VISUAL EMOTION RECOGNITION USING COMPACT FACIAL REPRESENTATIONS AND VISEME INFORMATION

Angeliki Metallinou, Carlos Busso, Sungbok Lee and Shrikanth Narayanan

Department of Electrical Engineering, University of Southern California,
Los Angeles, CA 90089-2560

metallin@usc.edu, busso@utdallas.edu, sungbokl@usc.edu, shri@sipi.usc.edu

ABSTRACT

Emotion expression is an essential part of human interaction. Rich emotional information is conveyed through the human face. In this study, we analyze detailed motion-captured facial information of ten speakers of both genders during emotional speech. We derive compact facial representations using methods motivated by Principal Component Analysis and speaker face normalization. Moreover, we model emotional facial movements by conditioning on knowledge of speech-related movements (articulation). We achieve average classification accuracies on the order of 75% for happiness, 50-60% for anger and sadness and 35% for neutrality in speaker independent experiments. We also find that dynamic modeling and the use of viseme information improves recognition accuracy for anger, happiness and sadness, as well as for the overall unweighted performance.

Index Terms— Emotion recognition, Principal Component Analysis, Principal Feature Analysis, Fisher Criterion, visemes

1. INTRODUCTION

During emotional speech, the face conveys rich and diverse information. Facial gestures are affected by several factors, including the underlying articulatory speech production and the emotional state of the subject. The specific goal of this study is to investigate the role of static and dynamic information conveyed by the face during emotional speech particularly from the perspective of automatic recognition. Our focus is two fold; first, we compute compact facial representations which preserve the useful information of the facial shape and movements in terms of emotion recognition performance. Second, we model and recognize emotions by conditioning on knowledge of speech-related lip movements (visemes), which occur in parallel. The use of direct facial marker data enables us to overcome some of the present challenges in feature processing from video data, and focus on establishing feasibility bounds for emotion classification using visual features. The insights gained from emotional face analysis could be applied to improve automatic emotion recognition and create more natural human-computer interfaces. This is increasingly relevant given the widespread use of cameras, incorporated in laptops and other portable devices that enables the use of facial cues for processing and recognition.

We use facial information obtained from multiple markers across the face. This information is redundant; neighboring markers tend to be highly correlated because they are controlled by the same underlying muscle movements. Moreover, the human face has a specific configuration and the possible range of physical movement of each facial marker is limited. Dealing with this redundancy in data becomes especially important in view of their use for pattern classification applications. We apply Principal Component Analysis (PCA) for dimensionality reduction, which is a widely used technique with various applications, including eigenface analysis for computer vision [1],[2]. Alternatively, we select face markers using either Principal Feature Selection (PFA), a recently proposed technique motivated by PCA [3], or we apply Fisher Criterion in order to select features that better discriminate between different emotional classes [4].

In order to constrain the speech-related variability of facial movement we use the concept of viseme, which represents the lip shape during the articulation of a phoneme. Visemes are widely used for speech analysis [5], audio-visual recognition of speech, especially under noisy conditions [6] and animation [7]. In [8] authors use averaging in order to smooth the speech-related face movements. In contrast, we incorporate those movements in our analysis by modeling the evolution of emotional visemes. Dynamic modeling of information streams using HMMs has been shown to be a powerful method for audiovisual recognition [9].

In the presented work we use a multispeaker database and perform speaker independent cross validations. Facial features resulting from averaged, decorrelated and normalized marker information (PFA features) achieve good performance. Happiness is the most well-recognized emotion using facial cues, with a recognition performance on the order of 75%, in leave-one-speaker-out cross validation experiments. Anger and happiness have performance on the order of 50-60% while neutrality has performance on the order of 35%. We find that emotion recognition accuracy is highly speaker-dependent. Also, the lower face seems to convey more information compared to the upper face. Explicitly modeling articulation movements improves recognition for anger, happiness and neutrality but decreases performance for sadness.

2. METHODOLOGY

2.1. Database

We use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [10]. This database contains approximately 12 hours of audiovisual data from five mixed gender pairs of actors, male and female. IEMOCAP contains detailed facial information obtained from motion capture as well as video, audio and transcripts of each session. In comparison to other acted emotion databases where actors are asked to read out sentences displaying a specific emotion,

This work was supported in part by funds from NSF and the Army

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

in IEMOCAP two techniques of actor training are used in order to elicit emotional displays; scripts and improvisation of hypothetical scenarios. The sessions are approximately 5 minutes in length. During these sessions actors displayed various emotions according to the content of the session and the course of the interaction. The sessions were later manually segmented into utterances and annotated into categorical (anger, happiness, neutrality, etc) and dimensional tags (valence, activation, dominance).

In this study, we use the facial motion capture data as well as the transcripts, from all 10 speakers in the corpus. We examine classes of anger, happiness, excitation, neutrality and sadness. We have merged classes of happiness and excitation into a single class which we will refer to as happiness. All utterances examined have been tagged by at least three annotators across which there is majority consensus regarding the emotional tag.

2.2. Feature Extraction

The IEMOCAP data contain detailed facial marker coordinates from the actors during their emotional interaction. The positions of the facial markers can be seen in figure 1(a). The markers were normalized for head rotation and translation. The nose marker is defined as the local coordinate center of each frame. The five nose markers were excluded because of their limited movement. In total, information from 46 facial markers is used, the (x,y,z) coordinates. This results in a 138-dimensional facial representation, which tends to be redundant because it does not exploit the correlations of neighboring marker movements and the structure of the human face. We examine various feature extraction approaches in order to find compact facial representations well suited for emotion recognition applications in terms of recognition accuracy.

2.2.1. Speaker Face Normalization

When we examine various speakers it is important to smooth out individual speaker face characteristics that are not related to emotion. Our speaker normalization approach consists of finding a mapping from the individual average face to the general average face. This is achieved by shifting the mean value of each marker coordinate of each speaker to the mean value of that marker coordinate across all speakers. Specifically, for each speaker we compute the mean of each face feature (marker coordinate) across all emotions, m_{ij}, where *i* is the speaker index and *j* is feature index. Also, we compute the mean of each feature across all speakers and all emotions, M_j, where *j* is the marker coordinate index. Each feature is then normalized by multiplying it with the coefficient $c_{i,j} = \frac{M_j}{m_{ij}}$.

2.2.2. Principal Component Analysis

PCA is a widely used dimensionality reduction method that finds the projection of data into a lower dimensional linear space such that the variance of the projected data is maximized. The application of PCA in this paper is motivated by the technique of eigenfaces [1]. In eigenfaces, a feature vector is constructed by facial image pixel values. PCA finds the principal faces, which can be linearly combined to reconstruct any face. Similarly, in our approach the feature vector consists of the facial marker coordinates, and the principal projections can be interpreted as the directions of facial movement along which the variance is maximum.

In our analysis, we perform PCA and then reconstruct the face from the first 30 principal components (because they encode more than 95% of the total variance). In an attempt to interpret those principal components, we change the value of each component in the low-dimensional domain and observe the change in the facial marker domain. For instance in figures 1(b),(c) and (d), we can see a neutral configuration of the face of a female speaker, reconstructed from the 30 first PCA projections, and the face appearance when we decrease and increase the second projection coefficient, keeping all other coefficients constant. The resulting movement roughly corresponds to the facial configuration of a smile. A second example is provided in figures 1(e) and (f) when, respectively, decreasing and increasing the 17th projection coefficient results in the up and down movement of the right eyebrow. In general, some projections correspond to recognizable directions of facial movement, affecting either the lower or the upper facial parts or both.

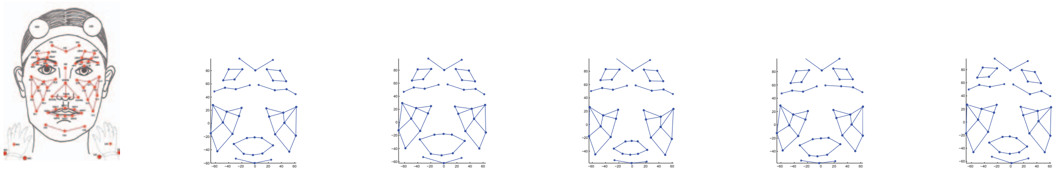
We use the first 30 PCA projections for efficient description of the face, since they convey more than 95% of the total data variance. The PCA transformation matrix is computed using data from all available speakers, therefore individual speaker characteristics are indirectly taken into accout. We find that speaker normalization, either prior or after the PCA transformation, does not improve recognition performance, therefore we do not normalize. The window used for feature extraction is 25msec with overlap of about 16msec. The choice of a short window enables further dynamic modeling of the visemes (since the average phoneme lasts about 100msec). We append the first derivatives to the feature vector resulting in a 60-dim facial representation.

2.2.3. Principal Feature Selection

The PCA transformation space is a linear combination of the initial space of face marker coordinates, with no inherent intuitive interpretation. Although, it is sometimes possible to interpret these projections as directions of specific face gestures and movements, generally it is difficult to find a meaning behind each projection. In order to find more meaningful facial representations we use Pricipal Feature Analysis [3]. This method computes the PCA transformation matrix as a first step and uses this matrix to cluster together facial marker coordinates that are highly correlated. Then it selects a representative feature from each cluster, thus performing feature selection while using similar criteria as PCA.

We find experimentally that it is beneficial to perform some adhoc averaging of neighboring facial markers prior to applying PFA. That way, highly correlated markers are averaged and the face markers are reduced from 46 to 28. Then we perform PFA, which selects the least correlated averaged marker coordinates, and finally we normalize the selected coordinates. The position of the marker coordinates is affected by the facial configuration of each speaker. Normalization is important in order to smooth out individual face characteristics which are unrelated to emotion and focus on emotional modulations. We select 30 features for similar reasons as the PCA features (PCA transformation explains more that 95% of total data variability). We use the same window and we append the first derivatives to the feature vector resulting in a 60-dim facial representation.

An analysis of the PFA process, shows that the facial features are clustered together in a meaningful way. For example same coordinates of neighboring or mirroring markers (e.g., left and right cheek) are clustered together. When repeating PFA 100 times, we find that on average 28% x-coordinates, 39% y-coordinates and 33% z-coordinates are selected, showing that all 3 coordinates have important variability in emotional speech. The com-



(a) Face markers (b) Neutral face (c) After descreasing projection 2 (d) After increasing projection 2 (e) After decreasing projection 17 (f) After increasing projection 17

Fig. 1. Facial marker positions and marker position reconstruction using the first 30 PCA projections (data from speaker 1, female)

paratively high percentage of y-coordinates selected is expected because of the jaw movements in the vertical direction. Indeed, on average 22% of the selected y-coordinates come from mouth markers while only 14% of the initial markers are placed around the mouth. Selection of z-coordinates can be attributed to lip protrusion during articulation. The distribution of initial markers across the face regions is (chin,mouth,cheeks,eyebrows,forehead) = (11%,14%,28%,36%,11%) while the distribution of the selected markers is (13%,23%,25%,31%,8%). This indicates a bias towards selecting lower face marker coordinates (especially mouth), which is expected since the movement of the jaw conveys a great amount of the variability. This is an encouraging result since the mouth can be automatically tracked more reliably than other face regions like cheeks and forehead.

2.2.4. Feature Selection using Fisher Criterion

The previously described features have been selected so as to capture maximal data variability. However, those features do not necessarily separate the different emotion classes well. We also extract a set of features using the Fisher criterion, which maximizes the between class variability and minimizes the within class variability [4]. For those features we first perform ad-hoc averaging of neighboring markers to reduce from 46 to 28, then perform speaker face normalization and finally we select the 30 best marker coordinates according to the Fisher criterion. The Fisher criterion value of each feature is computed on the train set, where the emotion classes are known. In each fold the Fisher criterion values are slightly different, so the features that are selected in different folds may vary slightly. The choice of 30 is ad-hoc so that this feature set is comparable with the previous two. Again, we append the first derivatives to the feature vector resulting in a 60-dim facial representation. From the selected features, across the 10 folds, on average 29% are x, 34% are y and 37% are z coordinates. Also, on average, 34% of the markers come from upper face (eyebrows and forehead) and 66% come from lower face. In general, we observe similar tendencies with PFA concerning the feature selection.

2.3. Viseme Information

A viseme specifies the lip shape during the articulation of a phoneme. The purpose of conditioning on visemes, is to constrain the variability related to speech, in order to better recognize the underlying emotion. Besides incorporating speech-related information, visemes provide a reasonable time unit for HMM training, if we want to dynamically model facial movement. The phoneme to viseme mappings are many to one, and various such mappings exist in the literature depending on the desired detail. Here we used the mapping presented in [6] and [11], in which the authors used 14 visemes. The phoneme to viseme mapping that we used is summarized in Table 1. For each utterance we have the word transcription and through forced alignment we obtain the phoneme-level transcription. We use this transcription to group facial data that correspond to each viseme.

Table 1. *Phoneme to viseme mapping.*

Visemes	Phonemes	Visemes	Phonemes
v1	P,B,M	v8	AE,AW,EH,EY
v2	F,V	v9	AH,AX,AY
v3	T,D,S,Z,TH,DH,CH,SH,ZH	v10	AA
v4	W,R	v11	AXR,ER
v5	CH,SH,ZH	v12	AO,OW,OW
v6	K,G,N,L,HH,NG,Y	v13	UH,UW
v7	IY,IH, IX	v14	SIL

3. EMOTION RECOGNITION EXPERIMENTS

3.1. Experimental Setup

We organize our emotion recognition experiments using a 10-fold leave-one-speaker-out cross validation. The mean and standard deviation of the number of test utterances across the folds is 59 ± 28 angry, 79 ± 25 happy, 56 ± 22 neutral and 62 ± 23 sad utterances. The presented recognition results are the speaker-independent averages over the 10 folds. For each of the feature sets, we examine whether explicitly modeling the visemes is beneficial. We train a static GMM for each emotion (no viseme information), a GMM for each emotional viseme in order to model viseme information and an HMM for each emotional viseme in order to model both viseme information and its dynamic evolution. For model training we use the HTK Toolkit [12]. For each model we try various number of mixtures, ranging from 4 to 32, and report the best performance. The results are per utterance and majority rule is used to obtain the utterance-level decision.

3.2. Results

In Table 2 we present the emotion classification percentages per utterance, using the PCA, PFA and Fisher selection features. For each emotion, we report the mean of the 10-fold cross validation. We also report the mean and standard deviation of the total unweighted performance (%UW). The parentheses next to each model name indicate the number of gaussian mixtures used. In Table 3 we present the detailed classification performance of our best performing classifier (HMM with normalized PFA features) per speaker.

4. DISCUSSION

Emotion recognition peformance follows similar trends across all feature sets. Specifically, happiness seems to be well trasmitted

Table 2. *Emotion Classification Percentages per utterance, using PCA, PFA and Fisher features. The mean classification percentage is computed over 10 leave-one-speaker-out cross validations.*

PCA	%ANG	%HAP	%NEU	%SAD	%UW
GMM(8)	35.42	72.10	25.06	**64.34**	49.23 ± 7.17
viseme-GMM(8)	52.70	73.67	**33.27**	44.85	51.12 ± 5.30
viseme-HMM(4)	**57.59**	**76.70**	23.13	52.67	**52.52** ± **4.82**
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**

Table 3. *Detailed Emotion Classification Percentages per speaker, using normalized PFA features and HMMs trained on emotional visemes*

HMM(16)-PFA	%ANG	%HAP	%NEU	%SAD	%UW
sp1 (female)	34.72	75.00	71.21	40.48	55.35
sp2 (male)	61.54	72.97	21.05	52.05	51.90
sp3 (female)	71.43	80.26	15.38	26.67	48.44
sp4 (male)	55.00	67.37	17.14	83.61	55.78
sp5 (female)	73.68	95.83	18.87	18.63	51.75
sp6 (male)	75.27	75.61	25.29	59.09	58.81
sp7 (female)	57.76	85.71	50.00	45.16	59.66
sp8 (male)	33.85	61.02	31.43	78.38	51.17
sp9 (female)	30.00	97.62	48.61	89.33	66.39
sp10 (male)	82.00	58.39	48.94	43.40	58.18

by the face since we achieve an average recognition performance around 75%. This result agrees with our intuition that the face can portray obvious expressions of happiness, e.g., through a smile. Our worst performance is for neutrality, which ranges from 25% (chance) to around 36% depending on the feature set. This may be due to the wide variability in the definition of neutrality. Finally, for anger and sadness we achieve performance on the order of 50-60% depending on the feature set. Overall, the PFA features are the best performing both in terms of total unweighted performance and neutral classification, while Fisher features perform slightly worse. This suggests that by selecting decorrelated marker features that represent maximal data variance, we essentially keep most of the emotion related information.

We observe similar trends across all feature sets, when modeling face at the viseme level. Using viseme information works well for anger and happiness. This may happen because these are emotions with high activation and high facial movement variability; thus constraining some of the variability that is related to speech seems to improve emotion recognition. Viseme modeling decreases recognition accuracies for sadness and slighly improves recognition of the neutral state. In Table 3, we observe large differences in recognition of various classes between speakers, which suggests that emotion expression is highly speaker dependent. Finally, total unweighted performance improves, across all feature sets, when conditioning on visemes. Constraining on visemes usually helps to reduce some of the speaker variability, as it is indicated by the decrease in the standard deviation of the total unweighted recognition performance (Table 2).

Finally, the database that we examine contains emotional expressions that are not caricatures but have been elicited so as to resemble natural emotional expression. Also, it is a multimodal database; annotators tagged each utterance taking into account not only the visual information, but also audio, content and context information. This means that there may be mismatches between the multimodal presentation of emotion and the emotional expression contained in the face alone. For example, one display of sarcasm consists of an angry voice and happy-looking face. Humans usually recognize the overall emotion as anger, however a model trained only on face data will most probably detect happiness. Such issues limit the performance of individual modality emotion classification and can be resolved by combining information from multiple modalities in our analysis.

5. CONCLUSION AND FUTURE WORK

In this work, we used a large, multi-speaker emotional database for the challenging speaker-independent emotion classification task. We examined a variety of facial representations and found that the use of averaged, decorrelated and normalized marker information leads to average accuracies on the order to 75% for happiness, 50-60% for anger and sadness and 35% for neutrality. We also found that explicitly modeling articulation movements improves recognition performace for anger and happiness and increases the overall total unweighted performance.

There are many potential directions for future work. One immediate step is to include multiple modalities in order to improve emotion recognition performance. The dynamic statistical modeling of multiple modalities and their effective fusion is an interesting and challenging problem. Concerning the use of PCA to find directions of facial movement, we could use the low dimensional PCA space to effectively reconstruct the face and control facial gestures. Manipulating the PCA projections has the advantage of causing more smooth and natural facial movements than just moving each marker separately. Therefore, those projections could be applied for realistic emotional speech animation.

6. REFERENCES

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

[2] N. Muller, L. Magaia, and B.M. Herbst, "Singular value decomposition, eigenfaces and 3d reconstructions," *SIAM Review*, vol. 46, no. 3, pp. 518–545, 2004.

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

[4] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley and Sons, 2001.

[5] M.R. MacEachern, "On the visual distinctiveness of words in the english lexicon," *Journal of Phonetics*, vol. 28, pp. 367–376, 2000.

[6] P. Lucey, T. Martin, and S. Sridharan, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in *Proc. Australian Int. Conf. on Speech Science and Technology*, 2004.

[7] J.M. DeMartino, L.P. Magalhaes, and F. Violaro, "Facial animation based on context-dependent visemes," *Computers and Graphics*, vol. 30, pp. 971–980, 2006.

[8] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T.S. Huang, and S. Levinson, "Audio-visual affect recognition through multi-stream fused HMM for HCI," in *CVPR*, 2005.

[9] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[10] 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.

[11] S. Lee and D. Yook, "Audio-to-visual conversion using hidden markov models," in *Proc. 7th Pacific Rim Int. Conf. on Artificial Intelligence*, 2002.

[12] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Cambridge Research Laboratory, Cambridge, England, 2006.