A New Feature Extraction Technique for Human Facial Expression Recognition Systems using Depth Camera

Muhammad Hameed Siddiqi[1], Rahman Ali[1], Byeong Ho Kang[2], Sungyoung Lee[1]⋆

[1]Department of Computer Engineering, Kyung Hee University
(Global Campus), Suwon, Rep. of Korea.
{siddiqi, rahmanali, sylee}@oslab.khu.ac.kr
[2]Department of Computing and Information Systems, University of Tasmania, Australia.
byeong.kang@utas.edu.au

Abstract. The analysis of facial expressions in telemedicine and healthcare plays a significant role in providing sufficient information about patients like stroke and cardiac in monitoring their expressions for better management of their diseases. Due to some privacy concerns, depth camera is a good candidate in such domains over RGB camera for facial expression recognition (FER). The accuracy of such FER systems are completely reliant on the extraction of the informative features. In this work, we have tested and validated the accuracy of a new feature extraction method based on symlet wavelet transform. In this method, the human face is divided into number of regions and in each region the movement of pixels have been traced in order to create the feature vectors. Each expression frame is decomposed up to 4 levels. In each decomposition level, the distance between the two corresponding pixels is found by using the distance formula in order to extract the most informative coefficients. After feature vector creation, Linear Discriminant Analysis (LDA) has been employed to reduce the dimensions of the feature space. Lastly, Hidden Markov Model (HMM) has been exploited for expression recognition. Most of the previous FER systems used existing available standard datasets and all the datasets were pose-based datasets. Therefore, we have collected our own depth data of 15 subjects by employing the dept camera. For the whole experiments, 10-fold cross validation scheme was utilized for the experiments. The proposed technique showed a significant improvement in accuracy against the existing works.

Keywords: Facial Expressions, Depth Camera, Multilevel Wavelet Decomposition, Linear Discriminant Analysis, Hidden Markov Model

⋆ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2013-067321)). This research was also 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-2014-(H0301-14-1003)).

1 Introduction

Telemedicine and healthcare applications that employ video technologies raise privacy concerns since it can lead to situations where subjects may not know that their private information is being shared and thus become exposed to a threat [13]. Unlike RGB-cameras, depth-cameras only capture the depth information and do not reveal the identity of the subject or other sensitive information, which makes them a superior choice over RGB-cameras. Therefore, we choose the depth-camera over RGB-cameras for the proposed FER system. To the best of our knowledge, no sufficient work has been done to study the expression recognition with depth camera.

There are three basic modules in a typical FER system: preprocessing, feature extraction and recognition. Much work has been done so far for preprocessing and recognition modules, and we also employed well-known method such as histogram equalization (HE) and hidden Markov model (HMM) for preprocessing and recognition modules respectively.

Regarding to the feature extraction, huge amount of methods have been proposed; however, most of them have their own limitations. These methods include global feature-based methods such as Nearest Features Line-based Subspace Analysis [12], Eigenfaces and Eigenvector [2, 8] and [7], Fisherfaces [1], global features [11], neural network, and Independent Component Analysis (ICA) [10]. However, these techniques are poor at handling data in which the classes do not follow the Gaussian distribution. Also, these techniques do not work well in case of a small sample size [5]. On the other hand, local feature-based methods have been proposed to compute the local descriptors from parts of the face and then integrate this information into one descriptor. These methods include Local Feature Analysis (LFA) [9], Gabor features [6], Non-negative Matrix Factorization (NMF), Local non-negative Matrix Factorization (LNMF) [4], and Local Binary Pattern (LBP) [15]. Among these methods, LBP is the most commonly employed feature extraction technique. However, LBP does not provide the directional information of the facial frame [16].

Accordingly, in this work, we have proposed a new feature extraction technique based on symlet wavelet transform. 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. Once the feature vectors have been created, the dimension of feature space is reduced by employing 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 feature extraction methods. Finally, the paper will be concluded after some future direction in Section 4.

2 Material and Method

2.1 Symlet Wavelet Transform-based Feature Extraction

In real-life scenarios, some environmental parameters (such as lighting effects) may produce some noise in the expression frames that could reduce the recognition rate. The proposed method employs symlet wavelet to reduce such noise. Facial frames are converted to grey scale prior to applying this step. In the first feature extraction, the decomposition process has been applied, 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 an expression frame X in the decomposition process, and it breaks up into the 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 respectively. If a facial frame is decomposed up to multiple levels, then Eq. 2 can 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. 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 $j = 4$, because by exceeding the value of $j > 4$, 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

$$\begin{aligned} X &= A_4 + D_4 + D_3 + D_2 + D_1 \\ &= A_4 + [(D_h)_4 + (D_v)_4 + (D_d)_4] \\ &+ [(D_h)_3 + (D_v)_3 + (D_d)_3] \\ &+ [(D_h)_2 + (D_v)_2 + (D_d)_2] \\ &+ [(D_h)_1 + (D_v)_1 + (D_d)_1] \end{aligned} \tag{4}$$

where D_h, D_v and D_d are known as horizontal, vertical and diagonal coefficients respectively. 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 using the distance formula and by this way some of the informative coefficients are extracted and hence the feature vector has been created.

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

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 bandwidth wavelet transform, the frequency is estimated. The signal (i.e., facial frame) is analyzed by using the wavelet transform [17].

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

where a_i is the scale of the wavelet between lower and upper frequency bounds to get high decision for frequency estimation, and b_j is the position of the wavelet from the start to the end of the time window with the specified signal sampling period, t is the time, the wavelet function $\Psi_{f.e}$ is used for frequency estimation, and $C(a_i, b_i)$ 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]. At last, the expressions are recognized by employing HMM for which the parameters were 64, 4, and 4, respectively. For more details on HMM, please refer to [14].

3 Results and Discussion

In order to validate the performance of the proposed feature extraction technique, we have created our own data by utilizing Intel creative depth data. The dataset was collected from 15 subjects (university students) that displays frontal view of the face and each expression is composed of several sequences of expression frames. During each experiment, we reduced the size of each input image (expression frame) to 60×60, where the images were first converted to a zero-mean vector of size 1×3600 for feature extraction. All the experiments were performed in Matlab using an Intel® Pentium® Dual-Core^{TM} (2.5 GHz) with a RAM capacity of 3 GB. For all the experiments, a 10−fold cross-validation scheme (based on subjects) was used. In other words, out of 10 subjects data from a single subject was used as the validation data, whereas data for the remaining 9 subjects were used as the training data. This process was repeated 10 times with data from each subject used exactly once as the validation data. The total images utilized for the proposed system were 1,080 (6×15×12), where 6 represents the number of expressions, 15 indicates the number of subjects, and 12 shows the frames in each expression video.

The performance of the proposed feature extraction technique has been validated by comparing it with some of the previous widely used well-known techniques like: LBP, and LDP, and LTP. The experimental results of the proposed feature extraction technique are shown in Figure 1 and Table 1, while the results of the existing methods (LBP, and LDP) are described in Table 2.

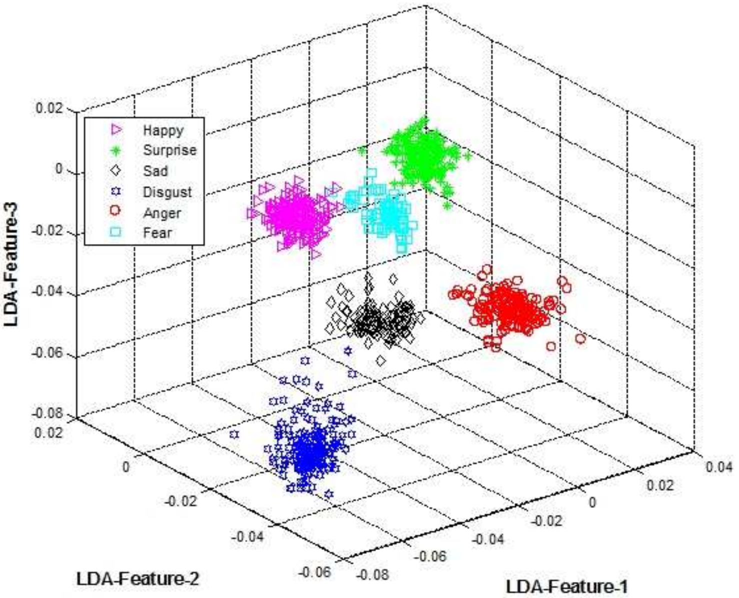

Fig. 1. 3D-feature plot for six different types of facial expressions. It is indicated that the proposed technique provides best classification rate on depth dataset of facial expression.

It is obvious from Figure 1 and Table 1 that the proposed technique achieved

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

	Happy	**Sad**	**Anger**	**Disgust**	**Surprise**	**Fear**
Happy	**93**	3	1	0	1	2
Sad	2	**94**	1	3	0	0
Anger	0	3	**92**	1	0	4
Disgust	2	1	1	**94**	2	0
Surprise	0	2	0	3	**92**	4
Fear	0	0	4	2	1	**93**
Average			**93.00**			

Table 2. Confusion matrix of (A) the LBP on our own dataset, and (B) the LDP on our own dataset of facial expression (Unit: %).

	Happy	**Sad**	**Anger**	**Disgust**	**Surprise**	**Fear**
Happy	**87**	3	2	2	3	3
Sad	4	**85**	3	3	3	2
Anger	2	2	**88**	2	4	2
Disgust	3	4	3	**84**	4	2
Surprise	3	2	5	4	**80**	6
Fear	1	2	5	3	2	**87**
Average			**85.17**			

(A)

	Happy	**Sad**	**Anger**	**Disgust**	**Surprise**	**Fear**
Happy	**79**	7	4	3	2	5
Sad	6	**80**	4	4	3	3
Anger	2	3	**83**	5	3	4
Disgust	0	2	4	**90**	1	3
Surprise	1	5	2	4	**85**	3
Fear	2	6	3	3	4	**82**
Average			**83.10**			

(B)

better recognition rate than that of the statistical methods as shown in Table 2. 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 statistical 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.

4 Conclusion

Facial Expressions Recognition (FER) has become an important research area for many applications over the last decade. Communication through facial expressions plays a significant role in telemedicine, and social interactions. In such applications, RGB cameras might not be used due some privacy concerns. A typical FER system consists of three basic modules such as preprocessing, feature extraction and recognition. Some very common methods such as histogram equalization (HE) and hidden Markov model (HMM) have been employed for preprocessing and recognition modules respectively. The 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 these features. Therefore, very less amount of work can be found on the feature extraction module in literature; however, most of them have their own limitations. Accordingly, in this work, we utilized Intel creative depth camera in order to tackle the privacy issue in the proposed FER system. Moreover, we proposed a new and robust feature extraction technique based on symlet wavelet for feature extraction. 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 the 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. In order to validate the performance of the proposed technique, we have collected our own data from 15 subjects (university students) in the frontal view of the camera. For all the experiments, we applied 10-fold cross validation scheme. The proposed system produced a significant improvement in the recognition rate (93%) against the existing methods. The proposed FER system has been trained and tested in laboratory. The next step will be the implementation of the proposed feature extraction technique either in smarthomes or in smartphones.

References

1. Zaenal Abidin and Agus Harjoko. A neural network based facial expression recognition using fisherface. *International Journal of Computer Applications*, 59(3), 2012.
2. Gualberto Aguilar-Torres, Karina Toscano-Medina, Gabriel Sanchez-Perez, Mariko Nakano-Miyatake, and Hector Perez-Meana. Eigenface-gabor algorithm for feature extraction in face recognition. *International Journal of Computers*, 3(1):20–30, 2009.

3. Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.
4. Ioan Buciu and Ioannis Pitas. Application of non-negative and local non negative matrix factorization to facial expression recognition. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 288–291. IEEE, 2004.
5. S Chitra and Dr G Balakrishnan. A survey of face recognition on feature extraction process of dimensionality reduction techniques. *Journal of Theoretical and Applied Information Technology*, 36(1):92–100, 2012.
6. Wenfei Gu, Cheng Xiang, YV Venkatesh, Dong Huang, and Hai Lin. Facial expression recognition using radial encoding of local gabor features and classifier synthesis. *Pattern Recognition*, 45(1):80–91, 2012.
7. Jeemoni Kalita and Karen Das. Recognition of facial expression using eigenvector based distributed features and euclidean distance based decision making technique. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 4(2):196–202, 2013.
8. Senthil Ragavan Valayapalayam Kittusamy and Venkatesh Chakrapani. Facial expressions recognition using eigenspaces. *Journal of Computer Science*, 8(10):1674–1679, 2012.
9. Stan Z Li, Xin Wen Hou, Hong Jiang Zhang, and Qian Sheng Cheng. Learning spatially localized, parts-based representation. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–207. IEEE, 2001.
10. Fei Long, Tingfan Wu, Javier R Movellan, Marian S Bartlett, and Gwen Littlewort. Learning spatiotemporal features by using independent component analysis with application to facial expression recognition. *Neurocomputing*, 2(1):126–132, 2012.
11. Vaibhavkumar J Mistry and Mahesh M Goyani. A literature survey on facial expression recognition using global features. *International Journal of Engineering and Advanced Technology*, 2(4):653–657, 2013.
12. Yanwei Pang, Yuan Yuan, and Xuelong Li. Iterative subspace analysis based on feature line distance. *Image Processing, IEEE Transactions on*, 18(4):903–907, 2009.
13. Rusyaizila Ramli, Nasriah Zakaria, and Putra Sumari. Privacy issues in pervasive healthcare monitoring system: A review. *World Acad. Sci. Eng. Technol*, 72:741–747, 2010.
14. Ferdinando Silvestro Samaria. *Face recognition using hidden Markov models.* PhD thesis, University of Cambridge, 1994.
15. Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.
16. Muhammad Hameed Siddiqi, Faisal Farooq, and Sungyoung Lee. A robust feature extraction method for human facial expressions recognition systems. In *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*, pages 464–468. ACM, 2012.
17. Jukka Turunen et al. A wavelet-based method for estimating damping in power systems. 2011.