A Comparison of Facial Features and Fusion Methods for Emotion Recognition

Demiyan V. Smirnov[1], Rajani Muraleedharan[2], and Ravi P. Ramachandran[1](✉)

[1] Rowan University, Glassboro, NJ, USA
smirno59@students.rowan.edu, ravi@rowan.edu
[2] Saginaw Valley State University, University Center, MI, USA
rmuralee@svsu.edu

Abstract. Emotion recognition is an important part of human behavior analysis. It finds many applications including human-computer interaction, driver safety, health care, stress detection, psychological analysis, forensics, law enforcement and customer care. The focus of this paper is to use a pattern recognition framework based on facial expression features and two classifiers (linear discriminant analysis and *k*-nearest neighbor) for emotion recognition. The extended Cohn-Kanade database is used to classify 5 emotions, namely, 'neutral, angry, disgust, happy, and surprise'. The Discrete Cosine Transform (DCT), Discrete Sine Transform (DST), the Walsh-Hadamard Transform (FWHT) and a new 7-dimensional feature based on condensing the Facial Action Coding System (FACS) are compared. Ensemble systems using decision level, score fusion and Borda count are also studied. Fusion of the four features leads to slightly more than a 90 % accuracy.

Keywords: Emotion recognition · Facial expression · Feature extraction · Linear discriminant analysis · K-nearest neighbor · Fusion

1 Introduction and Motivation

Emotion recognition is an important part of human behavior analysis, where applications including human-computer interaction (particularly the brain-computer interface) [1], human-robot interaction [2], computer games [3]. Driver safety [3,4], health care [5], stress detection [6] can be benefited. In addition, emotion recognition when combined with biometrics [7] is crucial for security purposes and improves law enforcement and forensic applications [8]. Humans express emotion using vocal, facial, gesture, body language, handwriting, and sign language. Identifying emotions through contents in an electronic conversation (like email) can maximize customer satisfaction, and is currently an important topic of research in marketing and customer service [9].

Face [2,3,10], speech [11,12] and physiological modalities are often used in emotion recognition. The use of physiological signals include the electrocardiogram (ECG) [13], electroencephalogram (EEG) [1], skin temperature and resistance, blood pressure and respiration [1]. Physiological modalities are highly

S. Arik et al. (Eds.): ICONIP 2015, Part IV, LNCS 9492, pp. 574–582, 2015.
DOI: 10.1007/978-3-319-26561-2_68

intrusive and cannot be measured remotely. However, speech or face signals can be acquired through remote audio and video surveillance, which is non-invasive and requires less cooperation from users.

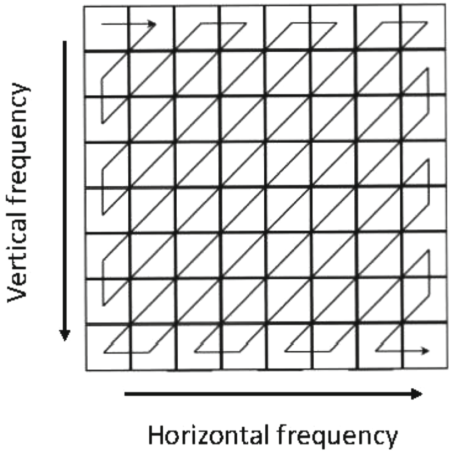

Fig. 1. Zig-zag scanning method for extracting a one-dimensional feature vector from a 2-D frequency transform

This paper compares the performance of various facial expression feature extraction and machine learning approaches for emotion recognition. The extended Cohn-Kanade database [4] is used to classify 5 emotions, namely, 'neutral, angry, disgust, happy, and surprise'. The main assumptions are (1) that the subject can be captured by at least one camera sensor to yield a digital image, (2) the subject is cooperative and the quality of the image is improved to remove illumination and noise effects and (3) biometric recognition of the face image is useful but not required for emotion identification (any biometric analysis would be in parallel with the system proposed in this paper). A pattern recognition [14,15] framework is used involving feature extraction, classification and a jackknife (or m-fold) strategy [14,16] with multiple trials for performance evaluation. The classifiers considered are linear discriminant analysis (LDA), and the k-nearest neighbor (kNN). Three of the facial expression features are the Discrete Cosine Transform (DCT), Discrete Sine Transform (DST) and the Walsh-Hadamard Transform (FWHT) [17]. A new 7-dimensional feature (recently proposed in [18]) based on condensing the Facial Action Coding System (FACS) [4] using 14 points is also analyzed. Ensemble systems using decision level, score fusion and Borda count are studied.

2 Facial Expression Feature Extraction

Each facial image in the database is first processed by the Viola-Jones face detector [7,19] and any non-face background portions of the image are removed.

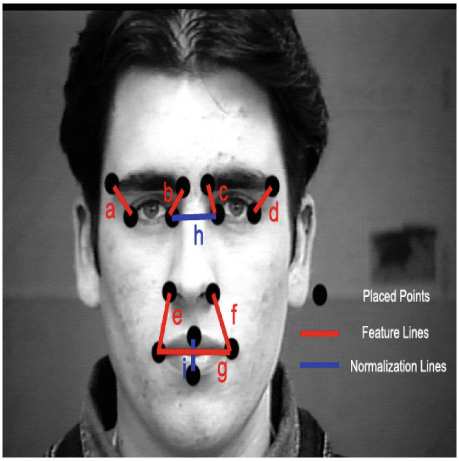

Fig. 2. The 14 point feature extraction method

Given a background removed face image, the 2-D DCT, 2-D DST and 2-D FWHT are calculated and scanned in a zig-zag fashion [17] as shown in Fig. 1 to get a one-dimensional feature vector.

Figure 2 shows a set of fourteen points, placed on a subject's face, based on the Facial Action Coding System (FACS) [4]. The FACS is used to quantify emotions by labeling muscle movements and facial feature changes by an Action Unit number [4]. The fourteen points and nine distances between the points are labeled as 'a' through 'i' in Fig. 2. The feature vector consists of 7 dimensions. The first six are the distances 'a' through 'f' each divided by the distance 'h' (accounts for image scaling and rotation). The seventh dimension is the horizontal distance at the mouth (labeled as 'g') divided by the vertical distance at the mouth (labeled as 'i'). This 7 dimensional feature vector is also is referred to as the 14 point feature extraction method.

The motivation of using this 14 point feature extraction method is that (1) the need to determine which Action Units a subject is/are displaying is avoided and (2) a vector of low dimension is configured by using normalized distances between key points corresponding to important facial attributes that indicate emotion. The specific points and distances were chosen as they represent movements of important muscles. The inner and outer eyebrows are accounted for, both when they are raised or lowered. In the same vein, the corners of the mouth produce longer lines when they are pulled down for a frown and shorter lines when pulled up for a smile. The status of the mouth (open or closed) can also be detected by the lines crossing on the lips.

3 Emotion Recognition Classifiers

Classifiers like neural networks and Gaussian mixture models require much data to obtain a good performance. In the extended Cohn-Kanade database, only the

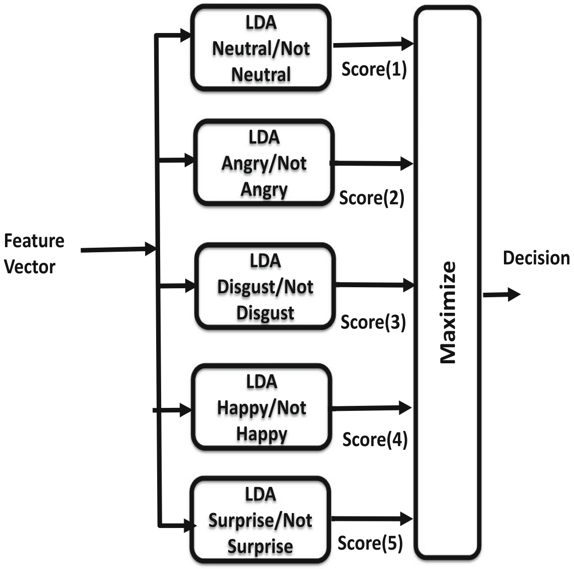

Fig. 3. LDAPA: Parallel arrangement of LDA classifiers

'neutral' emotion has much data (593 images). This is followed by 'happy' (69 images). The 'anger' emotion is only depicted by 45 images. In lieu of the limited amount of data, two simple classifiers, namely, linear discriminant analysis (LDA) and the k-nearest neighbor (kNN) are considered.

The LDA method is performed without dimensionality reduction. It is implemented in two ways. First, the LDA is trained using feature vectors from all 5 class labels (emotions) so that the feature space is partitioned into 5 distinct regions. It is assumed that the feature vectors for each of the five classes has a multivariate Gaussian density. For a linear discriminant function, each class has a different mean vector but the same covariance matrix. Each test image results in one test feature vector. The output of the LDA consists of 5 scores, each indicating the posterior probability that the test feature vector belongs to a particular 'emotion' class. The maximum score identifies the emotion.

The second LDA implementation is depicted in Fig. 3. The test feature vector is passed through a parallel arrangement of 5 LDA classifiers with each making a decision between two classes (emotion and not the emotion). This is referred to as LDA with a parallel arrangement (LDAPA). Each LDA classifier is trained using feature vectors representing the 'emotion' (like neutral) and feature vectors representing 'not the emotion' (like all emotions except neutral). Again, it is assumed that the feature vectors for each of the two classes has a multivariate Gaussian density. As before, each class has a different mean vector but the same covariance matrix. For a two class problem, a hyperplane divides the feature space into two regions that distinguish each class. The only score considered for each LDA classifier is the posterior probability that the test feature vector belongs to the ‘emotion’ class. The maximum score among the 5 classifiers identifies the emotion.

The kNN classifier stores all the training data and the associated class labels for the 5 emotions. The class labels of the k nearest neighbors to the test feature vector (in terms of squared Euclidean distance) are recorded. A majority vote among the class labels identifies the emotion. In the event of a tie, only the class labels involved in the tie are considered with the class label of the training vector closest to the test feature vector identifying the emotion.

4 Experimental Protocol

A jackknife (or m-fold strategy with $m = 7$) is used to randomly partition the image data into 7 non-overlapping subsets such that the amount of data for each emotion is about the same in each subset. Six of the subsets are used for training the classifier and one subset is used for testing. The subsets are revolving with each test so that each subset is used as the testing subset exactly once. Hence, seven test runs are performed for this specific partition. The Average Class Identification Rate (ACIR) is the number of times a test feature vector from a particular class is identified correctly with an average taken over the 7 test runs. There will be 5 ACIR values, one for each class or emotion. This process depicts one trial.

There are five trials performed, each with a different random partition of the image data into 7 subsets. The Average Trial Class Identification Rate (ATCIR) is the average of the ACIR values taken over the five trials. Again, there will be 5 ATCIR values, one for each class or emotion. The identification success rate (ISR) is the average of the ATCIR values taken over the 5 emotions.

Table 1. Identification success rate for individual features

Feature	Classifier	No SMOTE		With SMOTE	
		Best dimension	ISR (%)	Best dimension	ISR (%)
DCT	LDA	115	88.6	120	84.6
DCT	LDAPA	145	85.1	115	81.9
DCT	kNN	130	28.4	85	35.1
DST	LDA	120	88.1	95	85.2
DST	LDAPA	150	85.1	115	83.7
DST	kNN	85	28.2	130	35.8
FWHT	LDA	120	88.9	115	85.4
FWHT	LDAPA	150	86.0	115	82.5
FWHT	kNN	130	31.5	145	34.2
14 Point	LDA	7	49.9	7	63.3
14 Point	LDAPA	7	40.6	7	60.7
14 Point	kNN	7	58.1	7	61.6

The four features were each used with LDA, LDAPA and kNN to get an ISR value. The ISR values were obtained in two ways. The first was by using the unbalanced data in which there were many more samples of the 'neutral' emotion (593 feature vectors). The second was by using the Synthetic Minority Oversampling Technique (SMOTE) [20] to increase the number of feature vectors from each class (except 'neutral') such that the data is balanced.

5 Performance of the Individual Features

Table 1 shows the ISR for the various feature/classifier combinations with and without SMOTE. The best dimension given in Table 1 is that which results in the maximum ISR for the cases when SMOTE is used and not used. For the DCT, DST and FWHT (with and without SMOTE), the performance was evaluated for dimensions ranging from 5 to 150. Figure 4 depicts the results for the case when the LDA classifier is used without SMOTE. For the kNN classifier, values of k equal to 1, 3, 5 and 7 were attempted and $k = 1$ gave the best performance.

The DCT, DST and FWHT show the best results using LDA or LDAPA. A one-tailed t-test with unequal variances [21] based on five trials and a 95 % confidence interval confirms the following for the DCT, DST and FWHT:

1. There is no statistical significant difference in performance a, omg the three features.
2. The LDA shows the best performance with statistical significance.
3. Not applying SMOTE shows a better performance with statistical significance.

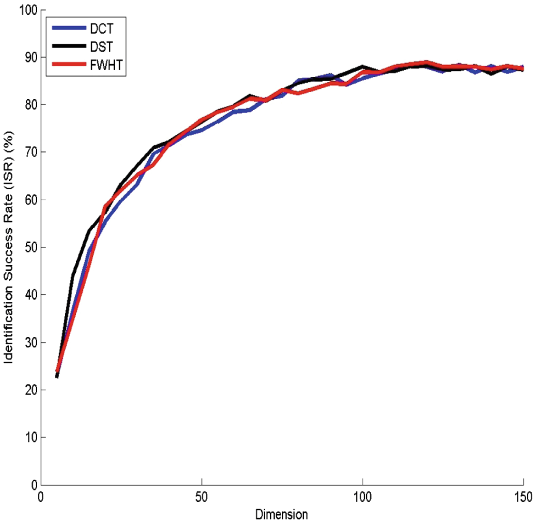

Fig. 4. ISR of DCT, DST and FWHT as a function of vector dimension for the LDA classifier (SMOTE not used)

The dimension of the vector resulting from the 14 point feature extraction method is fixed at 7. The performance of this feature is better when SMOTE is applied (with statistical significance). However, the performance is not as good as the DCT, DST or FWHT implemented with LDA or LDAPA. However, to achieve this better performance, the DCT, DST and FWHT require a much higher dimension. Future work is aimed at using more than fourteen points (based on the (FACS)), acquiring feature vectors of different dimensions (by labeling different points and taking distances between points as in Fig. 2) and investigating the ISR versus dimension. This will give a more clear comparison with the DCT, DST and FWHT. The aim is to get the high performance that DCT, DST and FWHT achieve but at a lower dimension that an FACS based method can potentially achieve.

6 Fusion

Since different feature/classifier combinations are used, an ensemble system [16] results, which naturally leads to the investigation of fusion. Decision level fusion is the simplest technique and involves taking a majority vote of the different features to get a final decision. For score fusion, the scores (or posterior probabilities) for each emotion of a single feature/classifier pair are converted to normalized scores such that their sum equals 1. For a particular emotion, the normalized scores generated by the different feature/classifier pairs considered are added to get a combined score. The maximum combined score identifies the emotion. The third fusion method is to use Borda count based on the normalized scores.

Table 2. Fusion results (Numbers expressed as a %)

Features and Classifiers	Fusion type	ISR (%) No SMOTE	ISR (%) SMOTE	ISR (%) Partial SMOTE
DCT/LDA and DST/LDA	Score	88.3	85.4	88.9
DCT/LDA and DST/LDA	Borda	89.1	82.8	89.8
DCT/LDA, DST/LDA and FWHT/LDA	Score	89.3	86.1	89.7
DCT/LDA, DST/LDA and FWHT/LDA	Borda	89.2	82.8	89.9
DCT/LDA, DST/LDA and FHWT/LDA	Decision	89.2	86.3	89.9
DCT/LDA, DST/LDA and 14 Point/kNN	Decision	88.8	85.5	90.4
DCT/LDA, DST/LDA, FWHT/LDA and 14 Point/kNN	Decision	89.7	86.6	90.9

Fusion experiments were performed using DCT/LDA, DST/LDA, FWHT/LDA and the 14 point feature extraction method with kNN. In the event of a tie due to fusion, the decision of the DCT/LDA is taken. Table 2 gives the results of the best approaches for the following cases:

1. No SMOTE: SMOTE not used for any of the features
2. SMOTE: SMOTE used for all of the features
3. Partial SMOTE: SMOTE not used for the DCT, DST and FWHT but used for the 14 point feature extraction method.

The best method is the decision level fusion of DCT/LDA, DST/LDA, FWHT/LDA and the 7 dimensional feature resulting from the 14 point feature extraction method (denoted as 14 Point/kNN). For this fusion method, partial SMOTE is the best (with statistical significance). This illustrates that the 7 dimensional feature is useful and should be explored further.

7 Summary and Conclusions

Various feature/classifier combinations have been compared for emotion recognition using facial features. Five emotions, namely, ‘neutral, angry, disgust, happy, and surprise’, from the extended Cohn-Kanade database are used in the experiments. Fusion of the features results in slightly more than a 90 % accuracy. Future work will aim to achieve a high performance with a low feature dimension.

Acknowledgement. This work was supported by the National Science Foundation through Grant DUE-1122296.

References

1. Huang, D., Zhang, H., Ang, K., Guan, C., Pan, Y., Wang, C., Yu, J.: Fast emotion detection from EEG using asymmetric spatial filtering. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Kyoto, Japan, pp. 589–592 (2012)
2. Tariq, U., Huang, T.S.: Features and fusion for expression recognition - A comparative analysis. In: IEEE Computer Vision and Pattern Recognition Workshop, pp. 146–152 (2012)
3. Tsai, H.-H., Lai, Y.-S., Zhang, Y.-C.: Using SVM to design facial expression recognition for shape and texture features. In: International Conference on Machine Learning and Cybernetics, Qingdao, China, pp. 2697–2704 (2010)
4. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: IEEE Computer Vision and Pattern Recognition Workshop, pp. 94–101 (2010)
5. Lucey, P., Cohn, J. F., Lucey, S., Matthews, I., Sridharan, S., Prkachin, K.: Automatically detecting pain using facial actions. In: International Conference on Affective Computing and Intelligent Interaction, pp. 1–8 (2009)

6. Nehra, D.K., Sharma, V., Mushtaq, H., Sharma, N., Sharma, M., Nehra, S.: Emotional intelligence and self esteem in cannabis abusers. J. Indian Acad. Appl. Psychol. **38**(2), 385–393 (2012)
7. Jain, A.K., Ross A., Nandakumar K.: Introduction to Biometrics. Springer, New York (2011)
8. Campbell, J.P., Shen, W., Campbell, W.M., Schwartz, R., Bonastre, J.-F., Matrouf, D.: Forensic speaker recognition. IEEE Signal Process. Mag. **26**, 95–103 (2009)
9. Gupta, N., Gilbert, M., Di Fabrizio, G.: Emotion detection in email customer care. Comput. Intell. 59 (2010)
10. Gupta, S.K., Agrwal, S., Meena Y.K., Nain, N.: A hybrid method of feature extraction for facial expression recognition. In: International Conference on Signal Image Technology and Internet-Based Systems, pp. 422–425 (2011)
11. Koolagudi, S.G., Kumar, N., Rao, K.S.: Speech emotion recognition using segmental level prosodic analysis. In: International Conference on Devices and Communications, pp. 1–5 (2011)
12. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal Information. In: ACM International Conference on Multimodal Interfaces, State College, Pennsylvania (2004)
13. Cai, J., Liu, G., Hao, M.: The research on emotion recognition from ECG Signal. In: International Conference on Information Technology and Computer Science, pp. 497–500 (2009)
14. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, New York (2001)
15. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
16. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits Syst. Mag. **6**, 21–45 (2006)
17. Bose, T.: Digital Signal and Image Processing. Wiley, New York (2004)
18. Smirnov, D.V., Banger, S., Davis, S.H., Muraleedharan, R., Ramachandran, R.P.: Automated human behavioral analysis framework using facial feature extraction and machine learning. In: 47th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California (2013)
19. Viola, P.A., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision **57**, 137–154 (2004)
20. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. **16**, 321–357 (2002)
21. Yates, R.D., Goodman, D.J.: Probability and Stochastic Processes. Wiley, New York (1999)