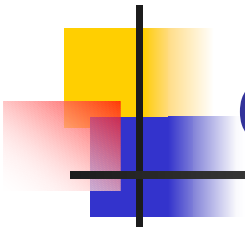
Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information

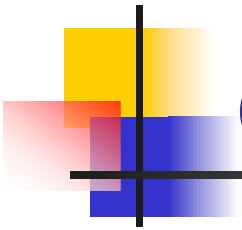
C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan

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



Outline

- Overview of Emotion Recognition
- Methodology
- Experiments and Results
- Discussion and Future Work



Overview: Emotion Recognition

Why do we need to recognize emotions?

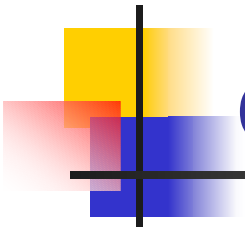
- Emotions are an important element of human-human interaction
- Design improved human-machine interfaces
- Give specific and appropriate help to user based on emotional state assessment

How can we recognize emotions from human communication cues?

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

Why is it necessary to use a multimodal approach?

- Modalities give complementary information [Chen98]
- Some emotions are better recognized by speech (sadness and fear) while others by facial expression (anger and happiness) [DeSilva97]
- Better performance and more robustness [Pantic03]



Overview: Emotion Recognition (2)

Previous Work

- Decision-level fusion systems (rule-based system) [Chen98] [DeSilva00] [Yoshitomi00]
- Feature-level fusion systems [Chen98_2] [Huang98]

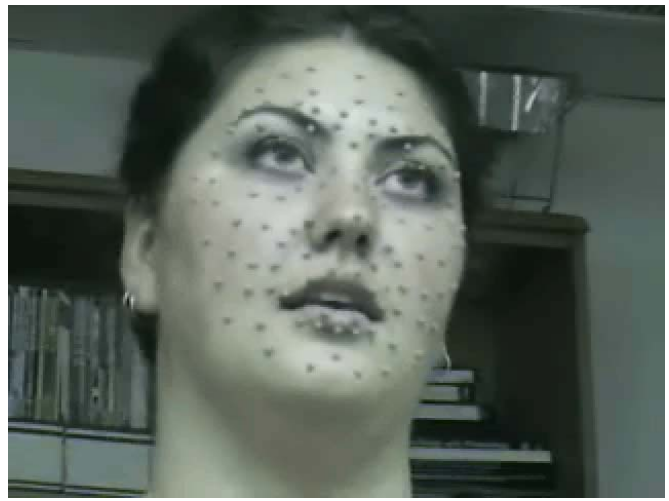
Purpose of this project

- Analyze the strength and limitation of unimodal systems to recognize emotion states
- Evaluate two fusion approaches, in terms of the performance of the system

Methodology

Database

- Four emotions are targeted, single subject
  - Neutral state
  - Anger
  - Sadness
  - Happiness



- 102 Markers to track facial expressions

- Facial motion and speech are simultaneously captured

- Phoneme balanced corpus (258 sentences)

Methodology (2)

Recognition Systems

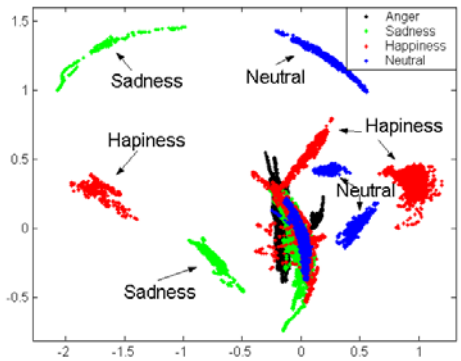
- Three recognition systems based on:
  - Speech features
  - Facial motion features
  - Multimodal data features
- Support vector machine classifier (SVC) is used in all the systems

Features from Speech

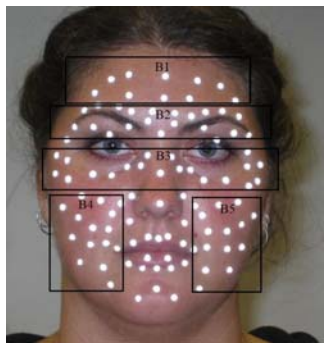
- Global-level prosodic features
  - Pitch and energy statistic (mean, median, std, max, min and range)
  - Voiced speech and Unvoiced speech ratio
- Sequential backward features selection (11-D feature vector)

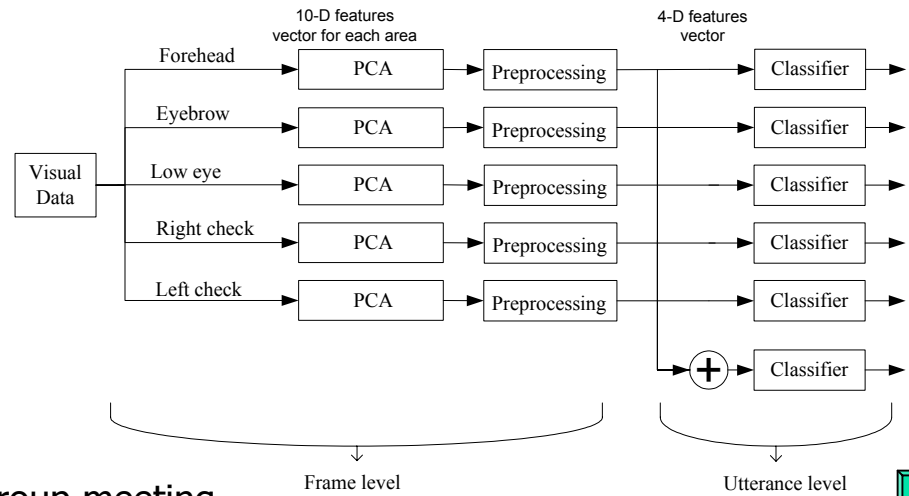
Methodology (3)

Features from Facial Expression



- 4-D feature vector
  - Data is normalized to remove head motion
  - Five facial areas are defined
  - 3-D coordinates are concatenated
  - PCA is used to reduce to 10-D vector
  - The points are clustered (K-nearest neighbor)
  - The statistic of the frame at utterance level is used as 4-D feature vector (A1)

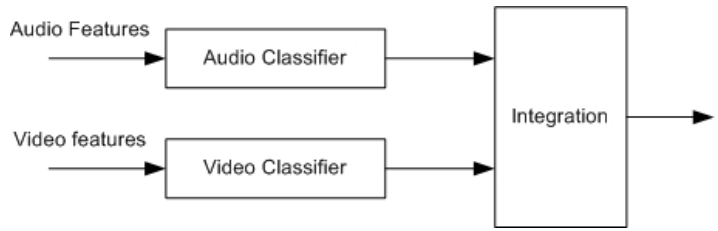




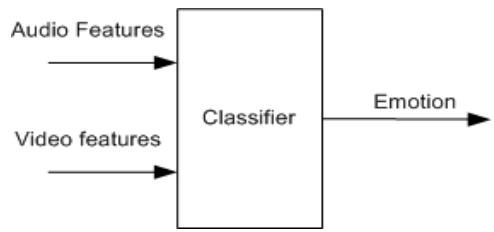
Methodology (4)

Multimodal techniques

- Decision-level integration
  - Maximum of the posterior probabilities
  - Average of the posterior probabilities
  - Product of the posterior probabilities
  - Weight of the posterior probabilities



- Feature-level integration
  - Sequential backward feature selection (10-D feature vector)



Experiments and Results

Unimodal Systems

- Emotion recognition system based on speech (70.9%)
  - Confusion sadness-neutral 22%
  - Confusion neutral-sadness 14%
  - Confusion happiness-anger 19%
  - Confusion anger-happiness 21% (A2)
  - Neutral-happiness and anger-sadness are well separated

	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

- Emotion recognition system based on facial expression (85.1%) (A3)
  - Happiness is recognized with high precision (M2)
  - Confusion anger-sadness 18%
  - Confusion sadness-neutral 15%
  - Confusion neutral-happiness 15%
  - Anger-happiness are well separated

	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

Experiments and Results (2)

Multimodal Systems

- Feature-level integration (89%)
  - High performance of anger, happiness and neutral state
  - Bad performance of sadness 79%
  - The performance of happiness decreased

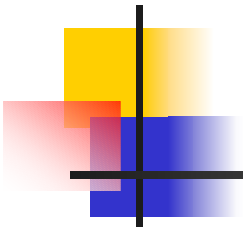
	Anger	Sadness	Happiness	Neutral
Anger	0.95	0.00	0.03	0.03
Sadness	0.00	0.79	0.03	0.18
Happiness	0.02	0.00	0.91	0.08
Neutral	0.01	0.05	0.02	0.92

- Decision-level integration (89%)
  - Product of the posterior probabilities was the best criterion
  - Product criterion: Same results, big differences

	Overall	Anger	Sadness	Hapiness	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

In detail

	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 future work

Discussion

- The multimodal systems give an improvement of 5% (absolute) compared to unimodal systems
- Some pair of emotions confused in one modality are easily separated in the other modality (E&R)
  - Sadness-Anger
  - Neutral-Happiness
  - Anger-Happiness
- Sadness and neutral are confused in both domains
  - Feature-level integration systems cannot separate them accurately
  - Decision-level integration systems maybe (in our experiments, yes) (E&R2)
- Feature and decision-level integration systems
  - Similar overall results
  - Analysis in detail show big differences

Discussion and future work (2)

Discussion (contd.)

• 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 (A3)

• 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)

Discussion and future work (3)

Future Work

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

References

[Chen98] Chen, L.S., T.S. Huang, T. Miyasato, and R. Nakatsu,(1998) "Multimodal human emotion / expression recognition," in Proc. of Int. Conf. on Automatic Face and Gesture Recognition, (Nara, Japan), IEEE Computer Soc., April 1998.

[Chen98_2] Chen, L.S.; Tao, H.; Huang, T.S.; Miyasato, T.; Nakatsu, R.; (1998). Emotion recognition from audiovisual information. Multimedia Signal Processing, 1998 IEEE Second Workshop on , 7-9 Dec. 1998 Pages:83 – 88.

[DeSilva97] De Silva, L. C., Miyasato, T., and Nakatsu, R., (1997). Facial Emotion Recognition Using Multimodal Information. in Proc. IEEE Int. Conf. on Information, Communications and Signal Processing (ICICS'97), Singapore, pp. 397-401, Sept. 1997.

[DeSilva00] De Silva, L.C.; Pei Chi Ng; (2000). Bimodal emotion recognition. Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on , 28-30 March 2000. Pages:332 – 335

[Huang98] Thomas S. Huang, Lawrence S. Chen, Hai Tao, Tsutomu Miyasato, Ryohei Nakatsu (1998). Bimodal Emotion Recognition by Man and Machine. Proceeding of ATR Workshop on Virtual Communication Environmnts, (Kyoto, Japan), April 1998.

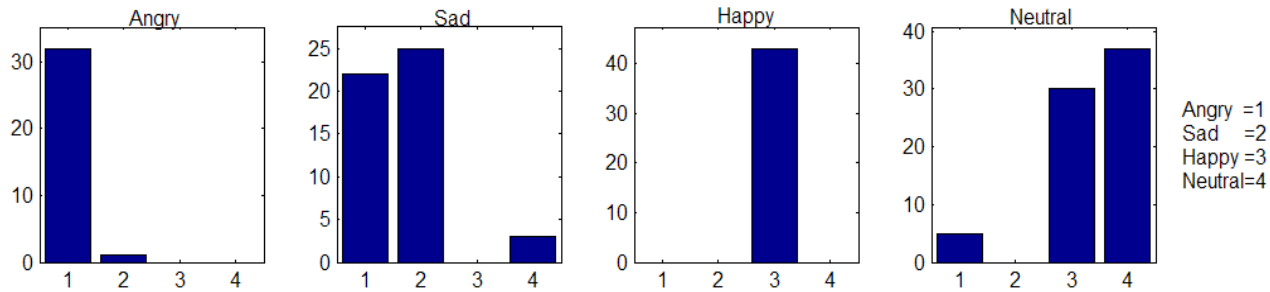
[Pantic03] Pantic, M., Rothkrantz, L.J.M. Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE , Volume: 91 Issue: 9 , Sept. 2003. Page(s): 1370 –1390.

[Yoshitomi00] Yoshitomi, Y., Sung-Ill Kim, Kawano, T., Kilazoe, T. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. Robot and Human Interactive Communication, 2000. RO-MAN 2000. Proceedings. 9th IEEE International Workshop on, 27-29 Sept. 2000. Pages: 178 – 18.

Why is the data preprocess?

Fusion techniques

- Global features are needed
- To take advantage of characteristic patterns of the frame data statistics



Why are these emotions confused?

Speech Analysis

- Results agree with human evaluations [DeSilva97]
- Speech associated with anger and happiness is characterized by
  - longer utterance duration
  - shorter inter-word silence
  - higher pitch and energy values with wider ranges
- Speech associated with neutral and sad sentences
  - Energy and the pitch are usually maintained at the same level.

Experiment and Results (ext)

Performance of Facial Expression Classifier by facial area

- Eyebrow give worse performance
- More experiments should be conducted to evaluate these results with other subjects

Area	Overall	Anger	Sadness	Hapiness	Neutral
Forehead	0.73	0.82	0.66	1.00	0.46
Eyebrow	0.68	0.55	0.67	1.00	0.49
Low eye	0.81	0.82	0.78	1.00	0.65
Right cheek	0.85	0.87	0.76	1.00	0.79
Left cheek	0.80	0.84	0.67	1.00	0.67
Combined classifier	0.85	0.79	0.81	1.00	0.81