Recurrent Neural Networks for Facial Action Unit Recognition from Image Sequences

HB Vadapalli
School of Computer Science
University of Witwatersrand
Private Bag 3, Wits 2050, South Africa
Hima.vadapalli@wits.ac.za

H Nyongesa
Department of Computer Science
University of the Western Cape
Bellville, South Africa
hnyongesa@uwc.ac.za

CWP Omlin
Middle East Technical University
Northern Cyprus Campus
Güzelyurt, Mersin10, Turkey
Omlin@metu.edu.tr

Abstract— **The Facial Action Coding System (FACS) has facilitated major advances in the area of Human Computer Interaction (HCI) through its ability to represent facial expressions. Facial expressions are viewed in terms of action units (AU's), which describe the subtle muscle movements involved in facial expressions. In this paper, we propose to use temporal data from image sequences for the recognition of a set of 6 upper face and 5 lower face action units. We use Gabor filters to extract salient features, apply dimensionality reduction in feature space and train recurrent neural network for the classification of action units. Recurrent neural networks help in handling the temporal dependencies present in image sequences. Our method achieves an average recognition rate of 85.4% for the 11 FACS AU's.**

Keywords-component; Gabor filters, feature extraction, recurrent neural networks, dimensionality reduction, optimization

I. INTRODUCTION

In the field of facial expression recognition (FER), the study of expressions as a set of muscle actions has gained major interest in recent years [2, 3, 13]. The Facial Action Coding System (FACS) introduced by Ekman and Friesen provides a framework that defines expressions in terms of subtle muscles movements, known as action units (AU's) [5]. The use of an image sequence rather than a single static image together with the classification of facial expressions in terms of AU's has given promising results in FER [2, 13].

Recurrent neural networks (RNN) have been successful in modeling, classifying and predicting time series for applications such as speech recognition [1], gesture recognition [11] and forecasting [10]. Recent work in [8] suggested that RNN's could also be successfully used for the recognition of facial expressions. Recurrent neural networks allow self-connections, which enable them to gain the knowledge of the past events and thus allowing them to be flexible with temporal data. Our work is aimed at using this ability of RNN's for the recognition of FACS AU's rather than expressions themselves.

Faces provide a high-dimensional input space. In order to make FACS AU's modeling and classification possible, we apply Gabor filters to extract salient features; we further reduce the dimensionality of the response vectors through the application of frequency scale selection [4, 13], local Gabor filters [3], and principle component analysis (PCA). We reduce the dimensionality of the model by applying weight decay.

II. FACS ACTION UNIT DATABASE

We use the Cohn-Kanade FACS Database [9], which is a collection of video frames of 97 subjects mimicking either single AU or combination of AU's. The subjects represent a range of ages, genders, and origins. The subjects were asked to directly face the camera and perform a series of 23 facial displays. Each sequence begins with a neutral facial expression that transforms to a target display where the AU's are at their peak intensity. The video images were digitized into 640*480 pixels arrays with 8-bit precision for grayscale images.

III. PREPROCESSING

Preprocessing of the facial images plays a vital role towards the systems performance in recognizing the action units. In this work we perform the following preprocessing steps: 1. Manual detection of facial features like eye centers, 2. Face normalization using these eye centers, and 3. Face detection and segmentation. In the first frame of every sequence we locate the eye coordinates. These coordinates are then used to rotate the face to line up the eye centers. Face detection and cropping is performed using MPISearch [7], which reduces the image size to 64*64 pixels. Further segmentation is performed to retrieve either upper or lower half of the face region. Segmentation of face into regions is based on the knowledge that, the AU's active in the upper face have little or no effect on the appearance changes in the lower face region and vice versa [5].

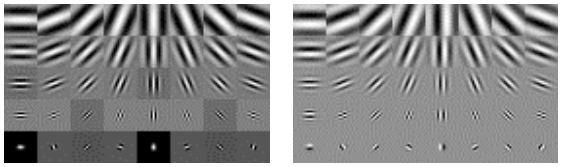
IV. Feature Extraction

While Gabor filters have shown to be more promising than geometry-based methods [16], they suffer from the disadvantage of incurring high computational costs. Here, we choose 5 frequencies and 8 orientations as in [13]. The real and imaginary components of a 5*8 Gabor jet are shown in Fig.1. In order to further reduce the dimensionality of the input, the output responses of the 40 Gabor filters were down sampled by a factor of 16 and normalized to unit length [2]. We chose a total of 300 video sequences from 78 subjects from the Cohn-Kanade database for the recognition of 6 upper face AU's and a total of 258 video sequences from 67 subjects for the lower face AU's. The criteria for selection of a sample from a subject is that it has at least 5 frames which, when put together will depict the AU of interest in action (from its absence to its presence at its peak intensity). The entire set is then divided into training and test sets using cross validation.

V. Classification

For facial expression modeling and classification, we use an Elman recurrent neural network [6] shown in Fig. 2. The network is trained using back propagation-through-time algorithm [15]. Recurrent neural networks are able to model, classify and predict dynamic systems with hidden states, i.e. they are able to develop internal representations of hidden states from observable states alone. Furthermore, they can accumulate and store information over time due to their internal feedback mechanism. In the context of facial action unit recognition, we hypothesize that, classification performance can be improved by the use of recurrent neural networks that use video sequences as inputs when compared to the use of modeling and classification methods which only use static images for facial AU classification. This advantage of recurrent neural networks over other classifiers may become more pronounced as we use video sequences where the position of the subject's head changes.

The recurrent neural network comprises of 5120 input neurons (size of the downsized image multiplied by the number of Gabor coefficients), 15 hidden neurons (optimized value obtained through experimentation), and a single output neuron (which represents the activation/non-activation of single AU's). The network weights were initialized to small random values in the range [-1, 1], the learning rate was set to 0.01 (through experimentation), and training is stopped after a maximum of 500 epochs have elapsed. An AU is recorded as present if the network output is greater than or equal to 0.5.



a) Real components b) Imaginary components

Figure 1. Real and imaginary components of a 5*8 Gabor jet.

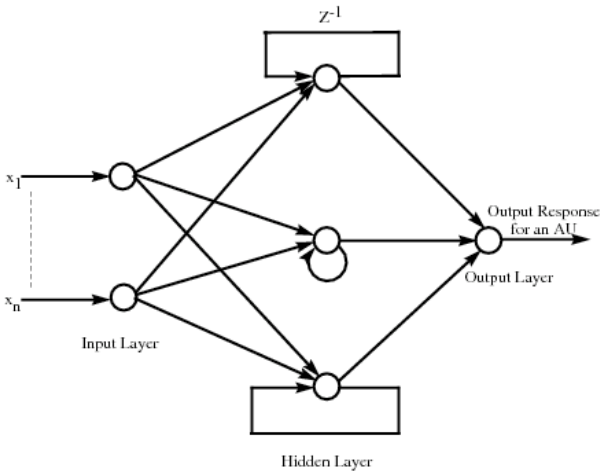


Figure 2. Elman recurrent neural network

In order to establish a baseline recognition rate, we classify the 6 upper face and 5 lower face AUs using all the Gabor coefficients obtained from the 40 Gabor filters without any dimensionality reduction or subsequent feature extraction techniques. This set of experiments will help us in determining whether the use of RNN's result in similar performance as that of other classic classifiers.

For the 6 upper face AU's we obtained an average recognition rate of 83.5% with a false positive rate of 6.7%. In a similar manner an average recognition rate of 81.98% with a false positive rate of 8.53% was obtained for the 5 lower face AU's. This supports the use of RNN's for FACS AU recognition. The average recognition rate reported by Tian [13] was 87.6% using geometric features alone and 32% using appearance based method alone for a set of 9 upper face FACS AU's employing no feature extraction techniques. The work by Barlett [2], gave an average hit rate of 80.1% with a false alarm rate of 8.2%. The comparison with the above systems shows that the classification performance obtained by the use of recurrent neural networks is inline with that obtained by the use of other famous classifiers like support vector machines and neural networks. The use of recurrent neural networks has the added advantage of handling the temporal dependencies, which is not present in other systems. The individual recognition rates are shown in Table 1 and Table 2 for the upper and lower face AU's respectively.

Table I. Recognition and false positive rate for 6 upper face AU's

FACS AU's	Recognition Rate	False Positive Rate
AU1	89.96	5.01
AU2	92.66	4.20
AU3	81.50	7.54
AU4	83.59	6.00
AU5	77.56	8.20
AU6	75.80	9.18
Average	83.51	6.68

TABLE II. RECOGNITION AND FALSE POSITIVE RATE FOR 5 LOWER FACE AU'S

FACS AU's	Recognition Rate	False Positive Rate
AU15	74.71	12.97
AU17	72.33	10.45
AU20	84.27	4.04
AU25	91.10	9.30
AU27	87.50	5.86
Average	81.98	8.53

VI. DIMENSIONALITY REDUCTION

Gabor filters are one of most successful feature extraction techniques used [4]. However, they suffer from a major drawback: When a Gabor filter with 5 frequencies and 8 orientations is applied to a single image; the extracted Gabor coefficients require 40 times more memory than the original image. This huge dimensionality of the Gabor coefficients generated has a detrimental effect on the classification performance. Dimensionality reduction techniques such as frequency selection [13], local Gabor filters [3], and principle component analysis (PCA) [3] have been widely used in conjunction with support vector machines and neural networks. We propose to use these famous methods in conjunction with recurrent neural networks and observe how the generalization performance is affected. The above methods need no or little extra preprocessing and thus do not add to the computational cost, which is a huge advantage in modeling real time facial action unit recognition systems. However, any selection/reduction technique used on features is said to be successful when it helps in increasing the generalization performance, in addition to decreasing the dimensionality.

A. Frequency selection

Frequency selection is one of the widely used methods for reducing the high dimensional response vectors generated by Gabor filters. Studies in [4] and [13] emphasize that not all frequency scales are required for the recognition of AUs. In [4], the set of three highest frequency scales performed similar to that of using all the frequency scales. However, in [13] the set of three middle frequency scales performed better for the recognition of upper face AUs. In this work, we divide the set of five frequency scales into three sets, as in [13]. We compare the performance of our system using these three frequency subsets. The overall performance of our system obtained using these three sets of frequency scales are given in Table 3. In this study, we confirm the results reported in [4] that the set of three highest frequency scales in case of upper face AU's leads to similar performance as that of using all the frequency components. In this case, there is no significant degradation in the generalization capability of the classifier. However, the performance obtained using the lower frequencies was the least. In case of lower face AU's the set of lower frequencies performed better than the set of middle frequencies. This emphasizes that not all the regions in the face act similar to a set of frequency components. The performance of frequency components for AU recognition is dependent on the face region and also the AU's under consideration.

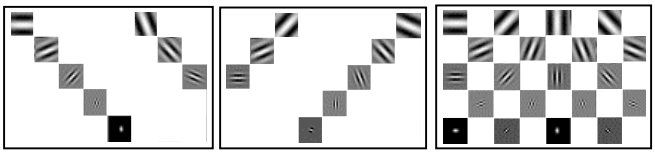
TABLE III. RECOGNITION RATES USING HIGHER, MIDDLE AND LOWER SET OF FREQUENCY SCALES FOR THE UPPER AND LOWER FACE AU'S

Face Region	Higher Frequencies RR%	Middle Frequencies RR%	Lower Frequencies RR%
Upper face AU's	82.07	80.10	77.40
Lower face AU's	79.56	75.03	76.84

B. Local Gabor filters

Disregarding some frequency scales reduces the dimensionality of the input space, but it may also lead to a loss of some information that is important for classification of AU's. Thus, a selection method is required which samples data from all the frequency scales, and yet helps in reducing the dimensionality of the input space by neglecting redundant data. It is evident from Fig.1, that Gabor representations obtained by two neighboring frequencies at the same given orientation are very much similar.

Taking the above-mentioned similarity into account, Deng [3] proposed a novel approach that uses a filter bank and a subset of all frequencies and orientations, so-called local Gabor filters. Local Gabor filters are denoted as LG (mxn) where m is the total number of spatial frequencies and n is the total number of orientations. Three variants of local Gabor filters suggested in [3] are LG1 (mxn), LG2 (mxn) and LG3 (mxn). LG1 (mxn) is formulated by repeatedly increasing the frequency from minimum to maximum, while orientation is incremented by 1 each time. LG2 (mxn) is the same as LG1 (mxn) except for a decrease in frequency, changing from maximum to minimum. For LG3 (mxn), the responses are selected with an interval between any two filters. The resultant feature vectors for LG1 (5x8), LG2 (5x8) and LG3 (5x8) are shown in Fig. 4. The memory requirements in terms of number of input neurons for the local Gabor filters LG1 (5*8), LG2 (5*8) and LG3 (5*8) are given in Table 4.



(a) LG1 (5x8) (b) LG2 (5x8) (c) LG3 (5x8)

Figure 3. Local Gabor filters

TABLE IV. MEMORY REQUIREMENTS OF GLOBAL AND LOCAL GABOR FILTERS

Gabor Jet	Original dimension	Upper face dimension	Sub sampling dimension	No. of hidden Units
G(5*8)	163840	81920	5120	15
LG1(5*8)	32768	16384	1024	10
LG2(5*8)	32768	16384	1024	10
LG3(5*8)	81920	40960	2560	10

TABLE V. RECOGNITION AND FALSE POSITIVE RATE FOR GLOBAL AND LOCAL GABOR FILTERS FOR THE 6 UPPER FACE FACS AU'S

Gabor Variants	Recognition Rate	False Positive Rate
G (5*8)	83.51	6.68
LG1 (5*8)	79.90	10.21
LG2 (5*8)	82.61	7.59
LG3 (5*8)	84.56	6.79

TABLE VI. RECOGNITION AND FALSE POSITIVE RATE FOR GLOBAL AND LOCAL GABOR FILTERS FOR THE 5 LOWER FACE FACS AU'S

Gabor Variants	Recognition Rate	False Positive Rate
G (5*8)	81.98	8.53
LG1 (5*8)	75.20	9.10
LG2 (5*8)	78.50	8.21
LG3 (5*8)	80.85	7.76

In this set of experiments, we trained our recurrent neural network using the three variants of the local Gabor filters mentioned above and empirically optimized our network architecture to 10 hidden neurons. The average recognition and false positive rate for the 6 upper face AU's using all the variants are given in Table 5. And that for the 5 lower face AU's are given in Table 6. When compared, LG3 (5*8) variant of the local Gabor filters performed slightly better than G (5*8) for the 6 upper face AU's. This can be attributed to the decrease in the number of input neurons and removal of ambiguous data. However, the same was not true in case of lower face AU's. Studying the results for the lower face AU's, it can be indicated that the AU's in the lower face region require more information for their detection. In general, LG3 variant helped in reducing the dimensionality of the system by 50%, and both LG2 (5*8) and LG3 (5*8) variants outperformed LG1 (5*8).

C. Feature selection using PCA

Principle component analysis (PCA) is one of most widely used techniques for feature selection and dimensionality reduction. PCA is effective in finding linear transformations and seeks a projection that best represents the original data in a least square sense [3]. In the present work, PCA is used to select features from the high dimensional response vectors generated by Gabor filters. However, the use of PCA has shown to be greatly affected by the lighting conditions of the available data [3]. It is thus advisable to eliminate the effects of lighting condition before applying PCA for feature selection. To control the illumination effect, we use histogram normalization [3].

Our recurrent neural network using PCA achieved an average recognition rate of 78.6% with a false positive rate of 11.7% for the 6 upper face AU's. The average recognition and false positive rate for the 5 lower face AU's was 71.1% and 8.3% respectively. Clearly, PCA failed to increase the generalization performance of our classifier; all PCA components perform no better than all Gabor coefficients. Even the removal of the first few components, which contain the information about the orientation and lighting conditions, failed to improve the performance.

VII. RNN OPTIMIZATION

Optimization of recurrent neural networks plays a vital role in their generalization performance. The most widely used regularization methods for network optimization are early stopping and weight decay. In optimization through early stopping, training data is divided into separate training and validation sets. While training the network using the new training set, it is also tested for the validation error on the validation set. The network, which shows the lowest validation error, is then used for testing. However, early stopping suffers from over-fitting of data where the generalization performance is good for one set and fails for another set. To overcome this problem, weight decay is used in many real world applications [14]. Weight decay is the process of adding a penalty term to the objective function. The common penalty term used is a constant multiple of sum of the squares of the weights. This regularization penalizes large weights, as these tend to increase the variance in neuron outputs. We experimented with different decay constants (0.01, 0.001, and 0.0001) to obtain the optimal decay constant value for our network (0.0001). The recognition and false positive rates for the six upper face and 5 lower face AU's are given in Table 7 and Table 8 respectively. Clearly, weight decay helped in improving the overall classification performance in case of both upper face and lower face AU's. The % increase in terms of recognition rate was more evident in case of lower face AU's, with more than 3% raise.

TABLE VII. RECOGNITION AND FALSE POSITIVE RATE FOR THE 6 UPPER FACE FACS AU'S WITH A DECAY CONSTANT OF 0.0001.

FACS AU's	Recognition Rate	False Positive Rate
AU1	89.57	3.49
AU2	94.72	3.95
AU4	81.18	10.66
AU5	90.10	2.17
AU6	80.59	6.65
AU7	78.85	11.53
Average	85.84	6.41

TABLE VIII. RECOGNITION AND FALSE POSITIVE RATE FOR THE 5 LOWER FACE FACS AU'S WITH A DECAY CONSTANT OF 0.0001.

FACS AU's	Recognition Rate	False Positive Rate
AU15	78.13	8.89
AU17	74.64	14.23
AU20	87.90	0.00
AU25	92.80	4.20
AU27	91.10	3.04
Average	84.91	6.07

VIII. Summary and Directions for Future Research

In this paper, we discussed the use of recurrent neural networks for the recognition of 6 upper face and 5 lower face FACS AU's using Gabor filters for feature extraction. We applied different dimensionality reduction techniques including frequency scale selection, local Gabor filters and principle component analysis (PCA) and studied their effect on the classification performance of recurrent neural networks. The LG3 variant of local Gabor filters performed the best (84.6% average performance) for the 6 upper face AU's, where as the use of all the Gabor filters performed the best (82% average performance) for the 5 lower face AU's. This emphasizes that lower face AU's require more information for their classification. Network regularization through weight decay improved the classification performance (85.8% average performance for 6 upper face AU's and 84.9% for the 5 lower face AU's). Future work will include the training of recurrent neural network on all AU's defined in FACS, comparison of training a single recurrent neural network for all AU's vs. training individual recurrent neural networks for individual AU's and the application of our methodology to scenarios where subjects change the position of their head relative to the camera, i.e. nodding and head turning. It is in this latter scenario where recurrent neural networks may show their real advantage over methods which base their classification on single static images.

Acknowledgment

The authors would like to thank Department of Computer Science, University of the Western Cape for their support in carrying out this research work. The first author (HB Vadapalli) is currently registered as a PhD student with the Department of Computer Science, University of the Western Cape.

References

[1] A.M. Ahmad, S. Ismaail, and D.F. Samaon, "Recurrent neural networks with back propagation through time for speech recognition," International Symposium on Communications and Information Technologies, ISCIT2004, Sapporo, Japan, Oct 26-29, 2004.

[2] M. Barlett, G. Littlewort, M.Frank, C. Iainseek, I.Fasel, and J. Movellan, "Fully automatic facial action recogtion in spontaneous behaviour", 7Th international Conference on Automatic Face and Gesture recognition, 223-230, 2006.

[3] H.B. Deng, L.W. Jin, L.X. Zhen and J.C. Huang, "A new facial expression recogntiion method based on local Gabor filter bank and PCA plus LDA", International Journal of Information Technologies, 11, 2005.

[4] D. Donato, M.S. Barlett, J.C. Hager, and T.J. Senjowski,"Classifying facial actions", IEEE Pattern analysis and Machine Intelligence, 21, No.10, 1999.

[5] P. Ekman, and W. Friesen,"Facial action coding system: A technique for the masurement of facial movement", Consulting Phychologists Press, Polo Alto, 1978.

[6] J. Elman,"Finding structure in time", Cognitive Science, 14, 179-211, 1990.

[7] I. Fasel, R. Dhal, J. Hershey, B. Susskind, and J.R. Movellan, Machine Perception Toolbox.

[8] A. Graves, C. Mayer, M. Wimmer, J. Schmidhuber and B. Radig,"Facial expression recognition with recurrent neural networks," Proceedings of the International Workshop on Cognition for Technical Systems, Munich, germany, Oct, 2008.

[9] T. Kanade, J. Cohn, and Y. Tian,"Comprehensive database for facial expression analysis", Proceedings of 7th International Conference on Automatic Face and Gesture Recognition, 46-53, 2000.

[10] V.V. Kondratenko, and Yu. Kuperin,"using recurrent neural networks to forecasting of forex", Disordered Systems and Neural Networks, April, 2003.

[11] A. Nikoloas,"Human computer interactionbased on hand gesture ontology", Proceedings of the 11th WSEAS Interational Conference on Computers, 26-31, 2007.

[12] T. Robinson,"An application of recurrent nets to phone probability estimate", IEEE Transactions on Neural Networks, Vol. 5, No.3, 1994.

[13] Y. Tian, T. Kanade, J. Cohn,"Evaluation of Gabor wavelets based facial action unit recognition in image sequences of increasing complexity", Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition, 2002.

[14] P. Werbos,"Backpropagation: Past and Future", Proceedings of the IEEE International Conference on Neural Networks, IEEE Press, 343-353, 1998.

[15] P. Werbos,"Backpropagation through time: what it does and how to do it", Proceedings of the IEEE, Vol.78, No. 10, October, 1990.

[16] Z. Zhang, M. Lyons, M. Schuster, S.Akamatsu,"Comparision between geometry based and Gabor wavelet based facial expression recognition using multilayer perceptron", Proceedings of 3rd IEEE International Conference on Automatic Face and Gesture Recognition, 454-459, 1998.