural sequence database. We used the audio streams from [14] for training our audio-based emotion recognition system. We achieve 83% of combined audio-visual accuracy at the score level for two-class person-independent emotion recognition. The individual visual and audio recognition rates were 50% and 67% respectively.

5. Conclusion

The proposed algorithm and the results presented in the previous sections confirm our assertion that temporal aggregation of the scores for the visual data increases the recognition rates by a maximum of 5% when compared to the single frame based visual classification. The recognition rate is also improved by combining the audio modality using the score level fusion by a maximum of 10%. For the score level fusion, it is observed that the minimum entropy criteria used for weighting the visual scores performs better than the maximum audio intensity criteria which was suggested in [8]. It can also be inferred from Figure 5 that the score level fusion always performed better than feature level fusion. Finally, the emotions: 'Anger' and 'Happy' were better recognized from the audio modality, whereas, the visual modality performed better at: 'Disgust', 'Sad' and 'Surprised' emotions.

6. Future work

The motivation for this work is derived from the Tele-Health care application where we would like to automatically analyze the emotional states of a patient in response to the type of interventions provided by a nurse practitioner. The automatic analysis thus obtained can be used to provide an offline feedback to the nurse in order to improve the nursing intervention protocols. It can also be useful in automatic generation of interventions learned from the patient-nurse interactive conversations.

As a part of our future work, we also intend to automatically analyze interactive conversations between two subjects in order to adaptively learn the combinations of the emotion states of the patients in response to the interventions provided by the nurse.

7. References

[1] R. W. Picard, "Affective Computing". *MIT Press*, 1997.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor. "Emotion Recognition in Human-Computer Interaction". *Signal Processing Magazine,* volume 18, pages 32-80, January 2001.

[3] L.C. De Silva, T. Miyasato and R. Nakatsu. "Facial Emotion Recognition using Multi-Modal Information". *Proceedings of International Conference on Information, Communications and Signal Processing,* 1997.

[4] L.C. De Silva and P.C. Ng. "Bimodal emotion recognition". *In Proceedings of 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition,* March 2000.

[5] Z. Zeng, Y. Hu, G.I. Roisman, Z. Wen, Y. Fu, and T.S. Huang, "Audio-visual Emotion Recognition in Adult Attachment Interview". *International Conference on Multimodal Interfaces,* 2006.

[6] D. Datcu and L.J.M. Rothkrantz. "Semantic Audio-Visual Data Fusion for Automatic Emotion Recognition". *Euromedia'2008*, April 2008.

[7] M. Song, C. Chen and M. You. "Audio-Visual based Emotion Recognition using Tripled Hidden Markov Model". *IEEE International Conference on Acoustics, Speech, and Signal Processing,* volume 5, May 2004.

[8] Y. Wang and L. Guan, "Recognizing Human Emotional State from Audio-Visual Signals". *IEEE Transactions on Multimedia*, volume 10, June 2008.

[9] B. Schuller, G. Rigoll and M. Lang. "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture". *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, May 2004.

[10] Boersma, Paul & Weenink, David (2009). Praat: doing phonetics by computer (Version 5.1.12) [Computer program]. Retrieved August 4, 2009, from http://www.praat.org/

[11] Intel, "Open CV: Open source Computer Vision Library", http://www.intel.com/research/mrl/research/opencv/.

[12] Guyon, J. Weston, S. Barnhill and V. Vapnik. "Gene selection for cancer classification using support vector machines". *In Machine Learning, Springer*, Volume 46, no. 1-3, pp. 389-422. 2002.

[13] J. Platt. "Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods". *In A. Smola, P. Bartlett, B. Schölkopf and D. Schuurmans (Eds.), Advances in large margin classifiers. Cambridge: MIT Press,* 2000.

[14] O. Martin, J. Adell, A. Huerta, I. Kotsia, A. Savran and R. Sebbe. "Multimodal Caricatural Mirror". *eINTERFACE'05-Summer Workshop on Multimodal Interfaces,* 2005.

[15] E Douglas-Cowie, R Cowie and M Schröder. "A New Emotion Database: Considerations, Sources and Scope". *Tutorial and Research Workshop (ITRW) on Speech and Emotion,* 2000.

[16] M. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba. "Coding Facial Expressions with Gabor Wavelets". *In Proceedings, IEEE International Conference on Automatic Face and Gesture Recognition*, April 1998.

[17] T. Kanade, J.F. Cohn and Y. Tian. "Comprehensive database for facial expression analysis". *In Proceedings of, IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46-53, 2000.

[18] P. Viola and M.J. Jones, "Robust real-time face detection". *International Journal of Computer Vision*, pages 137- 154, 2004.