Bimodal Information Analysis for Emotion Recognition

Malika Meghjani, Frank P. Ferrie and Gregory Dudek
Centre of Intelligent Machines, McGill University, Montreal, Quebec, Canada

Workshop on Applications of Computer Vision (WACV) - 2009

1. Introduction

Automatic emotion recognition is widely used for applications such as tele-health care monitoring, tele-teaching assistance, gaming, automobile driver alertness monitoring, stress detection, lie detection and user personality type detection.

In this work, we implemented a bimodal system for the study of voice patterns and facial expressions of human subjects to recognize five emotions: 'Anger', 'Disgust', 'Happiness', 'Sadness' and 'Surprise'.

The objective of our work is to create video summary of the audio-visual data by labeling the emotions present in a given sequence.

Figure 1: A Sample of emotions present in video sequences obtained from 'eNTERFACE 2005' and 'Belfast Naturalistic' databases respectively.

2. System Overview

Our bimodal emotion recognition system consist of three major components: audio analysis, visual analysis and data fusion (see Figure 2).

The audio component of our system is trained based on global statistics of features obtained from the speech signal.

The visual analysis is based on static peak emotions present in the key representative frame of the audio-visual sequences. The key frames are selected using a semi-supervised clustering algorithm.

The information obtained from the two modalities is combined using feature and score level fusion techniques.

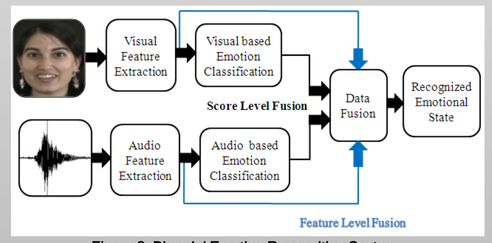

Figure 2: Bimodal Emotion Recognition System

3. Audio Feature Extraction

The audio-based emotion recognition is obtained by extracting paralinguistic features which is related to how the words are spoken based on variations of pitch, intensity and spectral properties of the audio signal.

A list of global statistical features is derived from the paralinguistic features (pitch, intensity and spectral properties) along with temporal features like the speech rate and the MFCCs which highlight the dynamic variations of the speech signal.

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.

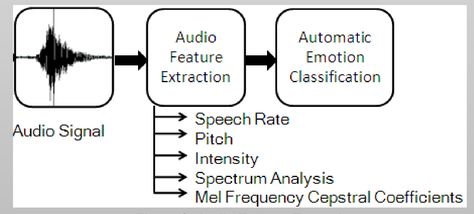

Figure 3: Audio Feature Extraction

4. Visual Feature Extraction

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 [1].

These detected facial regions are reshaped to equal sizes and spatially sampled using a bank of 20 Gabor filters (5 spatial frequencies and 4 orientations).

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.

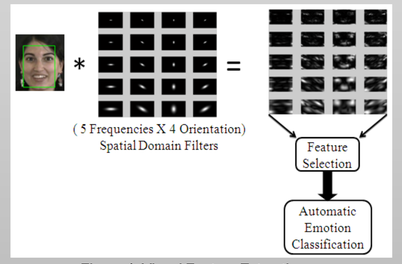

Figure 4: Visual Feature Extraction

5. Feature Reduction and Classification

The high dimensional feature vectors obtained from the two modalities are reduced using Recursive Feature Elimination (RFE) technique [2].

RFE iteratively removes the input features using a ranking criterion which 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 features selected from the above process are classified using SVM. The decision values obtained as the output of the SVM classifier is converted into probability estimates using the following formulation:

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

where,

$q \rightarrow Emotion\ Class$

$x \rightarrow Input\ Feature\ Vector$

$f(x) \rightarrow Decision\ Function$

$A, B \rightarrow Unknown\ Parameters\ to\ be\ evaluated\ using\ maximum\ likelihood\ between\ the\ training\ data\ and\ their\ decision\ values.$

6. Fusion Technique

We use two fusion techniques: (a) feature level fusion and (b) score level fusion.

In the feature level fusion a key representative visual frame from each of the test sequence is concatenated with the global audio feature vector for bimodal emotion recognition. The outline of the score level fusion technique is presented in Figure 5.

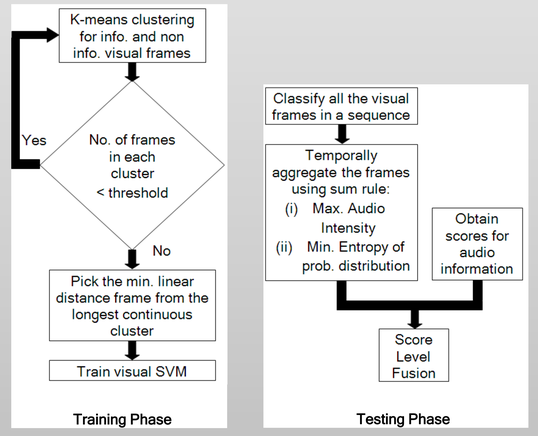

Figure 5: Algorithm for Score level fusion

7. Experimental Results

We evaluate our approach on two types of databases: posed [3] and natural [4], as illustrated in Figure 1. Results from score and feature level fusion techniques on the posed database are presented in Table 1.

A comparison of the recognition rates obtained using the semi-supervised and manually selected training sets are also summarized in the table below. The average audio-based emotion recognition rate for this database is 53%.

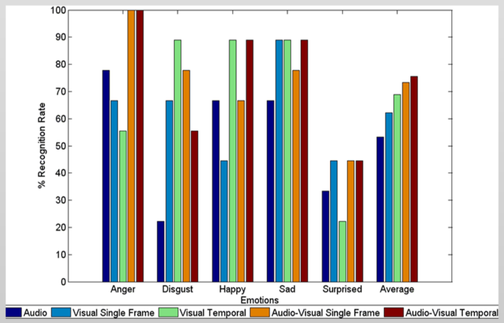

Figure 6: Manual Training based Recognition Rates

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

Table 1: Recognition rates (%) for posed audio-visual database

8. Conclusion

The results presented in the previous section suggest 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 and visual modalities by a maximum of 10% using the score level fusion technique.

9. References

[1] Viola et al. "Robust real-time face detection". *International Journal of Computer Vision* 2004.

[2] Guyon et al. "Gene selection for cancer classification using support vector machines". *Journal of Machine Learning* , 2002.

[3] O. Martin et al. "Multimodal Caricatural Mirror". *eINTERFACE'05-Summer Workshop on Multimodal Interfaces, 2005.*

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