21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

# Audio-Visual Data Association for Face Expression Analysis

Ashish Tawari and Mohan Trivedi
*University of California San Diego, Dept. of ECE*
*atawari@ucsd.edu, mtrivedi@soe.ucsd.edu*

## Abstract

*We present a novel facial expression recognition framework using audio-visual information analysis. In particular, we design a single good image representation of the image sequence by weighted sum of registered face images where the weights are derived using auditory features. We use a still image based technique for the expression recognition task. We performed experiments using eNTERFACE'05 audio-visual emotional database. The analysis shows that our framework can improve the recognition performance while significantly reducing the computational cost by avoiding redundant or insignificant frame processing by incorporating auditory information.*

## 1 Introduction

Automatic analysis of human affective behavior have been extensively studied in past several decades. Facial expression recognition systems, in particular, have matured to a level where automatic detection of small number of expressions in posed and controlled displays can be achieved with reasonably high accuracy. These emotions are often deliberate and exaggerated displays [10]. However, the deliberate and spontaneous behavior differs in their visual appearance, audio profile and the timing between the two modalities [1, 6]. Detecting these expressions in less constrained settings during spontaneous behavior is still a challenging problem. The research shift towards this direction suggests to utilize the multimodal data analysis approaches [9].

Audiovisual emotion recognition literature suggest that there exist number of works fusing multimodal data using early-, model-level- or late- fusion schemes. However, very little emphasis is given to using cross modal information in enhancing signal representation of any particular modality. From the audiovisual data of naturalistic conversation [7], it is evident that speech generation influences facial expression. Also, for expression recognition, the coupling between these two modalities is not so tight unlike other cases such as audio-visual speech recognition. We investigate how to exploit these knowledge to improve facial expression recognition in terms of computation cost and accuracy.

In this paper, we propose a novel facial expression recognition framework using bimodal information. Our contributions are two folds. First, our framework explicitly models the cross-modality data correlation while allowing them to be treated as asynchronous streams. Second, we generate a single image representing the key emotion in the video (image sequence) containing hundreds of frames. This helps to circumvent the complex and noisy dynamics in the video frames and at the same time enables to utilize image based facial expression approaches. We also show that by incorporating cross-modal information a significant reduction in computation cost can be achieved. On the other hand by avoiding spurious frames for further processing and there by reducing unwanted influence, classification accuracy can be improved.

## 2 Audio-Visual Data Association Approach

Figure 1 sketches an overview of the proposed recognition system. Salient feature of our framework is the introduction of cross-modal relevance feedback blocks and frame relevance measure blocks. The cross-modal relevance feedback block measures the importance of the current frame of the other modality from the analysis of its modality. The frame relevance block can potentially use cross-modal feedback and the analysis of its modality to finally assess the relevance of the current frame.

In our present work, frame relevance block utilizes only cross-modal feedback to highlight the importance of cross-modal information. Also, we have focused our discussion to facial expression recognition using visual features alone. Hence classification module only utilizes visual features. An audio-visual classification

978-4-9906441-1-6 ©2012 IAPR

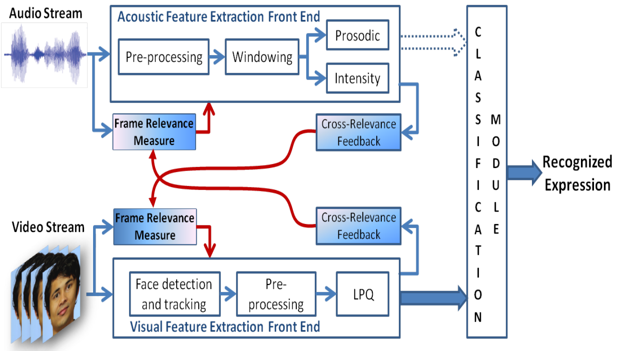

Figure 1: Overview of the proposed expression recognition system.

framework, however, can be devised to utilize standard fusion schemes (early-, model-level- or late-fusion). Important point to note is that the proposed method attempts to improve signal representation at the first place hence by reducing error propagation which, in general, is harder to deal at later stages.

A detailed approach to condense the visual expression information into a single image representation is presented in following sections.

## 2.1 Face tracking and alignment

The first step of visual processing involves face detection and tracking. This is accomplished using constraint local model (CLM) [5]. It is based on fitting a parameterized shape model to the location landmark points of the face. The fitting process on an image $I^{(m,n)}$ provides a row vector $P^{(m,n)}$ for each sequence $m$ and frame $n$ containing $l = 66$ detected landmark positions

$$P^{(m,n)} = [x_1, y_1; x_2, y_2; \cdots x_l, y_l]$$

The detected landmark is normalized by appropriate scaling, rotation and translation to make center of eyes 200 pixel apart and line joining the two centers horizontal. We denote the normalized shape vector as $P_N^{(m,n)}$. Further, a reference shape is calculated using Eq. 1.

$$P^{ref} = \sum_{m=1}^{M} \sum_{n=1}^{N_m} P_N^{(m,n)} \qquad (1)$$

where $N_m$ is total number of frames in sequence m and M is the total number of image sequences. Given this reference shape $P^{ref}$, image $I^{(m,n)}$ is aligned using affine transform to obtain the aligned image $I_{align}^{(m,n)}$. For alignment, we only considered the points which are relatively stable to track corresponding to the eyebrows, eyes, nose and mouth regions.

## 2.2 Visual sequence analysis: A Bimodal approach

Our aim is to provide segment level classification i.e. given a video segment, we would like to classify it as a particular expression class. However, a video segment has hundreds of frames and the question is how to utilize all or a subset of frames intelligently to come up with single image representation. For this, we propose to derive a weighted mean image $I_m^{rep}$ for the sequence $m$ which hopefully is representative of emotional content of the segment.

$$I_m^{rep} = \sum_{n=1}^{N_m} w(n) I_{align}^{(m,n)} \qquad (2)$$

where $\sum_{n=1}^{N_m} w(n) = 1$.

We design two rule based approaches to assign value of relevance measure $w(.)$ using auditory analysis. In the first approach, we assign $w(.)$ uniformly where the speech data is present. This removes preceding and trailing silence, and irrelevant frames where subject may not even be looking to camera for fare comparison with second approach. This is equivalent to keeping all the frames in the active video sequence; hence discarding any prosodic information available in audio stream. We call the resultant image as the '*mean*' image.

The second approach uses prosodic information related to pitch and intensity contour to chose only certain frames for the calculation of image $I^{rep}$. We use four sub-segments of the given video segment: two corresponding to start and end of the speech segment and two corresponding to maximum intensity and maximum pitch value. Each sub-segment is 200*ms* long, centered around mentioned events. Note that the later two segments may overlap. All the selected frames are assigned same weights. We call the resultant image as '*weighted-mean*' image. Figure 2 shows signal processing involved in a typical example of the two approaches and their *mean* and *weighted-mean* image output.

## 2.3 Appearance feature extraction

For the facial expression analysis, we use the blur insensitive Local Phase Quantization (LPQ) appearance descriptor proposed by Ojansivu et al. [3]. Due to space constraint, we encourage the reader to study [3] for the details. In our experiments, we used parameter $a = 1/3$. After histogram step, we get a 256 dimensional feature vector for a given image patch. We also use de-correlation process to eliminate the dependency of the neighboring pixels. In our experiment, we resize the representative face image $I_m^{rep}$ to $200 \times 200$ and further divided into non-overlapping tiles of $10 \times 10$ to extract local pattern. Thus the LPQ feature vector is of dimension $256 \times 10 \times 10 = 25600$.

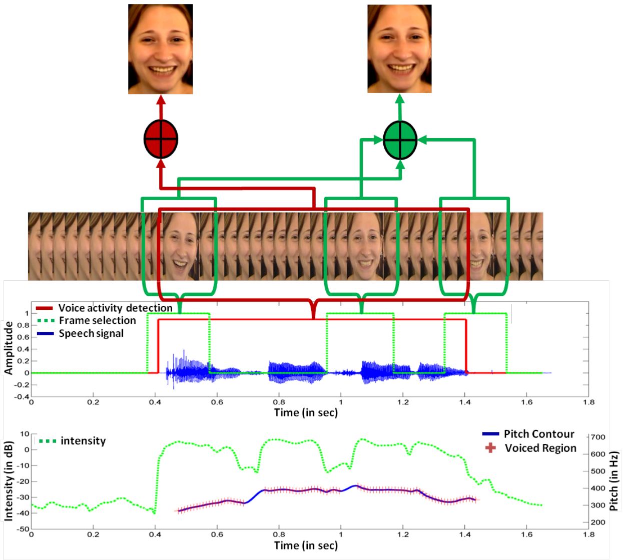

Figure 2: Signal processing involved in calculation of single image representation of the image sequence. The bottom curves are intensity (dotted green - associated with the left axis) and pitch contour (solid blue associated with right axis) along with voiced region depicted by the red cross. The middle plot is the speech signal showing the segments chosen by the two schemes '*weighted-mean*' in green and '*mean*' in red box. Finally the image sequence is shown next. All the plots and image sequence have the same time axis. Left and right image outputs are derived for the *Happy* expression class using '*mean*' and '*weighted-mean*' approaches respectively.

### 2.4 Auditory feature extraction

In our prior work [8], we have used prosodic and spectral features to model emotional states. In this paper, we use subset of these features for cross-modal relevance calculation. In particular we use the pitch and intensity contours to derive weights $w(n)$ for the $n^{th}$ frame in visual stream as described in Section 2.2.

For pitch calculation, we use the auto-correlation algorithm similar to [4]. The speech signal is divided into overlapping frames with overlap of $10ms$ and frame length of $60ms$ to span 3 periods of minimum pitch (50Hz). We futher use a dynamic programming approach to get the final pitch contour from the pitch candidates calculated over each frame. Log-intensity coefficients are calculated using $30ms$ frames with shift interval of $10ms$. Figure 2 shows the interpolated pitch contour and voiced segment as well as the intensity contour.

## 3 Experimental Analysis

For the evaluation of the proposed framework, we use eNTERFACE'05 [2] audio-visual affective database. It contains the six archetypal emotions: happy (ha), sad (sa), surprise (su), anger (an), disgust (di) and fear (fe). The database is collected in a controlled recording environment from 42 subjects.

In our experiments, we perform binary classification using Support Vector Machines (SVMs) with linear kernel and default parameters available in Matlab implementation. We have 15 binary classification tasks corresponding to every possible pair of six expression classes available in the database. This is to emphasize the importance of bimodal data association in facial expression recognition using visual sequence data. Also, binary classification analysis helps us gain better insight on, specifically, the impact of our proposed framework and generally, the inherent confusion between two classes as discussed in the following section.

### 3.1 Results and Discussion

We conducted two experiments using different cross validation strategies: randomized 10 fold cross validation and leave-one-subject-out cross validation. The later provides subject-independent analysis while the formal attempts to provide subject-dependent analysis as training and testing set can contain same subjects. Using only single subject data for subject-dependent analysis may not have provided useful results since the database have only five instances of an emotion class per subject.

Firstly, it can be observed from Table 1 that the use of single image representation can provide high recognition accuracy. The best accuracy is obtained for the *Happy/Anger* binary classification with over 95% for randomized 10 fold cross validation. As expected, subject independent results show lower accuracy. Also, certain classes are more confusing in visual domian like the *Sad/Fear* or *Surprise/Fear* with recognition accuracy below 80%. It is important to point out, though, that we have not used any tuning of SVM parameters nor have we used any feature selection technique which often improves the performance greatly. Our focus is to compare the usefulness of auditory cross-modal feedack for frame selection which is also evident from the results.

Table 1 shows slight improvement on overall average performance by exploiting audio information. While the best improvement of 10% is obtained for binary classifiation task of *Surprise/Fear* in subject independent analysis (Table 1b). A closer look on the results suggests that emotion classes *Fear* and *Happy* have

| Method | Ha/Sa | Ha/Su | Ha/Fe | Ha/An | Ha/Di | Sa/Su | Sa/Fe | Sa/An | Sa/Di | Su/Fe | Su/An | Su/Di | Fe/An | Fe/Di | An/Di |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MI | 93.65 | 88.28 | 89.70 | 95.78 | 93.47 | 81.16 | 66.51 | 83.25 | 90.69 | 74.65 | 82.79 | 92.09 | 82.79 | 84.65 | 89.76 |
| WMI | 93.19 | 92.96 | 90.62 | 96.02 | 93.69 | 78.0 | 74.5 | 83.25 | 90.00 | 79.30 | 81.39 | 93.02 | 82.55 | 88.13 | 88.83 |
| Average Accuracy (%) - MI: 85.95 and WMI: 87.03 | | | | | | | | | | | | | | | |

(a)

| Method | Ha/Sa | Ha/Su | Ha/Fe | Ha/An | Ha/Di | Sa/Su | Sa/Fe | Sa/An | Sa/Di | Su/Fe | Su/An | Su/Di | Fe/An | Fe/Di | An/Di |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MI | 86.40 | 82.00 | 80.00 | 87.20 | 85.20 | 71.20 | 56.80 | 69.20 | 86.40 | 61.20 | 72.00 | 88.8 | 71.2 | 81.60 | 88.80 |
| WMI | 84.80 | 86.40 | 80.80 | 87.60 | 81.20 | 68.00 | 64.80 | 68.80 | 86.00 | 71.20 | 71.20 | 87.60 | 74.40 | 80.00 | 84.40 |
| Average Accuracy (%) - MI: 77.86 and WMI: 78.48 | | | | | | | | | | | | | | | |

(b)

Table 1: Classification accuracy for the possible 15 different combinations of the binary classification tasks over six basic emotions: happy (ha), Sad (sa), Surprise (Su), Fear (Fe), Anger (An) and Disgust (Di). (a) randomized 10 fold cross validation (b) leave-one-subject-out cross validation. Note that the computation cost associated with visual processing of *weighted-mean* image (WMI) is at least one third than that of *mean* image (MI) method.

shown the most improvements. On the other hand, emotion classes *Disgust* and *Sad* may have not been benefited and even showing opposite trend in some cases. This can be attributed to our rule based weight assignment for these emotion classes. Particularly, for *Sad* class having low arousal profile, region corresponding to high intensity and pitch may not provide representative frames. This encourages us to learn such bimodal association automatically from audio visual data.

Another important performance metric is the computation cost. Notice that audio assisted approach utilizes maximum of $4 \times 200ms = 800ms$ worth of visual data corresponding to the four segments as described in Section 2.2 while using all the frames on an average requires $2.5sec$ worth of visual frame processing. Hence using cross-modal information improved the visual computation cost by factor of $\sim 3$.

## 4 Concluding Remarks

We presented a novel approach of summarizing emotional content of the video frames by a single image using cross-modal data association. We then investigated two different rule based data association approach for face expression recognition task. Our results showed that use of audio data could improve the performance in terms of computation cost (since in general visual processing is costlier than audio processing) as well as recognition accuracy. Unlike various data fusion strategies, our approach attempted to better represent signal at feature extraction level by weighting frames by it's importance based on cross-relevance feedback.

In our future efforts, we will explore data driven approach to learn better and more realistic cross-modal relevance measure as opposed to simple uniform weights used in present study. We will also incorporate audio modality in classification module and examine the multi-class classification approach for the design of fully automatic audio-visual affect recognition system.

## References

[1] J. F. Cohn and K. L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *J. Wavelets, Multi-resolution and Information Processing*, 2:1–12, 2004.

[2] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The enterface'05 audio-visual emotion database. In *Proceedings of the 22nd International Conference on Data Engineering Workshops*, page 8. IEEE Computer Society, 2006.

[3] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, editors, *Image and Signal Processing*, volume 5099, pages 236–243. 2008.

[4] B. Paul. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Inst of Phonetic Sciences 17*, pages 97–110, 1993.

[5] J. Saragih, S. Lucey, and J. Cohn. Face alignment through subspace constrained mean-shifts. In *Int. Conf. on Computer Vision*, pages 1034 –1041, 2009.

[6] N. Sebe, M. Lew, I. Cohen, Y. Sun, T. Gevers, and T. Huang. Authentic facial expression analysis. In *Proceedings of International Conf. on Automatic Face and Gesture Recognition*, pages 517 – 522, may 2004.

[7] A. Tawari and M. M. Trivedi. Speech emotion analysis: Exploring the role of context. *IEEE Transactions on Multimedia*, 12(6):502 –509, oct. 2010.

[8] A. Tawari and M. M. Trivedi. Speech emotion analysis in noisy real world environment. In *Proceedings of the International Conference on Pattern Recognition*, 2010.

[9] A. Tawari and M. M. Trivedi. Audio visual cues in driver affect characterization: Issues and challenges in developing robust approaches. In *International Joint Conf. on Neural Networks*, pages 2997 –3002, 2011.

[10] Y. Tian, T. Kanade, and J. Cohn. Facial expression analysis. In S. Li and A. Jain, editors, *Handbook of Face Recognition*, pages 247–276. Springer, 2005.