AUTOMATIC FACE EMOTION RECOGNITION BY AUDIO-VISUAL CROSS MODEL DATA ASSOCIATION

[1] Megha Agarwal, [2] Pankaj Kumar
[1] Assistant Professor, [2] Assistant Professor
[1] Shri Ramswaroop Memorial College of Engineering and Management, Lucknow, India.

Abstract— Facial expression plays an important role in our daily activities. It can provide sensitive and meaningful cues about emotional response and plays a major role in human interaction and nonverbal communication. Facial expression analysis and recognition presents a significant challenge to the pattern analysis and human machine interface research community. This research aims to develop an automated and interactive computer vision system for human facial expression recognition and tracking based on the facial structure features and movement information. Most automatic expression analysis system attempt to recognize a small set of prototypic expressions, such as happiness, anger, surprise, fear, disgust and sad. 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. 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.

Keywords- Audio-visual expression recognition, key frames selection, multi-model expression recognition, emotion recognition, affective computing.

I. INTRODUCTION *(HEADING 1)*

Facial expression plays an important role in our daily activities. The human face is a rich and powerful source full of communicative information about human behavior and emotion. The most expressive way that humans display emotions is through facial expressions. Facial expression includes a lot of information about human emotion. It is one of the most important carriers of human emotion, and it is a significant way for understanding human emotion. It can provide sensitive and meaningful cues about emotional response and plays a major role in human interaction and nonverbal communication. Humans can detect faces and interpret facial expressions in a scene with little or no effort.

In recent years there has been a growing interest in developing more intelligent interface between humans and computers, and improving all aspects of the interaction.

This emerging field has attracted the attention of many researchers from several different scholastic tracks, i.e., computer science, engineering, psychology, and neuroscience. These studies focus not only on improving computer interfaces, but also on improving the actions the computer takes based on feedback from the user. There is a growing demand for multi-modal/media human computer interface (HCI). The main characteristics of human communication are: multiplicity and multi-modality of communication channels. A channel is a communication medium while a modality is a sense used to perceive signals from the outside world. Examples of human communication channels are: auditory channel that carries speech, auditory channel that carries vocal intonation, visual channel that carries facial expressions, and visual channel that carries body movements.

Recent advances in image analysis and pattern recognition open up the possibility of automatic detection and classification of emotional and conversational facial signals. Automating facial expression analysis could bring facial expressions into man-machine interaction as a new modality and make the interaction tighter and more efficient. Facial expression analysis and recognition are essential for intelligent and natural HCI, which presents a significant challenge to the pattern analysis and human-machine interface research community. To realize natural and harmonious HCI, computer must have the capability for understanding human emotion and intention effectively. Facial expression recognition is a problem which must be overcome for future prospective application such as: emotional interaction, interactive video, synthetic face animation, intelligent home robotics, 3D games and entertainment. An automatic facial expression analysis system mainly include three important parts: face detection, facial feature points extraction and facial expression classification.

In this paper, we propose a novel facial expression recognition framework using bimodal information. Our contributions are two folds. First, our frame-work 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 thereby reducing unwanted influence, classification accuracy can be improved.

II. 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.

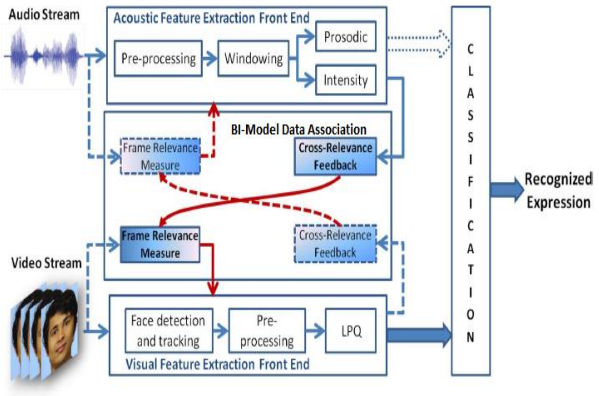


Figure1: Overview of the proposer expression recognition system.

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 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 at-tempts 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.

A. *Face Tracking and Alignment*

The first step of visual processing involves face detection and tracking. This is accomplished using constraint local model (CLM). 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) = [x1; y1; x2; y2;……… xl; yl]

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.

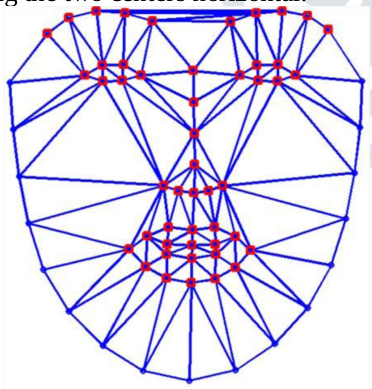


Figure2: Reference shape landmark position

We denote the normalized shape vector as PN(m;n) Further, a reference shape is calculated using Eq. 1.

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

Where Nm is total number of frames in sequence m and M is the total number of image sequences. Given this Reference shape Pref, image I(m;n) is aligned using affine transform to obtain the aligned image Ialign(m;n). For alignment, we only considered the points which are relatively stable to track corresponding to the eyebrows, eyes, and nose and mouth regions.

B. *Visual Sequence Analysis:A Bimodel 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_{rep} m 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 choose 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.

C. *Appearance Feature Extraction*

For the facial expression analysis, we use the blur insensitive Local Phase Quantization (LPQ) appearance descriptor proposed by Ojansivu et al. Due to space constraint, we encourage the reader to study 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 ImIrp to 200 X 200 and further divided into non-overlapping tiles of 10 X 10 to extract local pattern. Thus the LPQ feature vector is of dimension 256 X 10 X 10 = 25600.

D. *Auditory Feature Extraction*

In our prior work, 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 nth .

For pitch calculation, we use the auto-correlation algorithm similar to. 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 further 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.

III. Expermintal Analysis

For the evaluation of the proposed framework, we use self-create 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 in-sight on, specifically, the impact of our proposed frame-work

IV. Result 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.

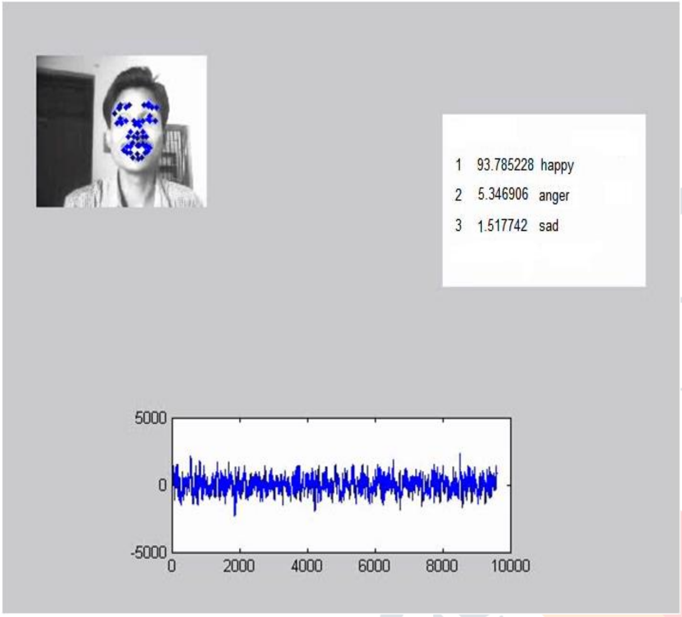
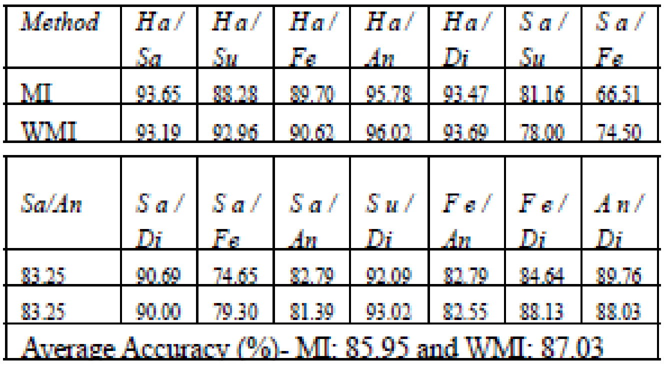


Figure3: Image derived for the expression class of Ha/An

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 folds cross validation. As expected, subject independent results show lower accuracy. Also, certain classes are more confusing in visual domain 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 feedback for frame selection which is also evident from the results.

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, the computation cost associated with visual processing of weighted-mean image (WMI) is at least one third than that of mean image (MI) method.



Method	Ha / Sa	Ha / Su	Ha / Fe	Ha / An	Ha / Di	Sa / Su	Sa / Fe
MI	93.65	88.28	89.70	95.78	93.47	81.16	66.51
WMI	93.19	92.96	90.62	96.02	93.69	78.00	74.50

Sa/An	Sa / Di	Sa / Fe	Sa / An	Su / Di	Fe / An	Fe / Di	An / Di
83.25	90.69	74.65	82.79	92.09	82.79	84.64	89.76
83.25	90.00	79.30	81.39	93.02	82.55	88.13	88.03
Average Accuracy (%)- MI: 85.95 and WMI: 87.03							

A closer look on the results suggests that emotion classes Fear and Happy have 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.

V. Conclusion

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 is 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 enter-face'05 audio-visual emotion database. In Proceedings of the 22nd International Conference on Data Engineering Workshops,

[3] V. Ojansivu and J. Heikkila¨. Blur insensitive texture classification using local phase quantization. In A.El-moataz, O. Lezoray, F. Nouboud, and D. Mammass, ed-itors, Image and Signal Processing, volume 5099, pages 236–243.

[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,

[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] 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.