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. 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 some 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 (302 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 (251 sentences).

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

• 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 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

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

Multimodal techniques

• Decision-level integration.
  – Maximum, Average, Product and Weight of probability abilities.
  – Feature-level integration.
    – Sequential backward feature selection (10-D feature vector).

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.

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 separated in the facial expression domain.
  – Sadness and neutral are confused in both domains, because their features are similar.
  – Decision-level integration systems cannot separate them accurately.
  – Decision-level integration systems may (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: features 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.