A Multilevel Fusion Approach for Audiovisual Emotion Recognition

Girija Chetty & Michael Wagner

National Centre for Biometric Studies
Faculty of Information Sciences and Engineering
University of Canberra, Australia

girija.chetty@canberra.edu.au , michael.wagner@canberra.edu.au

Abstract

The human computer interaction will be more natural if computers are able to perceive and respond to human non-verbal communication such as emotions. Although several approaches have been proposed to recognize human emotions based on facial expressions or speech, relatively limited work has been done to fuse these two, improve the accuracy and robustness of the emotion recognition system. This paper analyzes the strengths and the limitations of systems based only on facial expressions or acoustic information. It also analyses two approaches used to fuse these two modalities: decision level and feature level integration, and proposes a new multilevel fusion approach for enhancing the person dependant and person independent classification performance for different emotions. Two different audiovisual emotion data corpora was used for the evaluating the proposed fusion approach - DaFEx[1,2] and ENTERFACE[3] comprising audiovisual emotion data from several actors eliciting five different emotions – anger, disgust, fear, happiness, sadness and surprise. The results of the experimental study reveal that the system based on fusion of facial expression with acoustic information yields better performance than the system based on just acoustic information or facial expressions, for the emotions considered. Results also show an improvement in classification performance of different emotions with a multilevel fusion approach as compared to either feature level or score-level fusion.

Index Terms: audiovisual, multilevel, fusion, emotion,

1. Introduction

The new trends in human computer interaction, which have evolved from conventional mouse and keyboard to automatic speech recognition systems and special interfaces designed for disabled people, do not take complete advantage of multiple channels of inter-personal human communicative abilities, involving both verbal and non cues such as facial expressions and tone of the voice, resulting often in a less than natural interaction. If computers could recognize these emotional inputs, they could give specific and appropriate help to users in ways that are more in tune with the user's needs and preferences.

It is widely accepted from psychological theory that human emotions can be classified into six archetypal emotions: anger, disgust, fear, happiness, sadness and surprise [4]. Facial motion and the tone of the speech play a major role in expressing these emotions. The muscles of the face can be changed and the tone and the energy in the production of the speech can be intentionally modified to communicate different feelings. Human beings can recognize these signals even if they are subtly displayed, by simultaneously processing information acquired by ears and eyes. Based on psychological studies, which show that visual information modifies the perception of speech [5], it is possible to assume that human emotion perception follows a similar trend.

Motivated by these clues, De Silva et al. conducted experiments, in which 18 people were required to recognize emotion using visual and acoustic information separately from an audio-visual database recorded from two subjects [6]. They concluded that some emotions are better identified with audio such as sadness and fear, and others with video, such as anger and happiness. Moreover, Chen et al.[7] showed that these two modalities give complementary information, by arguing that the performance of the system increased when both modalities were considered together. Although several automatic emotion recognition systems have explored the use of either facial expressions [8],[9],[10],[11],[12] or speech [13],[14],[15] to detect human emotion states, relatively few efforts have focused on emotion recognition using both modalities [7],[16]. These previous studies fused facial expressions and acoustic information either at a decision-level, in which the outputs of the unimodal systems are integrated by the use of suitable criteria, or at a feature-level, in which the data from both modalities are combined before classification.

However, none of these papers attempted to compare the relative merits of each fusion approach or investigate whether a multilevel fusion approach is more suitable for emotion recognition. This paper evaluates these two traditional fusion approaches and investigates multilevel fusion approach involving a both feature-level and score-level fusion, to enhance the classification performance of six different emotions. The paper is organised as follows. The details of the two audiovisual emotion corpora used is described in the next section. Section 3 describes the details of acoustic and visual feature extraction from the emotion video clips in the database, and the details of the proposed multilevel fusion approach is described in Section 4. Results of some preliminary experiments are described in Section 5 and the paper concludes with discussion in Section 7 and some conclusions and plan for further work in Section 8.

2. Audiovisual Emotion Data

Two different corpora - DaFEx [1,2] and ENTERFACE [3], with actors eliciting six different emotions were used for evaluating the proposed fusion approach.

DaFEx is an Italian audiovisual database of posed human facial expressions collected with the purpose of creating a valid benchmark for the evaluation of synthetic faces and embodied conversational agents. DaFEx can also be used as a general reference for research on emotions and facial expressions. DaFEx is composed by 1008 short videos in which Ekman's prototypic emotions (happiness, sadness, anger, fear, disgust and surprise) plus the neutral expression are shown. Facial expressions were recorded by 8 italian professional actors (4 male and 4 female) on 3 intensity levels (low, medium, high) and in 2 different conditions: The utterance subset comprised the actors playing these emotions while uttering a phonetically rich and visemically balanced sentence ("*In quella piccola stanza vuota c_era per_ltanto una sveglia*", Italian for: "*In that little empty room there was only an alarm clock*"). For non-utterance subset, the actors played emotions without pronouncing any sentence. In addition, each video started and ended with the actor showing a neutral expression. Both video and audio signals were recorded. Each actor recorded a sub-set of 126 videos, which includes all the emotions considered, at the three intensity levels and in the two different conditions. One utterance subset was used for experiments reported in this paper. The recording was done with a digital camera (Canon MV630i) and a directional microphone (Sennheiser MKH 406T). Videos were then compressed with Indeo 5.10 compression and audio signal was filtered in order to eliminate external noise. Finally, videos were made available as *.avi* files with 360 x 288 pixel images. Figure 1 shows some images from this corpus.

Figure 1: *DaFeX audiovisual emotion corpus images.*

Enterface corpus comprised English audiovisual emotion data collected by a particular project group for a European Similar Network of excellence workshop [3]. The audiovisual emotion data in this corpus was collected from 44 male and female subjects containing five utterances each for each emotion –anger, disgust, fear, happiness, sadness and surprise.

3. Acoustic and Visual Feature Extraction

Relatively few efforts have focused on using both facial expressions and acoustic information to recognize emotions. De Silva et al. proposed a rule-based audio-visual fusion approach, in which the outputs of the unimodal classifiers are fused at the decision-level [16-8]. From audio, they used prosodic features, and from video, they used the maximum distances and velocities between six specific facial points. A similar approach was also presented by Chen et al. [7-4], in which the dominant modality, according to the subjective experiments conducted in [6-7], was used to resolve discrepancies between the outputs of uni-modal systems. In both studies, they concluded that the performance of the system increased when both modalities were fused. Yoshitomi et al. proposed a multimodal approach that not only considers speech and visual information, but also the thermal distribution acquired by infrared camera [17]. They argue that infrared images are not sensitive to lighting conditions, which is one of the main problems when the facial expressions are acquired with conventional cameras. They used a database recorded from a female speaker that read a single word acted in five emotional states. They integrated these three modalities at decision-level using empirically determined weights. The performance of the system was better when three modalities were used together. In [18] and [19], a bimodal emotion recognition system was proposed to recognize six emotions, in which the audio-visual data was fused at feature-level. They used prosodic features from audio, and the position and movement of facial organs from video. The best features from both unimodal systems were used as input in the bimodal classifier. They showed that the performance significantly increased from 69.4% (video system) and 75% (audio system) to 97.2% (bimodal system). However they use a small database with only six clips per emotion, so the generalizability and robustness of the results should be tested with a larger data set.

For preliminary investigations reported in this study, we avoided using any automatic face detection and face tracking algorithms for simplicity, and performed the extraction of facial features by the use of markers.

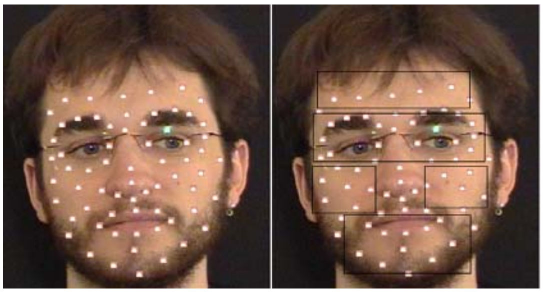


Figure 2: *Visual feature-markers and five face regions*

All markers were translated in order to make a nose marker be the local coordinate center of each frame, then one frame with neutral and close-mouth head pose was picked as the reference frame. The three approximately rigid markers (manually chosen and illustrated as white points in Figure 2) define a local coordinate origin for each frame, and each frame was rotated to align it with the reference frame. Each data frame is divided into five blocks: forehead, eye, lower mouth, right cheek and left cheek area (see Figure 2).

For each block, the 3D coordinate of markers in this block is concatenated together to form a data vector. Then, Principal Component Analysis (PCA) method is used to reduce the number of features per frame into a 10-dimensional vector for each area, covering more than 95% of the variation. Though The markers near the lips were shown in Figure 2, they were not considered for actual experiments, because the articulation of the speech might be recognized as a smile, confusing the recognition of that emotion state [20].

Notice that for each frame, a 10-dimensional feature vector is obtained in each block. This local information might be used to train dynamic models such as HMM. However, in this paper we decided to use global features at utterance level for both unimodal systems, so these feature vectors were preprocessed to obtain a low dimensional feature vector per utterance. In each of the 5 blocks, the 10-dimensional features at frame level were classified using a K-nearest neighbor classifier (k=3), exploiting the fact that different emotions appear in separate clusters. Then, the number of frames that were classified for each emotion was counted, obtaining a 4-dimensional vector at utterance level, for each block. These feature vectors at utterance level take advantage not only of the spatial position of facial points, but also of global patterns shown when emotions are displayed. For example, when happiness is displayed in more than 90 percent of the frames, they are classified as happy, but when sadness is displayed even more than 50 percent of the frames, they are classified as sad. The SVM classifiers use this kind of information, improving significantly the performance of the system. Also, with this approach the facial expression features and the global acoustic features do not need to be synchronized, so they can be easily combined in a feature-level fusion.

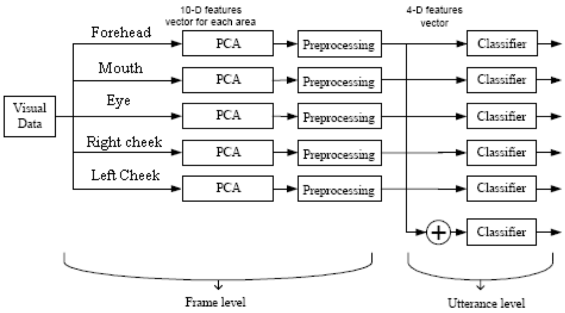


Figure 3: *Facial expressions feature extraction*

As shown in figure 3, a separate SVM classifier was implemented for each block, so it is possible to infer which facial area gives better emotion discrimination. In addition, the 4- dimensional features vectors of the 5 blocks were added before classification, as shown in figure 3.

The most widely used speech cues for audio emotion recognition are global-level prosodic features such as the statistics of the pitch and the intensity. Therefore, the means, the standard deviations, the ranges, the maximum values, the minimum values and the medians of the pitch and the energy were used as acoustic features. In addition, the voiced/speech and unvoiced/speech ratio were also estimated. By the use of sequential backward features selection technique, a 11-dimensional feature vector for each utterance was used as acoustic feature vector.

4. Audio-Visual Fusion

Six emotions – anger, disgust, fear, happiness, sadness and surprise were recognized by the use of four different approaches based on audio, facial expression bimodal, and multilevel bimodal fusion, respectively. The main purpose is to quantify the performance of unimodal systems, recognize the strengths and weaknesses of these approaches, compare different approaches to fuse these dissimilar modalities, and evaluate the proposed multilevel fusion approach to increase the overall recognition rate for different emotions. For all the experiments a support vector machine classifier (SVM) with 2nd order polynomial kernel functions [21]. Also, for all experiments, the training and testing using the corpora was done using leave-one-out cross validation method.

To fuse the facial expression features with acoustic features, three different approaches were used: feature-level fusion, in which a single classifier with features of both modalities are used (left of Figure 4a); score level fusion, in which a separate classifier is used for each modality, and the outputs are combined using some criteria (right of Figure 4a); and a multilevel fusion where feature fusion and score level fusion modules are fused with another level of fusion as shown in Figure 4b.

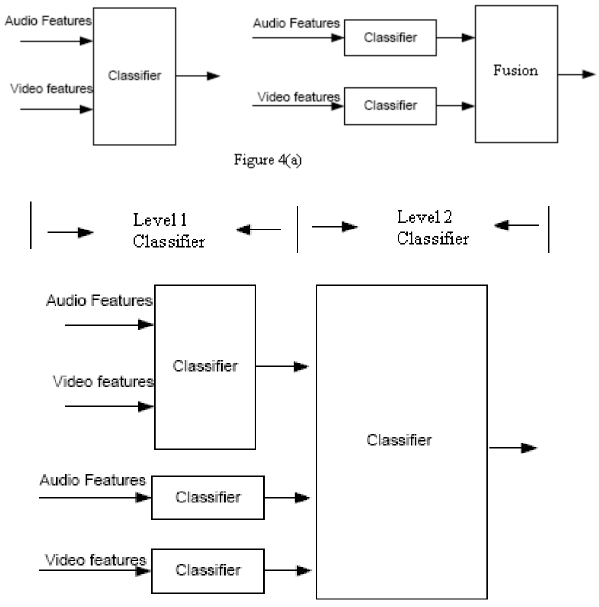


Figure 4(a)

Figure 4(b)

Figure 4:*(a) Feature-level and score level fusion (b) Multilevel fusion*

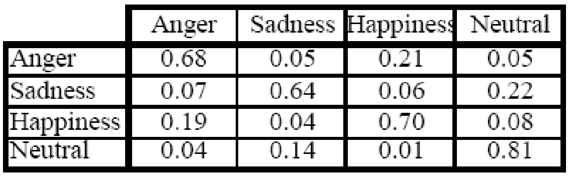
5. Experimental Results

In this section some of the preliminary emotion recognition experiments for the unimodal acoustic and visual expression features, and fusion of these features based on feature level, score level and multilevel approaches are described.

5.1. Emotion recognition using acoustic features

Table 1 shows the confusion matrix for recognizing four emotions based on acoustic information. For DaFEx corpus, the overall recognition performance was 70.9 %. The diagonal components of Table 1 reveal that all the emotions can be recognized with more than 64 percent of accuracy, by using only the features of the speech. However, Table 1 shows that some pairs of emotions are usually confused more. Sadness is misclassified as neutral state (22%) and vice versa (14%). The same trend appears between happiness and anger, which are mutually confused (19 %and 21 % respectively). These results agree with the human evaluations done by De Silva et al. [7], and can be explained by similarity patterns observed in acoustic parameters of these emotions [22]. For example, speech associated with anger and happiness is characterized by longer utterance duration, shorter inter-word silence, higher pitch and energy values with wider ranges. On the other hand, in neutral and sad sentences, the energy and the pitch are usually maintained at the same level. Therefore, these emotions were difficult to be classified. For ENTERFACE corpus, the recognition performance was 5-8% lower than the DaFEx corpus, though the performance trend for individual emotion states was similar to the DaFEx corpus.

Table 1. *Confusion matrix for emotion recognition based on acoustic features for DaFeX corpus*

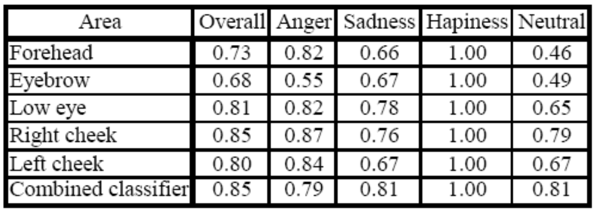


	Anger	Sadness	Happiness	Neutral
Anger	0.68	0.05	0.21	0.05
Sadness	0.07	0.64	0.06	0.22
Happiness	0.19	0.04	0.70	0.08
Neutral	0.04	0.14	0.01	0.81

5.2. Emotion recognition using facial expressions

Table 2 shows the performance of the emotion recognition for DaFEx corpus based on four facial expressions, for each of the five facial regions shown inn Figure 2, and the combined facial expression classifier. This table reveals that the cheek areas give valuable information for emotion classification. It also shows that the eyebrows, which have been widely used in facial expression recognition, give the poorest performance. Also happiness seems to be classified without any mistake easy to recognize. Table 2 also reveals that the combined facial expression classifier has an accuracy of 85%, which is higher than most of the 5 facial region classifiers. For ENTERFACE database, the performance for individual facial regions showed similar trend as for DaFeX, though the performance of combined classifier was 82%.

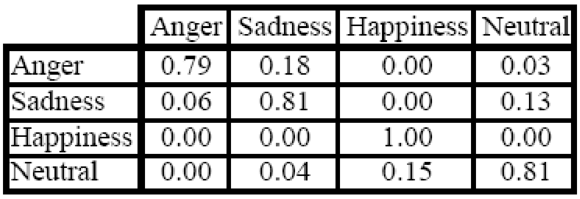
Table 2: *Performance of the facial expression classifiers for DaFeX corpus*



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

Table 3 shows the confusion matrix of the combined facial expression classifier to analyze in detail the limitation of this emotion recognition approach. The overall performance of this classifier was 85.1 percent for DaFEx corpus and 82% for ENTERFACE corpus. This table reveals that happiness is recognized with very high accuracy. The other three emotions are classified with 80 percent of accuracy, approximately. Table 3 also shows that in the facial expressions domain, anger is confused with sadness (18%) and neutral state is confused with happiness (15%). Notice that in the acoustic domain, sadness/anger and neutral /happiness can be separated with high accuracy, so it is expected that the bimodal fusion will give good performance for anger and neutral state. This table also shows that sadness is confused with neutral state (13%).

Table 3: *Confusion matrix of combined facial expression classifier for DaFeX corpus*



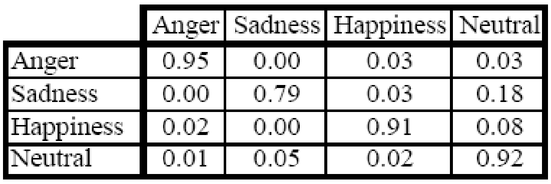
	Anger	Sadness	Happiness	Neutral
Anger	0.79	0.18	0.00	0.03
Sadness	0.06	0.81	0.00	0.13
Happiness	0.00	0.00	1.00	0.00
Neutral	0.00	0.04	0.15	0.81

Unfortunately, these two emotions are also confused in the acoustic domain (22%), so it is expected that the recognition rate of sadness in the bimodal classifiers will be poor. Other discriminating information such as contextual cues are needed.

5.3. Emotion recognition using audiovisual fusion

Table 4 displays the confusion matrix of the audiovisual fusion for DaFEx corpus when the facial expressions and acoustic information were fused at feature-level. The overall performance of this classifier was 89.1 percent. The overall performance of this classifier for ENTERFACE corpus was 86.1 percent. As can be observed, for both corpora, anger, happiness and neutral state are recognized with more than 90 percent of accuracy. As it was expected, the recognition rate of anger and neutral state was higher than unimodal systems. Sadness is the emotion with lower performance, which agrees with our previous analysis. This emotion is confused with neutral state (18%), because none of the modalities we considered can accurately separate these classes. Notice that the performance of happiness significantly decreased to 91 percent.

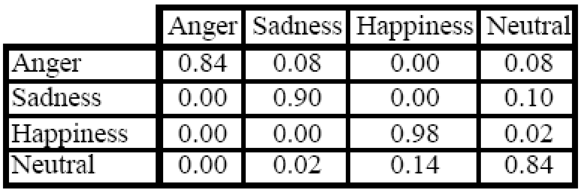
Table 4: *Confusion matrix of the feature-level fusion classifier for DaFeX corpus*



	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

Table 5 shows the confusion matrix of the score-level bimodal classifier when the product-combining criterion was used. The overall performance of this classifier was 89.0 percent for DaFEx corpus and 87.4% for ENTERFACE corpus, which is very close to the overall performance achieved by the feature-level bimodal classifier (Table 4). However, the confusion matrices of both classifiers show important differences. Table 5 for DaFEx corpus shows that in this classifier, the recognition rate of anger (84%) and neutral states (84%) are slightly better than in the facial expression classifier (79% and 81%, Table 4), and significantly worse than in the feature-level bimodal classifier (95%, 92%, Table 4). However, happiness (98%) and sadness (90%) are recognized with high accuracy compared to the feature-level bimodal classifier (91% and 79%, Table 4). These results suggest that in the score-level fusion approach, the recognition rate of each emotion is increased, improving the performance of the audiovisual fusion.

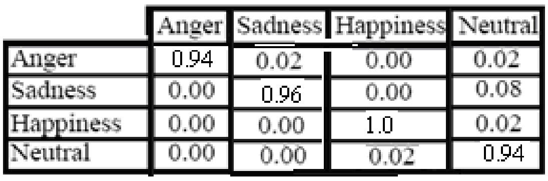
Table 5: *Confusion matrix of the score-level fusion classifier for DaFeX corpus*



	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

Table 6 shows the confusion matrix of the multilevel classifier shown in Figure 4(b) comprising a fusion of score-level bimodal classifier (product-combining criterion) with acoustic features and facial expression classifiers. The overall performance of this classifier was 97.0 % for DaFEx corpus and 96.4% for ENTERFACE corpus, which is better than the overall performance achieved by both the feature-level and score level audiovisual fusion (Table 5). However, the confusion matrices of both classifiers show important differences. The happiness (100 %) and sadness (96%) are recognized with high accuracy compared to other fusion approaches. These results suggest that for the multi-level fusion approach, the recognition performance of each emotion is increased, improving the overall performance of the system. Further, it was observed that the ENTERFACE data performed equally good as the DaFEx data for multilevel fusion mode, which means multilevel fusion makes the system more robust irrespective of the quality of the data and recording conditions

Table 5: *Confusion matrix of the multi-level fusion classifier for DaFeX corpus*



	Anger	Sadness	Happiness	Neutral
Anger	0.94	0.02	0.00	0.02
Sadness	0.00	0.96	0.00	0.08
Happiness	0.00	0.00	1.0	0.02
Neutral	0.00	0.00	0.02	0.94

6. Discussion

Humans use more than one modality to recognize emotions, so it can be expected that the performance of fusion of multiple modes at multiple levels will be higher than automatic unimodal systems. The results reported in this work confirm this hypothesis, since the multilevel fusion approach gave a significant improvement as compared to the performance of the acoustic or facial expression recognition schemes. The results show that pairs of emotions that were confused in one modality were easily classified in the other. For example, anger and happiness that were usually misclassified in the acoustic domain were separated with greater accuracy in the facial expression emotion classifier. Therefore, when these two modalities were fused at feature-level, these emotions were classified with high precision. Unfortunately, sadness is confused with neutral state in both domains, so its performance was poor. Although the overall performance of the feature-level and decision-level bimodal classifiers was similar, the multilevel fusion approach resulted in better emotion recognition performance. Also, an analysis of the confusion matrices of both classifiers revealed that the recognition rate for each emotion type was totally different. For the multilevel fusion classifier, the recognition rate of each emotion increased compared to the facial expression classifier, which was the best unimodal recognition system. In the feature-level bimodal classifier, the recognition rate of anger and neutral state significantly increased. However, the recognition rate of happiness decreased 9 percent. However, the preliminary results suggested in this paper suggest that the multilevel approach seems to be best approach to fuse the modalities. Also, the results reveal that, even though the system based on audio information had poorer performance than the facial expression emotion classifier, its features have valuable information about emotions that cannot be extracted from the visual information. These results agree with the finding reported by Chen et al. [7], which showed that audio and facial expressions data present complementary information. On the other hand, it is reasonable to expect that some characteristic patterns of the emotions can be obtained by the use of either audio or visual features. This redundant information is very valuable to improve the performance of the emotion recognition system when the features of one of the modal are inaccurately acquired as in two different emotion databases examined here. For example, if a person has beard, mustache or eyeglasses, the facial expressions will be extracted with high level of error. In that case, audio features can be used to overcome the limitation of the visual information.

Although the use of facial markers are not suitable for real applications, and an automatic face detection and feature tracking approach is needed, the preliminary analysis presented in this paper based on two emotion databases give important clues about emotion discrimination contained in different blocks of the face. The shapes and the movements of the eyebrows have been widely used for facial expression classification, the results presented in this paper show that this facial area does not provide good emotion discrimination as compared to other facial areas such as the cheeks. However, the results reported here were for just three emotion states and the neutral state - though the experiments were performed for all six emotions in the corpora. For other three emotions eyebrows do play an important role i.e. fear, disgust and surprise. Also, the experiments were conducted by using two emotion databases, where the quality of the data and the recording conditions was not similar.

7. Conclusions

In this paper, we reported the results of some preliminary experiments on analyzing the strengths and weaknesses of unimodal facial expression and acoustic emotion classification approaches, where some pairs of emotions are usually misclassified. We proposed an audiovisual fusion approach at multiple levels to show that most of these confusions could be resolved. The further plans for this research will be to find better methods to fuse audio-visual information that can model the dynamics of facial expressions and speech. Segmental level acoustic information can be used to trace the emotions at a frame level. Also, it might be useful to find other kind of features that describe the relationship between both modalities with respect to temporal progression. For example, the correlation between the facial motions and the contour of the pitch and the energy might be useful to discriminate emotions.

8. References

[1] Battocchi, A.; Pianesi, F.. 2004. DaFEx: Un Database di Espressioni Facciali Dinamiche. In Proceedings of the SLI-GSCP Workshop "Comunicazione Parlata e Manifestazione delle Emozioni", Padova (Italy) 30 Novembre - 1 Dicembre 2004.

[2] Mana N., Cosi P., Tisato G., Cavicchio F., Magno E. and Pianesi F., An Italian Database of Emotional Speech and Facial Expressions, In Proceedings of "Workshop on Emotion: Corpora for Research on Emotion and Affect", in association with "5th International Conference on Language, Resources and Evaluation (LREC2006), Genoa, Italy, 24-25-26 May 2006.

[3] Martin O., Adell J., Huerta A., Kotsia I., Savran A., Sebbe R., Multimodal Caricatural Mirror, Proceedings Enterface'05,Workshop, http://www.enterface.net/enterface05/docs/results/reports/project2.pdf

[4] Carlos, B., Zhigang D., Serdar Y., Murtaza B., Chul Min L., Abe K., Sungbok L., Ulrich N., Shrikanth N., Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information, Proc. of ACM 6th International Conference on Mutlmodal Interfaces (ICMI 2004), State College, PA, Oct 2004.

[5] Lee C. M., Narayanan, S.S., Pieraccini, R. Classifying emotions in human-machine spoken dialogs. Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International. Conference on , Volume: 1 , 26-29 Aug. 2002. Pages:737 - 740 vol.1.

[6] De Silva, L. C., Miyasato, T., and Nakatsu, R. 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.

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

[8] Black, M. J. and Yacoob, Y. Tracking and recognizing rigid and non-rigid facial motions using local parametric model of image motion. In Proceedings of the International Conference on Computer Vision, pages 374–381. IEEE Computer Society, Cambridge, MA, 1995.

[9] Essa, Pentland, A. P. Coding, analysis, interpretation, and recognition of facial expressions. IEEE Transc. On Pattern Analysis and Machine Intelligence, 19(7):757–763, JULY 1997.

[10] Mase K. Recognition of facial expression from optical flow. IEICE Transc., E. 74(10):3474–3483, 0ctober 1991.

[11] Tian, Ying-li, Kanade, T. and Cohn, J. Recognizing Lower Face Action Units for Facial Expression Analysis. Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), March, 2000, pp. 484 – 490.

[12] Yacoob, Y., Davis, L. Computing spatio-temporal representations of human faces. Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on , 21-23 June 1994 Page(s): 70 –75.

[13] Dellaert, F., Polzin, T., Waibel, A. Recognizing emotion in speech. Spoken Language, 1996. ICSLP 96. Proceedings. Fourth International Conference on, Volume: 3, 3-6 Oct. 1996. Pages: 1970 - 1973 vol.3.

[14] Nwe, T. L., Wei, F. S., De Silva, L.C. Speech based emotion classification. Electrical and Electronic Technology, 2001. TENCON. Proceedings of IEEE Region 10 International Conference on, Volume: 1 , 19-22 Aug. 2001. Pages: 297 - 301 vol.1.

[15] Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh A., Busso,C., Deng, Z., Lee, S., Narayanan, S.S. Emotion Recognition based on Phoneme Classes. Proc. ICSLP'04, 2004.

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

[17] 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.

[18] Huang, T. S., Chen, L. S., Tao, H., Miyasato, T., Nakatsu, R. Bimodal Emotion Recognition by Man and Machine. Proceeding of ATR Workshop on Virtual Communication Environments, (Kyoto, Japan), April 1998.

[19] Chen, L.S., Huang, T.S. Emotional expressions in audiovisual human computer interaction. Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on, Volume: 1, 30 July-2 Aug. 2000. Pages: 423 - 426 vol.1.

[20] Ekman, P., Friesen, W. V. Facial Action Coding System: A Technique for Measurement of Facial Movement. Consulting Psychologists Press Palo Alto, California, 1978.

[21] Burges, C. A tutorial on support vector machines for pattern recognition. Dat Mining and Know. Disc., vol. 2(2), pp. 1–47, 1998.

[22] Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Deng, Z., Busso, C., Lee, S., Narayanan, S.S., Analysis of acoustic correlates in emotional speech. in ICSLP'04, 2004.