Emotion Recognition from Decision Level Fusion of Visual and Acoustic Features using Hausdorff Classifier

H.D.Vankayallapati[1], K.R.Anne[2], and K. Kyamakya[1]

[1]Institute of Smart System Technologies, Transportation Informatics Group
University of Klagenfurt, Klagenfurt, Austria.
[2]Department of Information Technology, TIFAC-CORE in Telematics
VR Siddhartha Engineering College, Vijayawada, India.
raoanne@gmail.com

Abstract. The emotions are generally measured by analyzing either head movement patterns or eyelid movements or face expressions or all the lasts together. Concerning emotion recognition visual sensing of face expressions is helpful but generally not always sufficient. Therefore, one needs additional information that can be collected in a non-intrusive manner in order to increase the robustness. We find acoustic information to be appropriate, provided the human generates some vocal signals by speaking, shouting, crying, etc. In this paper, appropriate visual and acoustic features of the driver are identified based on the experimental analysis. For visual and acoustic features, Linear Discriminant Analysis (LDA) technique is used for dimensionality reduction and Hausdorff distance is used for emotion classification. The performance is evaluated by using the Vera am Mittag (VAM) emotional recognition database. We propose a decision level fusion technique, to fuse the combination of visual sensing of face expressions and pattern recognition from voice. The result of the proposed approach is evaluated over the VAM database with various conditions.

Key words: Driver Monitoring System, Acoustic features, Visual features, Hausdorff distance

1 Introduction

In the past years, research related to emotion recognition has been done in psychology and physiology. Here physiological characteristics of the person such as heart rate, brain activity, pupil size, skin conductance and production of stress hormones, and pulse rate are measured and interpreted for inferring the person state or emotion. These characteristics are measured in intrusive way by involving measurement systems such as the Electroencephalogram (EEG) which monitors brain activities [1, 2], the Electrocardiogram (ECG) which measures heart rate variation, the Electrooculogram (EOG) which monitors eye movement, the skin potential level measurement techniques, etc [3–5]. These systems need the person's cooperation as the electrodes are attached directly to the person's body. So non-intrusive techniques are introduced. These non-intrusive techniques uses cameras or sensors which are place before the person to measure the head, eye lid movements and face expressions etc. To measure these characteristics, several techniques are available in the literature.

Usually a monitoring system takes the sensor's input sample from the person, extracts features and compares them with the template of the existing emotions to find the best match. This match explains how well the extracted features from the sample match a given template. There has also been a similar increase in use of multimodal techniques to overcome the difficulties of single modal system and for performance improvement [6]. Several real world applications require higher performance than just one single measure to improve road safety. In multimodal techniques, the data collected from different sensors are integrated at different levels are explained in Section. 3. One of the important benefits in multimodal is if one input is highly noisy, then the other input might be helpful to make an overall reliable decision. Most recently, the study concerning the recognizing the emotional state of the person from the visual and acoustic information of the human has been introduced.

In this work also, we considered visual and acoustic information of the person. The visual features (such as global features) and acoustic features (such as pitch, zero crossing rate (ZCR), short time energy (STE) and Mel Frequency Cepstral Coefficients (MFCC) [7–9] are selected based on the experimental validation performed) are extracted based on our emotion recognition application.

Each person has many number of emotions. Here emotions like Happy, Anger, Sad, Disgust and Neutral are considered. The Linear Discriminant Analysis (LDA) is used for visual and acoustic feature reduction [10]. The emotions from the visual and acoustic features are classified separately based on the Hausdorff classifier is explained briefly in Section. 2. The emotion for visual information and emotion from acoustic information are recognized separately. Then we fused these two decisions/emotions based on the weighted majority voting rule in decision fusion explained briefly in Section. 4. We recognize the final emotion such as sad, angry, happy, disgust and neutral. The Vera Am Mittag (VAM) emotional database is used to evaluate the performance of the decision level fusing using Hausdorff classifier.

2 Hausdorff distance classifier

The Hausdorff Classifier is a non-linear operator, which measures the mismatch of the two sets [11]. The Hausdorff classifier measures the extent to which each point of a 'model' set lies near some point of an 'sample' set and vice versa. Unlike most vector comparison methods, the Hausdorff classifier is not based on finding corresponding mode and speech points. Thus, it is more tolerant of perturbations in the location of points because it measures proximity rather than exact superposition [12, 13]. However, the Hausdorff distance is extremely sensitive to outliers.

The distance between two points a and b is defined as shown in Equation (1).

$$d(a, b) = ||a - b|| \tag{1}$$

Here, we not only compute the distance between the point a in the finite point set A and the same value b in the finite point set $B = b_1, ..., b_{N_b}$, but also compute the distances between the a_t and its two neighbor values b_{t-1} and b_{t+1} in the finite point set B, respectively, and then minimize these three distances as shown in in Equation (2) [11].

$$d(a, B) = \min_{b \in B} d(a, b) = \min_{b \in B} ||a - b|| \tag{2}$$

The directed Hausdorff metric h(A,B) between the two finite point set $A = a_1, ..., a_{N_a}$ and $B = b_1, ..., b_{N_b}$is defined in Equation (3),(4) :

$$h(A, B) = \max_{a \in A} d(a, B) = \max_{a \in A} \min_{b \in B} d(a, b) \tag{3}$$

$$h(A, B) = \left\{ \max_{a \in A} \left\{ \min_{b \in B} ||a - b|| \right\} \right\} \tag{4}$$

3 Multimodal Fusion

Generally Fusions are of different levels, they are sensor fusion or data fusion, feature fusion, decision fusion. For emotion recognition of the person, we use multimodal system using both visual and acoustic sensors means camera and microphone respectively. In Data fusion, raw data from the multiple sensors are fused to generate the new data from which features are extracted. In visual information, different features are extracted and acoustic information different features are extracted [6]. This is the major limitation which, the fusion at sensor level is not possible in our work. Data fusion is possible only for same types of sensors (either two cameras or two microphone data).

Feature fusion means fusion after feature extraction. In this level we are integrating the feature vectors (extracted from the input sample) from multiple biometric sensors. If two different samples are of the same types (two samples of the same face), then it is possible to combine and create a new and more reliable feature vector from the two vectors. However, if the samples are of different types (face and voice data), the feature vectors can be concatenated into a new and more detailed feature vector.

Fusion at feature level is difficult in practice because of the following reasons [6]: (i) the feature sets of multiple modalities may be incompatible (e.g., eigen-coefficients of face and Mel-frequency cepstral coefficients (MFCCs) of voice); (ii) the relationship between the feature vectors of different biometric sensors may not be known; and (iii) concatenating two feature vectors may result in a feature vector with very large dimensionality leading to the 'curse of dimensionality' problem [14].

4 Decision level Fusion

For decision level fusion, we can use either multiple samples for the same of type of sensors or multiple sample from different types of sensors. Here multiple sensor information namely visual and acoustic are processed independently and their decisions are fused. Fig. 1 shows the fusion at decision level.

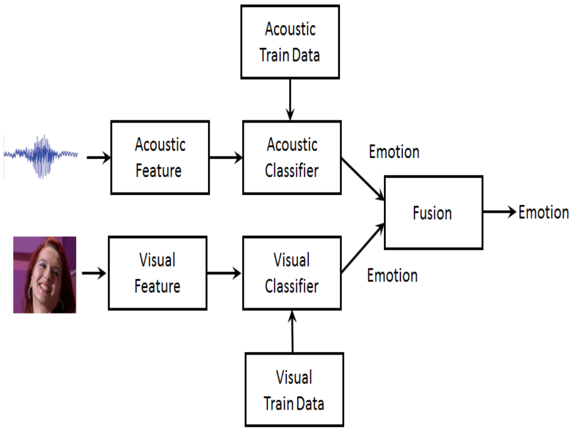

Fig. 1. Schematic description of fusion at decision level

In multimodal verification system, "AND" and "OR" rule is a simplest method to combine the decision output of different multimodal subsystems. The output of "AND" rule is the match when the both subsystems input samples matches with the train data template. By using "AND", the False Acceptance Rate (FAR) is lower than the FAR of individual subsystem [6]. The output of "OR" rule is the match of which one of the subsystem input sample matches the train data template. By using "OR", the False Reject Rate (FRR) is greater than the FRR of individual subsystem [6]. But this "OR" rule is not possible in emotion recognition applications. Because if the two subsystems has different emotions, if we apply "OR" rule, we can't decide which emotion is the final output. And another problem with this "AND" and "OR" rule is, in real world scenario, acoustic information is present in bursts based on the emotion of the person, where as visual information is present throughout the processing. Because of this limitation, we have not considered "AND" and "OR" rule for human emotion recognition.

The most common and simplest rule derived from the sum rule for decision level fusion is majority voting rule. In multimodal verification system, input samples are assigned to the subsystems and identifies the majority of the subsystems agrees the match or not. If the input samples are R, then atleast k are identified as matched then final output of the decision level is "match". k is shown in Equation (5). Atleast k matchers should agree that identity.

$$k = \{ \begin{array}{l} \frac{R}{2} + 1 \text{ if R is even} \\ \frac{R+1}{2} \quad \text{otherwise.} \end{array} \tag{5}$$

The major drawback in majority voting is all the subsystems are treated or weighted equally. In our emotion recognition, majority of the emotion obtain to either from acoustic subsystem or visual subsystem is the final output. But in real cases, it is not correct. Visual is more reliable than acoustic. To overcome this limitation weighted majority voting rule is used in this work.

In weighted majority voting rule, the weights w_k are assigned based on the reliability (during training process) and k is the number of times the subsystem is matched.

$$s_{j,k} = \{ \begin{array}{l} 1 \text{ if output of the } j^{th} \text{ mathcer is in class } w_k \\ 0 \text{ otherwise.} \end{array} \tag{6}$$

The discriminant function for class w_k computed using weighted voting is shown in Equation (7).

$$g_k = \sum_{j=1}^{R} w_j s_{j,k} \tag{7}$$

Where w_j is the weight assigned to the j^{th} matcher.

Emotions can be classified into different classes like anger, happiness, disgust or sadness. Neutral emotion means no expression/emotion is present. In this work, we classify different expression based on distance from the neutral as shown in Fig. 2.

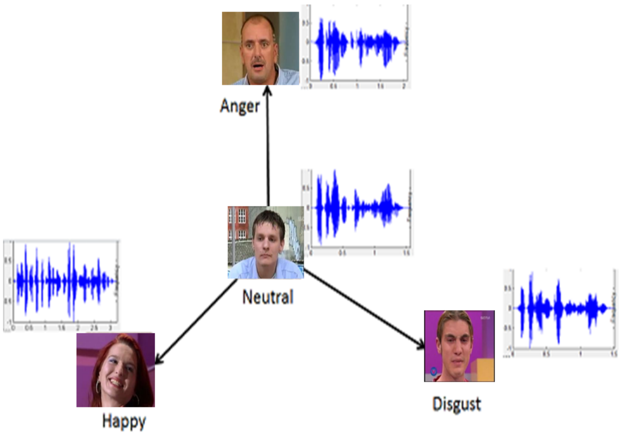

Fig. 2. Illustration of fusion based emotion classification

To validate the performance, the probability for each of the emotions was calculated for the audio and visual features separately and by applying weighted majority voting to get the final result as shown in Fig. 2.

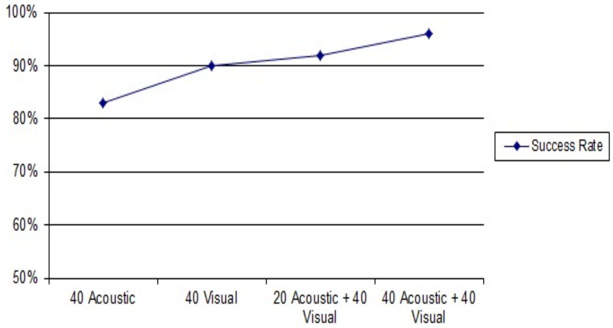

Fig. 3. Performance comparison of different visual-acoustic feature fusion at decision level

In order to evaluate the recognition algorithm with fused features, we have used the Vera am Mittag (VAM) emotional database. Performance comparison of different visual-acoustic feature fusion at decision level for emotion recognition application is shown in Fig. 3.

5 VAM German emotional speech Database

The Vera Am Mittag (VAM) is the audio visual emotional database consists of 12 hours of audio-visual recordings of the German TV talk show named as Vera am Mittag (in English "Vera at noon"), segmented into broadcasts, dialogue acts and utterances. The title of the database is taken from this TV show title. Because there is a reasonable amount of speech from the same speakers available in each session of the TV show [15]. The Sat.1 german tv channel broadcasted these recorded shows between December 2004 to February 2005. Speakers are at the age of 16 to 69 years at the time of recording.

VAM database has 3 different parts. They are VAM-Video database, VAM-Audio database and VAM-Faces Database as shown in Fig. 4 and Fig. 5 respectively. VAM-Video database contains 1421 videos of the 104 speakers. VAM-Faces database is extracted from the VAM-video by taking each frame as one still image. This database classified different emotions as anger (A), happiness (H), fear (F), disgust (D), sadness(Sa), surprise (Su) and neutral (N). VAM-Faces database contains 1867 images by 20 speakers.

VAM-Audio database contains 1018 emotional utterances by 47 speakers [15]. It mainly contains the complete sentences but sometime also incomplete sentences. Some examples of the dialogue in this database are

Fig. 4. Sample images in Vera am Mittag (VAM)-Faces database [15]

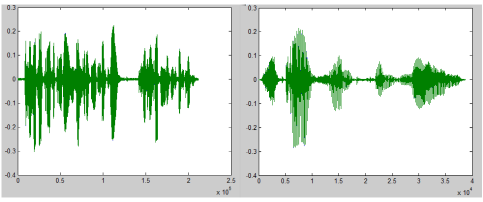

Fig. 5. Sample speech in Vera am Mittag (VAM)-Audio database [15]

- Melanie: "She said, she would leave him to me, and she would even help me." serious
- Kristin: "Melanie, I do know very well what I told you on the boat!" angry

6 Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) graphs are useful for analyzing the classifier performance. The ROC curve is a plot of the True Positive Fraction (TPF) versus the False Positive Fraction (FPF). It compares the two operating characteristics (TPF, FPF). So ROC is also called as Relative Operating Characteristic curve. These TPF and FPF are calculated by using True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) as shown in Equation (8),(9). Sensitivity is the true positive fraction, expressed as a percentage. Specificity is the true negative fraction, expressed as a percentage.

In Fig. 6 is the example roc curve, Shadow area (ROC space) gives the better classification [16].

$$TPF = \frac{TP}{TP + FN} \tag{8}$$

$$FPF = \frac{FP}{FP + TN} \tag{9}$$

$$AUC = \frac{TP + TN}{TP + FN + FP + TN} \tag{10}$$

Fig. 7 demonstrates the recognition performance of LDA based Hausdorff distance classifier on the Vera am Mittag (VAM) audio visual emotional database

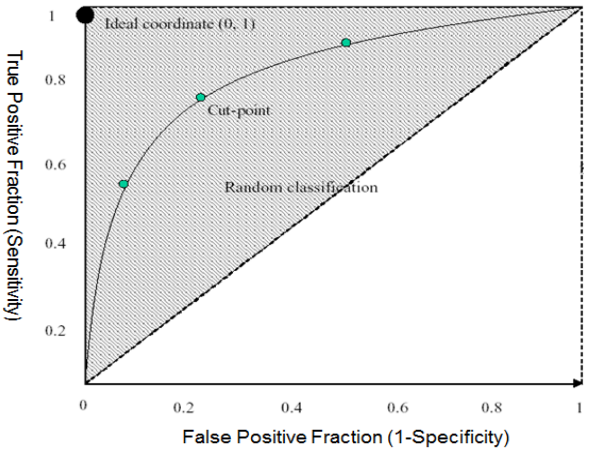

Fig. 6. Receiver Operating Characteristic (ROC) curve [16]

in which their ROC curves of False Positive Fraction (FPF) and True Positive Fraction (TPF) are shown. VAM Database has emotions like anger (A), happiness (H), disgust (D), sadness(Sa), and neutral (N). In the ordinary data format, each line represents one classifier case and each line has two numbers. In that the first number is either truly negative "0" or truly positive "1" [17]. The second number represents the output level with each case. In our emotion recognition algorithm, we analyze the hausdorff classifier performance by considering the different emotions in the VAM database. For this case, I used 5-point rating scale; the categories would have the following meaning:

- 1 - Definitely negative
- 2 - Probably negative
- 3 - Possibly negative
- 4 - Possibly positive
- 5 - Definitely positive

Each case has a rating from 1 to 5. If the first number in the case "1", a rating of 3 or greater than 3 is considered positive or correct. Remaining are treated as negative or wrong. If the first number is "0" then rating is in reverse order.

7 Conclusion

Multimodal emotion recognition system will be useful to understand the state and emotion of a person. In this work, we came to know that the acoustic information together with the visual information significantly increases the recognition rate. However, in real world scenario acoustic information is available only in burst. A good amount of stress is made in locking the audio burst and in

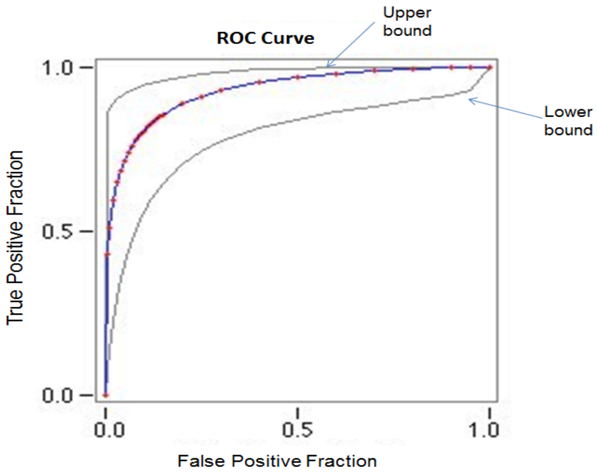

Fig. 7. Receiver Operating Characteristic (ROC) curve for emotion recognition

correlating the information at decision level with visual information. Also in this work, we gave weightage to features that predominantly have a role in emotion and omitted other minor features which may have role in recognizing the emotion in order to increase the performance in real time. In this work, we emphasized on performing the evaluation over different databases to check the robustness of the algorithms and to see the scalability of the algorithms.

References

1. A. C. Saroj K.L. Lal, Peter Boord, Les Kirkup, Hung Nguyen: Development of an algorithm for an EEG-based driver fatigue countermeasure, Journal of Safety Research, vol. 34, pp. 321- 328 (2003).
2. K. H. Roman Bittner, Lubomir Pousek, Pavel Smrcka, Petr Schreib and Petr Vysok: Detecting of Fatigue States of a Car Driver, vol. 1933: Springer-Verlag, London, Uk (2000).
3. A. Broggi: Vision-based driving assistance, IEEE Intelligent Transport Systems,vol 13,pages 22-23 (1998).
4. H. D. Vankayalapati and K. Kyamakya: Nonlinear feature extraction approaches for scalable face recognition applications, In ISAST Transactions on Computers and Intelligent Systems, volume 2 (2010).
5. H. D. Vankayalapati, K. Anne and K. Kyamakya: Extraction of visual and acoustic features of the driver for monitoring driver ergonomics applied to extended driver assistance systems, volume 81, pages 83-94, Springer Berlin / Heidelberg (2010).
6. Arun A. Ross, Karthik Nandakumar, Anil K. Jain: Handbook of Multibiometrics, Springer-Verlag New York, USA (2006).
7. Dimitrios Ververidis and Constantine Kotropoulos: Emotional speech recognition: Resources, features, and methods, Speech Communication, 48(9):1162-1181 (2006).

8. J. Arnott, I. Murray: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, Journal of the Acoustical Society of America, 93(2):1097-1108 (1993).
9. Thurid Vogt, Elisabeth Andr and Johannes Wagner: Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation, pages 75-91 (2008).
10. J. P. Hespanha P. N. Belhumeur and D. J. Kriegman: Eigenfaces vs. Fisherfaces:Recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19:711-720 (1997).
11. Oliver Jesorsky, Klaus J. Kirchberg, and Robert W. Frischholz: Robust face detection using the hausdorff distance, In Proc. Third International Conference on Audio- and Video-based Biometric Person Authentication, Springer, LNCS-2091, pages 9095, Halmstad, Sweden, 68 June (2001).
12. T. Barbu: Discrete speech recognition using a hausdorff-based metric, In Proceedings of the 1st Int. Conference of E-Business and Telecommunication Networks, ICETE 2004, volume 3, pages 363-368, Setubal, Portugal, Aug (2004).
13. M. P. Dubuisson, A. K. Jain: Modified Hausdorff distance for object matching, Proc. of IAPR International Conference on Pattern Recognition (ICPR'94, Jerusalem, IS), pages 566-568 (1994).
14. Arun A. Ross, Rohin Govindarajan: Feature Level Fusion in Biometric Systems, in proceedings of Biometric Consortium Conference (BCC), September (2004).
15. Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan: The Vera am Mittag German Audio-Visual Emotional Speech Database, In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hannover, Germany (2008).
16. Wen Zhu, Nancy Zeng, Ning Wang: Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations, Health Care and Life Sciences, NESUG (2010).
17. John Eng: ROC analysis: web-based calculator for ROC curves, Baltimore, Johns Hopkins University [updated 2006 May 17. Available from `http://www.jrocfit.org`.