Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information

Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, Shrikanth Narayanan

Emotion Research Group, Speech Analysis and Interpretation Lab
Integrated Media Systems Center, Department of Electrical Engineering, Department of Computer Science
Viterbi School of Engineering, University of Southern California, Los Angeles
http://sail.usc.edu

Abstract

The interaction between human beings and computers will be more natural if computers are able to perceive and respond to human non-verbal communication such as emotions. Although several approaches have been proposed to recognize human emotions based on facial expressions or speech, relatively limited work has been done to fuse these two, and other, modalities to improve the accuracy and robustness of the emotion recognition system. This paper analyzes the strengths and the limitations of systems based only on facial expressions or acoustic information. It also discusses two approaches used to fuse these two modalities: decision level and feature level integration. Using a database recorded from an actress, four emotions were classified: sadness, anger, happiness, and neutral state. By the use of markers on her face, detailed facial motions were captured with motion capture, in conjunction with simultaneous speech recordings. The results reveal that the system based on facial expression gave better performance than the system based on just acoustic information for the emotions considered. Results also show the complementarity of the two modalities and that when these two modalities are fused, the performance and the robustness of the emotion recognition system improve measurably.

Introduction

Why do we need to recognize emotions?

- Emotions are an important element of human-human interaction.
- Design improved human-machine interfaces able to give specific and appropriate help to user based on emotional state assessment.

How can we recognize emotions from human communication cues?

- From speech, facial expression, gesture, head movement, *etc.*
- Computer algorithms can use same inputs.

Why is it necessary to use a multimodal approach?

- Modalities give complementary information [Chen, 98]. Some emotions are better recognized by speech (sadness) while others by facial expression (anger and happiness)[De Silva, 97].
- Better performance and more robustness [Pantic, 03].

Previous Work

- Decision-level [Chen,98][De Silva,00] and feature-level fusion systems [Chen,99][Huang,98].

Purpose of this project

- Quantify the performance of unimodal systems to recognize emotion states, find the strengths and weaknesses of these approaches and compare different approaches to fuse these dissimilar modalities to increase the overall recognition rate of the system.

Methodology

Database

- Four emotions – sadness, happiness, anger and neutral state – are targeted, single subject.
- Facial motion and speech are simultaneously captured. A VICON motion capture system with three cameras was used to capture the expressive facial motion data with 120Hz sampling frequency (102 markers). The recording was made in a quiet room using a close talking SHURE microphone at the sampling rate of 48 kHz.
- Phoneme balanced corpus (258 sentences).

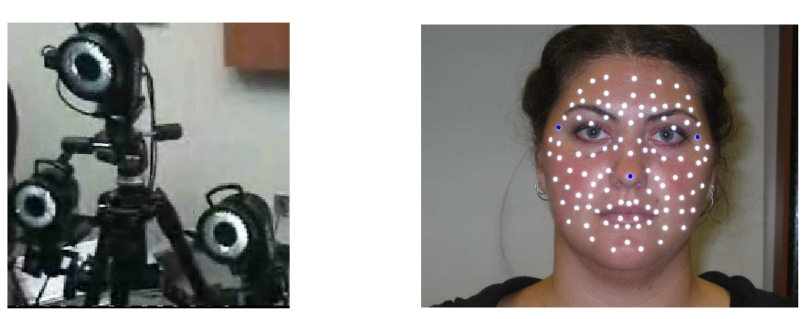


Figure 1: Data recording system

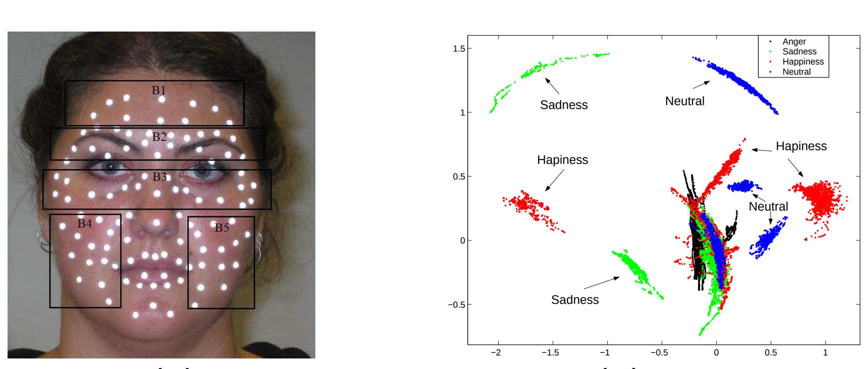
Three different systems based on speech, facial expression and bimodal information, respectively, were implemented using Support Vector Machine classifier (SVC) with 2nd order polynomial kernel functions. The database was trained and tested using the leave-one-out cross validation method.

Features from Speech

- Global-level prosodic features: Pitch and energy statistic (mean, median, std, max, min and range); and, Voiced speech and Unvoiced speech ratio.
- Sequential backward features selection (11-D feature vector).

Features from Facial Expression

- A 4-D feature vector at utterance level is extracted
  1. Data is normalized to remove head motion
  2. Five facial areas are defined
  3. 3-D coordinates are concatenated
  4. PCA is used to reduce to 10-D vector per frame and per area
  5. The points are clustered (K-nearest neighbor)
  6. The statistic of the frames at utterance level is used as 4-D feature vector



(a) (b)

Figure 2: a) Five facial areas considered in this study b) First two PCA components of low eye area

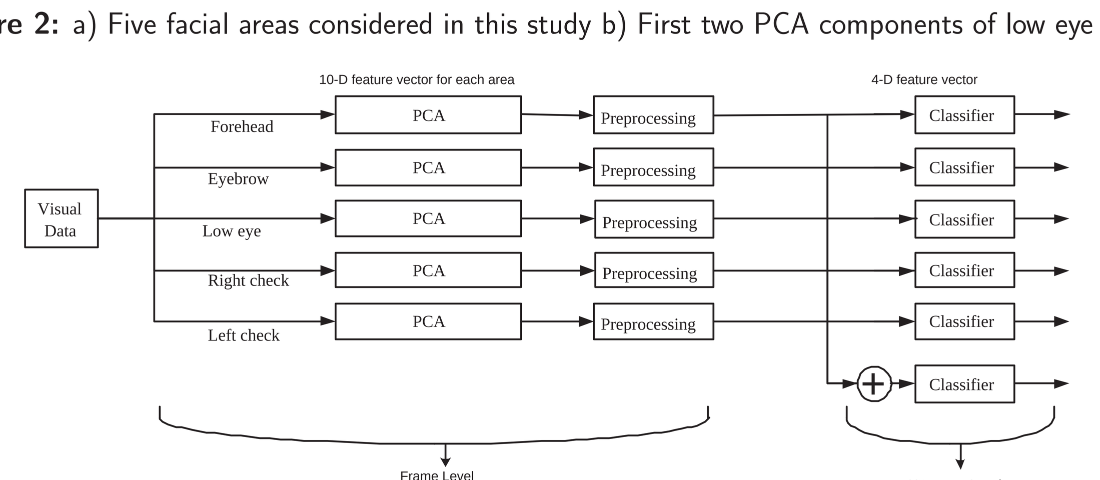
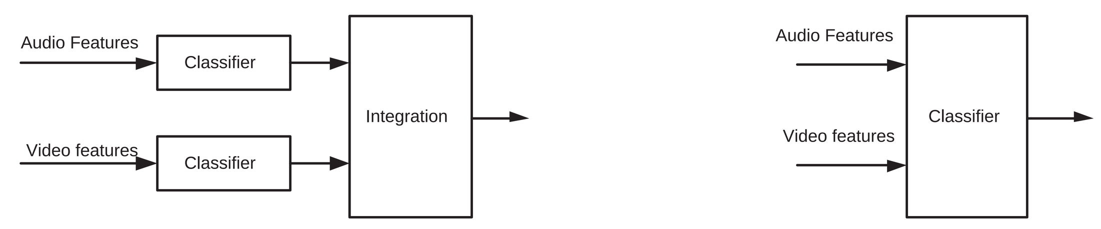


Figure 3: System based on facial expression

Multimodal techniques

- Decision-level integration.
  - Maximum, Average, Product and Weight of the posterior probabilities.
- Feature-level integration.
  - Sequential backward feature selection (10-D feature vector).



(a) (b)

Figure 4: a) Decision-level fusion b) Features-level fusion

Results

Tables 1 and 2 show the confusion matrix of the unimodal emotion recognition systems.

- The overall performance of the classifiers based on speech and facial motions were 70.9% and 85.1%, respectively.
- In the acoustic domain, sadness-anger and neutral-happiness can be separated with high accuracy. However, happiness-anger and sadness-neutral are mutually confused.
- In the facial expression domain, anger-happiness can be accurately separated. However, anger-sadness and neutral-happiness are confused.
- Note that sadness-neutral are confused in both domains, so it is expected that the recognition rate of sadness in the feature-level bimodal classifier will be poor. Other discriminating information such as contextual cues are needed.

Table 1: Emotio Recognition from Speech (70.9%)

	Anger	Sadness	Happiness	Neutral
Anger	0.68	0.05	0.21	0.05
Sadness	0.07	0.64	0.06	0.22
Happiness	0.19	0.04	0.70	0.08
Neutral	0.04	0.14	0.01	0.81

Table 2: Emotio Recognition from Facial Motion (85.1%)

	Anger	Sadness	Happiness	Neutral
Anger	0.79	0.18	0.00	0.03
Sadness	0.06	0.81	0.00	0.13
Happiness	0.00	0.00	1.00	0.00
Neutral	0.00	0.04	0.15	0.81

The table 3 shows the performance of the bimodal system at decision-level with different fusing criteria. In the weight-combining rule, the modalities are weighted according to rules extracted from the confusion matrices of unimodal classifiers. This table shows that the product-combining rule gives the best performance.

Table 3: Decision-level integration bimodal classifier with different fusing criteria

	Overal	Anger	Sadness	Happiness	Neutral
Maximum combining	0.84	0.82	0.81	0.92	0.81
Averaging combining	0.88	0.84	0.84	1.00	0.84
Product combining	0.89	0.84	0.90	0.98	0.84
Weight combining	0.86	0.89	0.75	1.00	0.81

Tables 4 and 5 show the confusion matrix of the decision and feature level bimodal classifiers.

- Feature-level integration (89.1%).
  - High performance of anger, happiness and neutral. Bad performance of sadness (79%).
  - The performance of happiness significantly decreased to 91 percent.
- Decision-level integration with product-combining rule (89.0%).
  - Although the overall results are similar, the confusion matrices show important differences.
  - The recognition rate of each emotion increased compared to unimodal systems (except happiness)
  - Sadness is recognized with high accuracy (90%).

Table 4: Feature-level bimodal classified

	Anger	Sadness	Happiness	Neutral
Anger	0.95	0.00	0.03	0.03
Sadness	0.00	0.79	0.03	0.18
Happiness	0.02	0.00	0.91	0.08
Neutral	0.01	0.05	0.02	0.92

Table 5: Decision-level bimodal classifier (product-combinig rule)

	Anger	Sadness	Happiness	Neutral
Anger	0.84	0.08	0.00	0.08
Sadness	0.00	0.90	0.00	0.10
Happiness	0.00	0.00	0.98	0.02
Neutral	0.00	0.02	0.14	0.84

Discussion and Summary

- The multimodal systems give 5% improvement (absolute) compared to unimodal systems.
- Some pair of emotions confused in one modality are easily separated in the other modality.
  - Sadness-anger can be separated in the acoustic domain, and neutral-happiness and anger-happiness can be separate in the facial expression domain.
- Sadness and neutral are confused in both domains, because their features are similar.
  - Feature-level integration systems cannot separate them accurately.
  - Decision-level integration systems maybe (in our experiments, yes).
- Feature and decision-level integration systems give similar overall results, but analysis in detail show differences.
- Although the system based on speech has worse performance than the system based on facial expression, the acoustic features provide valuable information about emotions.
  - Note that visual features were directly obtained from marker tracking and not video: feature extraction from video may introduce challenges.
  - Although the use of facial markers are not suitable for real applications, the analysis presented in this paper give important clues about emotion discrimination.
- Redundant information provided by modalities can be used to improve the performance of the emotion recognition system when the features of one of the modal are inaccurately acquired (e.g. beard, mustache, eyeglasses and noise).

Limitation of this work

- Marker based visual data for a single speaker.
- Global features (no dynamic information is used).
- Standard fusion approaches.

Future Work

- Collect more emotional data from other speakers.
- Use visual algorithms to extract facial expression features from video.
- Use segmental level information to trace the emotions at a frame level.
- Find better methods to fuse audio-visual information that model the dynamics of facial expressions and speech.