Human Facial Expression Recognition Using Wavelet Transform and Hidden Markov Model

Muhammad Hameed Siddiqi and Sungyoung Lee*

Department of Computer Engineering, Kyung Hee University
(Global Campus), Suwon, Rep. of Korea
{siddiqi,sylee}@oslab.khu.ac.kr

Abstract. The accuracy of the Facial Expression Recognition (FER) system is completely reliant on the extraction of the informative features. In this work, a new feature extraction method is proposed that has the capability to extract the most prominent features from the human face. The proposed technique has been tested and validated in order to achieve the best accuracy for FER systems. There are some regions in the face that have much contribution in achieving the best accuracy. Therefore, in this work, the human face is divided into number of regions and in each region the movement of pixels have been traced. For this purpose, one of the wavelet families named symlet wavelet is used and individual facial frame is decomposed up to 2 levels. In each decomposition level, the distances between the pixels is found by using the distance formula and by this way some of the informative coefficients are extracted and hence the feature vector has been created. Moreover, the dimension of the feature space is reduced by employing a well-known statistical technique such as Linear Discriminant Analysis (LDA). Finally, Hidden Markov Model (HMM) is exploited for training and testing the system in order to label the expressions. The proposed FER system has been tested and validated on Cohn-Kanade dataset. The resulting recognition accuracy of 94% illustrates the success of employing the proposed technique for FER.

Keywords: Facial Expression Recognition, Wavelet Transform, LDA, HMM.

1 Introduction

Facial expression recognition (FER) plays a significant role in daily life communication. In daily life, various types of communication are utilized for human-to-human interactions: for instance, verbal and non-verbal communication, mental states, and physiological activities [6]. Among these, verbal communication (such

* This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2013-(H0301-13-2001)).

C. Nugent, A. Coronato, and J. Bravo (Eds.): IWAAL 2013, LNCS 8277, pp. 112–119, 2013.
© Springer International Publishing Switzerland 2013

as speech) and non-verbal communication (such as facial expressions) [14] are most often employed. According to [14], during a face-to-face communication, the feelings of a person (such as like or dislike) depend just 7% on the spoken words, 38% on voice intonation, and an incredible 55% on facial expressions.

Generally, FER system consists of three basic modules: preprocessing, feature extraction and recognition. So far, there lots of works have been done for preprocessing including automatic face detection and for recognition modules. Most of the facial features are very sensitive regarding to noise and illumination and also there is very slight change in the facial pixels intensity, however, very limited work can be found for feature extraction in the literature.

Some of the previous works including [9, 12, 13] employed a well-known statistical technique like Principal Component Analysis (PCA) for facial feature extraction. However, PCA focuses only the global features and moreover computational wise PCA is much expensive [7]. In order to achieve high recognition rate, local facial features are very important, Therefore, to solve the problem of PCA, another higher-order statistical method named Independent Component Analysis (ICA) has been exploited by [1] and [4], which has capability to extract the informative local features from the face. However, if a huge amount of data is exploited as an input, ICA does not has the capability to handle these inputs and might lose informative features which we want.

Therefore, the authors of [3, 5, 11, 15] proposed Local Feature Analysis (LFA) and Local Non-negative Matrix Factorization (LNMF) in order to solve the limitations of statistical methods and to extract informative local facial features from the human face. However, LFA does not extract the local features when there are local distortion and partial occlusion in the pixels located in non-salient areas [10]. Similarly, one of the limitations of LNMF is that it does not assure the significant facial features in the localized area. Moreover, some time LNMF reduces the performance of FER systems because it has no ability to discriminate the features of cheek, forehead and jaw like areas [10].

The objective of this paper is to propose a new feature extraction technique based wavelet transform (especially symlet wavelet family). In this method, the human face is divided into number of regions and in each region the distance between the two pixels has been calculated by employing the distance formula. After that the average distance of each region is calculated and by this way the feature vector is calculated. In the second step of this method, the dimension of the feature space is reduced by exploiting a well-known linear classifier named Linear Discriminant Analysis (LDA), and finally, each expression is labeled by employing a well-known classifier like Hidden Markov Model (HMM).

We already described some related work about this field. The rest of the paper is organized as follows. Section 2 delivers an overview of the proposed feature extraction technique. Section 3 provides some experimental results along with some discussion on the results and a comparison with some of the widely used statistical feature extraction methods. Finally, the paper will be concluded after some future direction in Section 4.

2 Material and Method

2.1 Feature Extraction Using Wavelet Transform

Feature extraction deals with getting the distinguishable features from each facial expression shape and quantizing it as a discrete symbol.

In this stage, the decomposition process has been applied using symlet wavelet, for which the facial frames were in grey scale. The wavelet decomposition could be interpreted as signal decomposition in a set of independent feature vector. Each vector consists of sub-vectors like

$$V_0^{2D} = V_0^{2D-1}, V_0^{2D-2}, V_0^{2D-3}, ..., V_0^{2D-n} \tag{1}$$

where V represents the 2D feature vector. If we have 2D frame X it breaks up into orthogonal sub images corresponding to different visualization. The following equation shows one level of decomposition.

$$X = A_1 + D_1 \tag{2}$$

where X indicates the decomposed image and A_1 and D_1 are called approximation and detail coefficient vectors. If a facial frame is decomposed up to multiple levels, the Eq. 2 can then be written as

$$X = A_j + D_j + D_{j-1} + D_{j-2} + ... + D_2 + D_1 \tag{3}$$

where j represents the level of decomposition, and A and D represent the approximation and detail coefficients respectively. The detail coefficients mostly consist of noise, so for feature extraction only the approximation coefficients are used. In the proposed algorithm, each facial frame is decomposed up to two levels, i.e., the value of $j = 2$, because by exceeding the value of $j = 2$, the facial frame looses significant information, due to which the informative coefficients cannot be detected properly, which may cause misclassification. The detail coefficients further consist of three sub-coefficients, so the Eq. 3 can be written as

$$\begin{aligned} X &= A_2 + D_2 + D_1 \\ &= A_2 + [(D_h)_2 + (D_v)_2 + (D_d)_2] \\ &+ \qquad [(D_h)_1 + (D_v)_1 + (D_d)_1] \end{aligned} \tag{4}$$

where D_h, D_v and D_d are known as horizontal, vertical and diagonal coefficients respectively. It means that all the coefficients are connected with each other like a chain. Note that at each decomposition step, approximation and detail coefficient vectors are obtained by passing the signal through a low-pass filter and high-pass filter respectively.

In each decomposition level, the distance between the pixels is found by using the distance formula and by this way some of the informative coefficients are extracted and hence the feature vector has been created.

$$Dist = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{5}$$

where (x_1, y_1) and (x_2, y_2) are the location of the two pixels respectively.

In a specified time window and frequency bandwith wavelet transform, the frequency is guesstimated. The signal (i.e., facial frame) is analyzed by using the wavelet transform [17].

$$C\left(a_i,\, b_j\right) = \frac{1}{\sqrt{a_i}} \int_{-\infty}^{\infty} y\left(t\right) \Psi_{f.e}^{*} \left(\frac{t - b_j}{a_i}\right) dt \tag{6}$$

where a_i is the scale of the wavelet between lower frequency and upper frequency bounds to get high decision for frequency estimation, and b_j is the position of the wavelet from the start and end of the time window with the spacing of signal sampling period. Other parameters include: time t; the wavelet function $\Psi_{f.e}$ is used for frequency estimation; and $C(a_i, b_i)$ that are the wavelet coefficients with the specified scale and position parameters. Finally, the scale is converted to the mode frequency, f_m for each facial frame:

$$f_m = \frac{f_a\left(\Psi_{f.e}\right)}{a_m\left(\Psi_{f.e}\right).\Delta} \tag{7}$$

where $f_a\left(\Psi_{f.e}\right)$ is the average frequency of the wavelet function, and Δ is the signal sampling period. The feature vector is obtained by taking the average of the whole pixels distance for each facial frame that is given as:

$$f_{dist} = \frac{f_1 + f_2 + f_3 + ... + f_K}{N} \tag{8}$$

where f_{dist} indicates the average distance of each facial frame which is known as a feature vector of that expressions, f_1 f_2 f_3 .... f_K are the mode frequencies for each individual frame, K is the last frame of the current expression, and N represents the whole number of frames in each expression video.

In next step, the dimension of the feature space is reduced by employing a well-known technique Linear Discriminant Analysis (LDA) that maximizes the ratio of between-class variance to within-class variance in any particular data set, thereby guaranteeing maximal separability. For more details on LDA, please refer to [3].

2.2 Expression Modeling and Training Using HMM

HMM is a well-known method that provides a statistical model λ for a set of observation sequences. Sometimes, the observations are called "frames" in facial expression recognition applications. Suppose there are sequence of observations of length T that are denoted by $O_1, O_2, ..., O_T$. An HMM also consists of particular sequences of states, S, whose lengths range from 1 to N $(S = S_1, S_2, ..., S_N)$, where N is the number of states in the model, and the time t for each state is denoted $Q = q_1, q_2, ..., q_N$. The likelihood $P\left(O|\lambda\right)$ can be evaluated by summing over all possible state sequences:

$$P\left(O|\lambda\right) = \sum_{Q} P(O, Q|\lambda) \tag{9}$$

A simple procedure for finding the parameters λ that maximize the above equation for HMMs, introduced in [2] depends on the forward and backward algorithms $\alpha_t(j) = P(O_1...O_t, q_t = j|\lambda)$ and $\beta_t(j) = P(O_(t+1)...O_T|q_t = j, \lambda)$, respectively, such that these variables can be initiated inductively by the following three processes:

$$\alpha_1(j) = \pi_j b_j(O_1), 1 \leq j \leq N \tag{10}$$

$$\beta_T(j) = 1, 1 \leq j \leq N \tag{11}$$

During testing, the appropriate HMMs can then be determined by mean of likelihood estimation for the sequence observations O calculated based on the trained λ as

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{12}$$

The maximum likelihood for the observations provided by the trained HMMs indicates the recognized label. For more details on HMM, please refer to [16].

3 Experimental Results and Discussion

We have tested the idea of employing symlet wavelet transform for human facial expression recognition in the study. The tests were found to be successful and we have achieved significant improvement in recognition rate. The proposed feature extraction has been tested and validated on publicly available standard dataset named Cohn-Kanade [8]. Six basic expressions were collected for experiments such as happy, anger, surprise, sad, disgust, and fear from this dataset that were performed by 40 different subjects. All these expressions in this dataset display the frontal view of the face. The size of each frame was 60x60, where the images were first converted to a zero-mean vector of size 1x3600 for feature extraction.

The system was trained and tested by employing n-fold cross validation rule based on subjects. It means that out of n subjects, data from a single subject was retained as the validation data for testing the proposed scheme, whereas the data for the remaining $n - 1$ subjects were used as the training data. This process was repeated n times, with data from each subject used exactly once as the validation data. The value of n varied according to the dataset used. The total 2,880 ((6 x 40 x12), where 6 represents the number of expressions, 40 indicates the number of subjects, and 12 shows the frames in each expression video) frames are used for the whole experiments.

The performance of the proposed facial feature extraction technique has been validated by comparing it with some of the previous widely used well-known statistical techniques like: PCA, and ICA. The experimental results of the proposed technique along with results of the statistical methods are shown in Figure 1 and in Table 1 and 2.

It is obvious from Figure 1 and Table 1 that the proposed technique achieved best recognition rate than of the statistical methods as shown in Table 2.

Table 1. Confusion matrix of the proposed method on Cohn-Kanade database of facial expressions (Unit: %)

	Happy	**Sad**	**Anger**	**Disgust**	**Surprise**	**Fear**
Happy	**94**	3	0	0	1	2
Sad	2	**95**	0	3	0	0
Anger	0	3	**93**	0	0	4
Disgust	0	1	1	**96**	2	0
Surprise	0	2	0	3	**92**	4
Fear	0	0	4	2	0	**94**
Average			**94**			

Table 2. Confusion matrix of the statistical methods (like: PCA, ICA, LDA) with HMM using Cohn-Kanade database of facial expressions (Unit: %)

	Happy	**Sad**	**Anger**	**Disgust**	**Surprise**	**Fear**
Happy	**86.1**	6	2.5	3.4	2	0
Sad	5	**89**	0	3	0	3
Anger	0	0	**91**	3	0	6
Disgust	0	0	0	**90.1**	9.9	0
Surprise	2	0	0	6	**88**	4
Fear	2	1	15	2	0	**80**
Average			**87.37**			

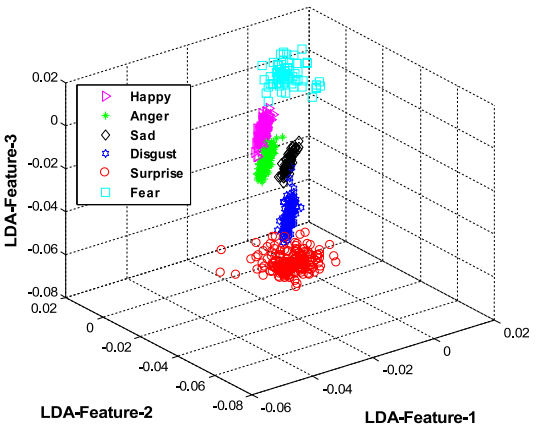

Fig. 1. 3D-feature plot for six different types of facial expressions after LDA. It is indicated that LDA provides best classification rate on the proposed feature extraction technique.

This is because symlet wavelet is a compactly supported wavelet on gray scale images with the least asymmetry and highest number of vanishing moments for a given support width. The symlet wavelet has the capability to support

the characteristics of orthogonal, biorthogonal, and reverse biorthogonal of gray scale images, thats why it provides better classification results. The frequency-based assumption is supported in our experiments. We measure the statistic dependency of wavelet coefficients for all the facial frames of gray scale. Joint probability of a grey scale frame is computed by collecting geometrically aligned frames of the expression for each wavelet coefficient. Mutual information for the wavelet coefficients computed using these distributions is used to estimate the strength of statistical dependency between the two facial frames. Moreover, wavelet transform is capable to extract prominent features from gray scale images with the aid of locality in frequency, orientation and in space as well. Since wavelet is a multi-resolution that helps us to efficiently find the images in coarse-to-find way. Moreover, it is obvious from Figure 1 that applying LDA to features from all the classes provides best separation of the expressions. LDA is a linear technique, which limits its flexibility when applied to complex datasets. Moreover, LDA maximizes the total scatter of the data while minimizing the within scatter of the classes.

4 Conclusion

Facial Expressions Recognition (FER) has become an important research area for many applications over the last decade. A typical FER system consists of three basic modules such as preprocessing module that is used to improve the quality of the image by diminishing the illumination noise and by eliminating the unnecessary details from the background, feature extraction module that deals with getting the distinguishable features each expression and quantizing it as a discrete symbol, and recognition module, in which a classifier is first trained with training data and then used to generate the label of human facial expression contained in the incoming video data. Mostly, facial features are very sensitive to noise and illumination and quite merge with each other in the feature space, that's why in the feature space, it is very hard to separate the different facial expression features. Therefore, very less amount of work can be found be found on feature extraction module in literature. In this work, we proposed a new technique based on symlet wavelet for feature extraction module. In this technique, the human face is divided into number of regions and in each region the distance between the two pixels were calculated based on distance formula. After that, the average distance was found for each region and hence by this way the feature vectors were created. To reduce the dimensions of the feature vectors in the feature space, LDA was exploited. Finally, the expressions were labeled by employing HMM. The proposed system achieved an average recognition accuracy of 94% over Cohn-Kanade dataset, illustrating the successful employment of the proposed method for FER system. In the proposed technique, n-fold cross validation rule has been exploited to achieve best accuracy. The proposed FER system has been trained and tested in laboratory. The next step will be the implementation of the proposed feature extraction technique in smarthomes or in smartphones for real healthcare environment.

References

1. Bartlett, M.S., Donato, G., Movellan, J.R., Hager, J.C., Ekman, P., Sejnowski, T.J.: Face image analysis for expression measurement and detection of deceit. In: Proceedings of the 6th Annual Joint Symposium on Neural Computation (1999)
2. Baum, L.E.: An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. Inequalities 3, 1–8 (1972)
3. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 711–720 (1997)
4. Chuang, C.-F., Shih, F.Y.: Recognizing facial action units using independent component analysis and support vector machine. Pattern Recognition 39(9), 1795–1798 (2006)
5. Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J.: Classifying facial actions. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(10), 974–989 (1999)
6. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognition 36(1), 259–275 (2003)
7. Feng, G.C., Yuen, P.C., Dai, D.Q.: Human face recognition using pca on wavelet subband. Journal of Electronic Imaging 9(2), 226–233 (2000)
8. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53. IEEE (2000)
9. Kaur, M., Vashisht, R.: Comparative study of facial expression recognition techniques. International Journal of Computer Applications 13(1) (2011)
10. Kim, J., Choi, J., Yi, J., Turk, M.: Effective representation using ica for face recognition robust to local distortion and partial occlusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(12), 1977–1981 (2005)
11. Li, S.Z., Hou, X.W., Zhang, H.J., Cheng, Q.S.: Learning spatially localized, parts-based representation. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, p. I-207. IEEE (2001)
12. Lin, D.-T.: Facial expression classification using pca and hierarchical radial basis function network. Journal of Information Science and Engineering 22(5), 1033–1046 (2006)
13. Lin, D.-T.: Human facial expression recognition using hybrid network of PCA and RBFN. In: Kollias, S.D., Stafylopatis, A., Duch, W., Oja, E. (eds.) ICANN 2006. LNCS, vol. 4132, pp. 624–633. Springer, Heidelberg (2006)
14. Mehrabian, A.: Communication without words. Psychological Today 2, 53–55 (1968)
15. Penev, P.S., Atick, J.J.: Local feature analysis: A general statistical theory for object representation. Network: Computation in Neural Systems 7(3), 477–500 (1996)
16. Samaria, F.S.: Face recognition using hidden Markov models. PhD thesis, University of Cambridge (1994)
17. Turunen, J., et al.: A wavelet-based method for estimating damping in power systems (2011)