Visual Emotion Recognition Using Compact Facial Representations and Viseme Information

Angeliki Metallinou, Carlos Busso†, Sungbok Lee, Shrikanth Narayanan

Signal Analysis and Interpretation Lab, University of Southern California.
† Electrical Engineering Department, University of Texas Dallas.

Face during Emotional Speech

emotional gestures + articulation movements

Focus

- Multi-speaker emotional database
- Detailed Motion Capture (MOCAP) facial data
- Low dimensional facial representations
  - Decorrelated markers, PCA, Fisher criterion
- Dynamically model articulation using visemes
  - Improvement in recognition performance

IEMOCAP Database

- Dyadic acted emotional database
- Multimodal (audio, video, MOCAP, text)
- Multi-speaker: 10 actors
- Improvisations + scripts
- Annotations
  - Categorical attributes
  - Dimensional attributes
- **http://sail.usc.edu/data.php**

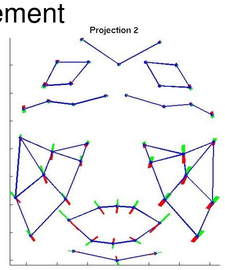
Emotions examined

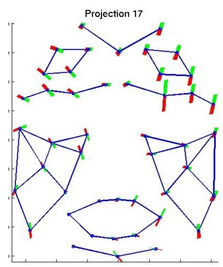
- angry, happy (+excited), neutral, sad

Facial Feature Extraction

Principal Component Analysis-PCA

- Linear projection to maximize variance
- 30 first principal components (out of 138)
  - Keep >95% of variance
- Derivatives of projections: 60-dimensions
- Interpretation: principal directions of facial movement

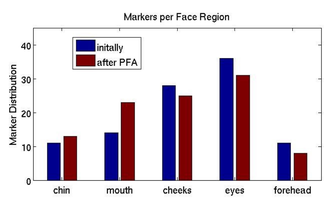




Principal Feature Analysis-PFA

- Decorrelate markers using PCA criteria
- Select a reduced set of decorrelated markers
- Method:
  - Average neighboring markers
  - PFA to select 30 features
  - Normalization
  - Plus derivatives
    →60 dimensions

Lower face bias

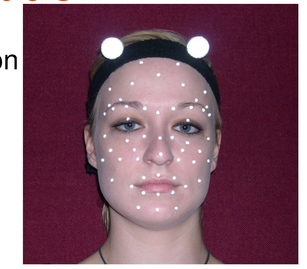


Fisher Feature Selection

- Select features that:
  - Maximize between class variability
  - Minimize within class variability
- Method:
  - Average neighboring markers
  - Normalization
  - Fisher feature to select 30 features
  - Plus derivatives
    →60 dimensions

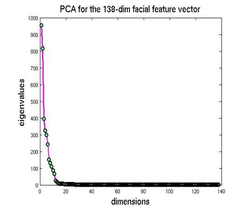
Facial Marker Information

- Motion Capture
- Normalize head rotation/translation
- Nose is the coordinate center
- 46 markers * 3 coordinates
  - 138 dimensions
- Redundant:
  - Underlying muscles
  - Facial configuration



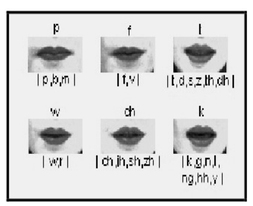
Speaker Face Normalization

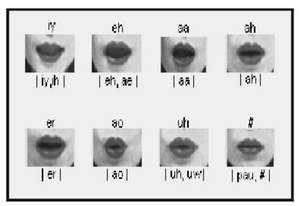
- Shift each marker such that:
- Speaker mean marker positions
  → global mean marker positions



Viseme Modeling

- Viseme: Lip shape during the voicing of a phoneme
- Modeling emotional visemes:
  - Constraint the speech-related variability
- Use available phoneme-level transcriptions
- Dynamic modeling using HMMs





Experiments

PFA	ANG%	HAP%	NEU%	SAD%	UW %
GMM(16)	47.03	73.37	36.55	58.35	53.83±6.09
Viseme-GMM(16)	58.44	71.71	37.22	49.28	54.16±6.24
Viseme-HMM(16)	57.52	76.98	34.79	53.68	55.74±5.26
Fisher	ANG%	HAP%	NEU %	SAD %	UW %
GMM(8)	49.87	72.82	27.35	55.27	51.33±7.23
Viseme-GMM(8)	62.78	73.76	29.57	42.62	52.18±7.05
Viseme-HMM(8)	62.92	75.97	27.65	46.46	53.25±8.30

- 10-fold leave-one-speaker-out cross validations
- Report the average over the 10 folds
- Decisions are per sentence
  - Majority rule used

Discussion and Conclusion

Emotions:

- Happiness is the best recognized emotion
- Neutrality is the lowest recognized state
- Great performance differences between speakers
- PFA and Fisher features have similar performance

Visemes:

- Dynamic articulation modeling is beneficial:
  - Total unweighted performance (UW)
  - Anger and happiness
- ...but sadness recognition performance decreases

Limitations:

- Multimodal nature of emotional expression

References

1) C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP:Interactive Emotional Dyadic Motion Capture Database," Journal of Language Resources and Evaluation, vol. 42, pp. 335–359, 2008.

2) M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71–86, 1991.

3) I. Cohen, Q. T. Xiang, S. Zhou, X. Sean, Z. Thomas, and T. S. Huang, "Feature selection using principal feature analysis," 2002.

4) S. Lee and D. Yook, "Audio-to-visual conversion using Hidden Markov Models," in Proc. 7th Pacific Rim Int. Conf. on Artificial Intelligence, 2002.

This research was supported by funds from NSF