Facial Action Unit Recognition from Video Streams with Recurrent Neural Networks

Hima Vadapalli
University of the Witwatersrand
Braamfontein, Johannesburg
South Africa

ABSTRACT
Facial expressions are one of the parameters for accessing individual behavioral processes. Their recognition and verification can be framed as the identification of states of dynamical systems generated by physiological processes. Whereas a snap shot of a dynamical system gives information about its current state, a time series of past states captures its trajectory in state space. The description and recognition of facial expressions using atomic muscle movements, so-called action units provide an extensive framework. The temporal modeling and recognition of these muscle movements promises a broader and more generic approach for recognizing subtle changes on the facial region. This paper proposes the use of recurrent neural networks for modeling facial action unit activity. Recurrent neural networks are able to model actions based on their previous and current states, unlike other dynamic classifiers such as hidden Markov models. A detailed comparative analysis with the recognition performance of a static classifier such as support vector machines suggests that recurrent neural networks gain more knowledge about the action unit activation when presented with a sequence of images. On average our model achieved a positive hit rate of 85.8% for upper face action units and 84.9% for lower face action units.

General Terms
Pattern Recognition, Machine Learning.

Keywords
Computer Vision, Face and Gesture Recognition, Feature Extraction, Neural Nets.

1. INTRODUCTION
There has been a significant increase in interest in real-time recognition of facial expressions for applications ranging from human-computer interaction to computer vision and robotics. Facial expression recognition (FER) is focused on the analysis and assessment of the activation of facial muscles. In the past, automatic facial expression recognition was mainly focused on the recognition of canonical expressions such as joy, sadness etc. [7], [13], [19], [25], [30], [35], [36], [45]. However, recent work is focused on the recognition of facial expressions using Facial Action Coding System (FACS) [10], [11] action units (AUs) as basic recognition entities [28], [39], [40]. Facial expressions describe global changes in a face caused by a combination of muscle movements, where as AUs describe subtle local appearance changes in a face due to action of few individual muscles.

Early FER systems extracted features based on the displacements of predefined models in a facial region. A typical application was the recognition of canonical facial expressions [7]. The recognition performance could be improved by fitting physical 3D models onto an individual's face and measuring the displacement of motion field on the 3D model [13]. Other popular researchers used grid tracking and deformation systems, which used the difference in the node coordinates from the first and highest intensity frames.

Advances in feature extraction from images led to further improvements in recognition rates. A comparison of different feature extraction methods showed that the use of Gabor wavelets resulted in the best recognition performance [9]. Detailed studies on the selection of frequencies and orientations of Gabor jets were performed by [30], [45]; a comparison of the extraction of global vs. regional feature extraction using Gabor filters was given in [45]. The use of Haar features classified with an Adaboost classifier showed performance comparable to the use of Gabor filters with support vector machines (SVMs) as classifiers [20].

In addition to feature extraction using different approaches mentioned above, different classification schemes such as static and dynamic classifiers have been investigated for facial expression recognition. Static classifiers classify a single image or a frame from a video and map it onto a facial expression. Static classifiers such neural networks, Support Vector Machines (SVMs) and principle component analysis (PCA) usually work with frames displaying an expression at its peak intensity. These classifiers have been widely used for both expression classification [4], [14], [25], [29] and AU classification [4], [39]. Bartlett et al [4] classified Gabor features using SVMs, however Ligang and Dian [29] classified localized Gabor features. Detailed studies on the use of neural networks for AU classification were performed in [39] using both geometric features and Gabor filters.

Recent work is focused on the modeling and recognition of facial expressions from video streams. One of the rationales of using video streams over individual static images is the availability of temporal data. The temporal characteristic of video streams for facial expression recognition has been explored by [19], [27], [32], [33] using hidden Markov models (HMMs). Studies using temporal data with static classifiers were also explored, but they fundamentally lack the ability to learn from previous events. Comparative analyses between the use of static and dynamic classifiers was performed by [19], suggesting that dynamic classifiers are more suited for systems that are person-dependent. A comparative study of using SVMs and HMMs for classifying frustrated and delighted smile expression classification demonstrated that the SVMs outperformed HMMs [18]. The recognition of prototypical facial expressions from time-variant image sequences was also investigated using recurrent neural networks (RNNs) [16], [17], [23], [34], [36]. Kobayashi and Hara [23] indicated that RNNs based dynamic recognition of facial expressions had similar performance as that of facial expressions recognized by humans. Peter and Aggelos [34] used facial action parameters from the MPEG-4 standard to model six basic expressions using

HMMs. Graves et al., [16] came up with a complete system for the automatic recognition of facial expressions using model based image interpretation and sequence labeling using RNNs. Recently dynamic Bayesian networks (DBNs) were also used to account for the temporal changes in facial action development [44,45].

The use of dynamic classifiers such as HMMs in previous works suffered from a well-known drawback. Hidden Markov model assumes that the probability of each observation depends only on the current state, which limits the understanding of a temporal observation. We hypothesize that learning from temporal data using RNNs will improve the performance of a recognition model as the model gains knowledge from the past and current states. Recurrent neural networks will be able to take the contextual information in to account while recognizing an AU.

In this paper, we classified 11 commonly depicted FACS AUs of upper and lower face regions. The selection of AUs is based on the availability of minimum number of samples per AU in the Cohn-Kanade database [22] and also on their presence on the facial region. As the facial regions were segmented into upper and lower faces, AUs lying in the mid region such as AU 9 (nose wrinkle) were omitted. The first task involves the collection of frames from continuous video streams, which depict an AU. Preprocessing, feature extraction and classification are then performed.

Feature extraction is performed using Gabor filters on the images. Resultant feature vectors are then passed on to the classifiers. Classification of AUs is performed using both SVMs and RNNs. The advantage of using SVMs is in their ability to handle high dimensional data without any effect on the models performance [4]. The use of SVMs also provides a baseline for the evaluation of our RNNs based FACS AU recognition.

2. FACS AU DATABASE

The Cohn-Kanade FACS database [22] has served as a benchmark database to evaluate performance of different approaches to facial expression recognition. The database contains video streams from 97 subjects performing a single AU or combination of AUs. The age of the subjects ranged from 18 to 30 years. 65% of the subjects were females, 15% were African-American and 3% Asian. The subjects were asked to directly face the camera and perform the assigned 23 facial displays. Each facial display may be a single AU or a combination of AUs. Each video sequence begins with a neutral face and ends with a target display of the assigned AU or AUs. The last frames of the video stream display an activated AU at its peak intensity. Image sequences were digitized into 640*480 pixel arrays with 8-bit precision for gray-scale images. Over 17% of the data is comparison coded by a second FACS coder.

3. BACKGROUND AND TECHNIQUES

3.1 Gabor Filters

In the past decade or so, wavelet decompositions have been proposed as an alternative to holistic analysis in the field of facial expression analysis. The overall efficiency of a FER model is highly influenced by the feature extraction from the raw data. Gabor filters are widely used for feature extraction owing to their effectiveness compared to geometry-based methods [9].

They are computed by filtering the input image with a Gabor filter that is tuned to a particular frequency and orientation. The response image of the Gabor filter when applied on an input image $I(x)$ using the Gabor kernel is

$$a_k(x_0) = \int I(x) p_k(x - x_0) dx$$

where the Gabor filter $p_k(x)$ is given as

$$p_k(x) = \frac{k^2}{\sigma^2} exp\left(-\frac{(k^2x^2)}{2\sigma^2}\right)\left(exp(ik.x) - exp\left(\frac{-\sigma^2}{2}\right)\right)$$

where k is the characteristic wave vector. $p_k(x)$ is a complex quantity where the magnitude changes very slowly with position where as the phases are very sensitive. For this reason only the magnitude is used for FER.

Facial features for expression recognition are extracted as a set of multiple Gabor filters tuned to different characteristic frequencies and orientations. The combined response from all the Gabor filters is called a *Gabor jet*. In FER, different combinations of frequencies and orientations can be used. However, the use of five different spatial frequencies and eight different orientations were suggested to be optimal for the extraction of the facial features [9]. For a $5 * 8$ Gabor jet, orientations range from 0 to π differing by $\frac{\pi}{8}$ and five spatial frequencies with wave number $k_i = |k| = \left\{\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{8}, \frac{\pi}{16}, \frac{\pi}{32}\right\}$.

Feature extraction through Gabor filters is usually carried out on a whole face image or at a few selected fiducial points. The success of Gabor approach is governed by (i) the amount of data a classification model can handle, (ii) the inter AU correlation between the AUs in the facial expression database and (iii) the set of AUs that are classified. Recognition models employing feature extraction on the whole face image [4], [37] and at few fiducial points [31], [39], [47] were found to be successful in recognizing facial expressions.

3.2 Support Vector Machines

Support vector machines were introduced by Vladimir et al, [8]. They are one of the highly researched upon models in machine learning literature. The basic idea is to maximize the distance between two classes in the input space that one wishes to classify. Support vector machines enjoy several advantages such as handling high dimensional data without any considerable affect on the training time, power and flexibility offered through the use of kernel trick. In the field of FER, this capability of SVMs proved to be very helpful in handling high dimensional data generated with feature extraction techniques such as Gabor filters [11].

Support vector machines try to classify whether a test sample belongs to one of the two classes, defined by the given training data. Instead of single data points, data point is viewed as a p-dimensional vector. Training samples are of the form: (x_i, y_i) where $i = \{1,2, \ldots n\}$ and $x_i \epsilon \Re^d$, $y_i \epsilon (-1, +1)$ where (x_i) are called the co-variants or input vector and (y_i) the response variables or labels. Assuming that the data is linearly separable, there exists a hyperplane H consisting of a set of points x such that

$$H: w.x + b = 0$$

where . denotes the dot product. Vector w is a normal vector perpendicular to the hyperplane and the parameter $b/||w||$ is the offset of the hyperplane H from the origin along the normal vector w. This hyperplane is able to divide the points having $y_i = 1$ from those having $y_i = -1$. For the above hyperplane H, we can formulate an infinite number of equations by scaling both w and b. Here we need to choose the maximum margin (the distance between the hyper planes) so that they are as far as possible while still separating the data.

3.3 Recurrent Neural Networks

Recurrent neural networks have dynamic characteristics and

can model time variant data. The ability of RNNs to treat and store time depend information enables them to learn space-time relationships. This ability to learn space-time relations makes RNNs useful in speech recognition [1], signature verification [2] and forex forecasting [24], where the examples are space-time patterns. First-order RNNs are one of the most widely used RNN architectures. These networks use context units to store previous time step data with a time lag as shown in Fig. 1.

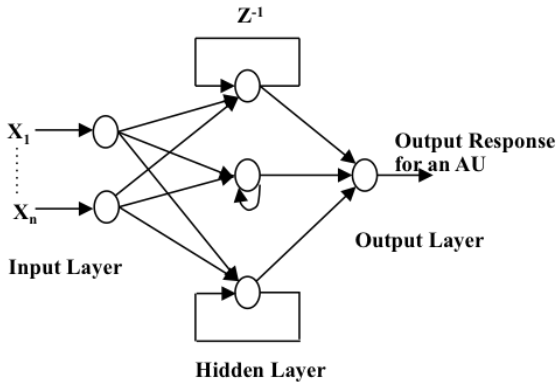

Fig 1: First order Elman recurrent neural network

The hidden unit activations at time $t+1$ is given as

$$S_i^{t+1} = g\left(\sum V_{ij} I_j(t) + \sum W_{ij} S_j(t)\right)$$

where k is the number of input units, and N the number of state units, V_{ii} and W_{ii} are the weights associated with the input and state neurons respectively. $g()$ is the transfer function. I_j and S_j are the output values of the input and state neurons at time t.

Commonly implemented network variants of first-order RNNs are Elman [12] and Jordan [21] based on the values taken by the context units. In Elman architecture [12], context units take the output values of the state units with a time lag, as shown in Fig. 1. Popular network training algorithms for RNNs include back-propagation through time (BPTT) learning algorithm [42] that implements gradient descent and real time recurrent learning (RTRL) algorithm [43]. In supervised learning methods the role of the training algorithm is to adjust the system weights so that the output values at the output nodes are equal to the target values at specific time. In the recent times RNNs have been used for the recognition of dynamic recognition of facial expressions [16], [17]. It was shown that the performance of RNNs is similar for both dynamic recognition of facial expressions and human recognition of facial expressions.

4. FACIAL ACTION RECOGNITION

This section provides a detailed overview of our experiments conducted using SVMs and RNNs. A comparative analysis on the performance of RNNs over SVMs is performed to highlight the use of RNNs.

4.1 Data Collection

In this work, we used a subset of the Cohn-Kanade database consisting of 300 video sequences from 78 human subjects for the recognition of upper face AUs and a total of 258 video sequences from 67 subjects for the recognition of lower face AUs. The number of sample video sequences is a reflection of the limited number of AUs that were classified by our model. The first and last frames of the video sequences are collected to represent static data, as shown in Fig. 2. A static image is represented by one single snap shot depicting a neutral face or an AU at its peak intensity. In the sample shown, the first frame is an instance of neutral face and the last frame is an instance of an activated AU at its peak intensity.

Fig 2: First and last two frames from a video sequence.

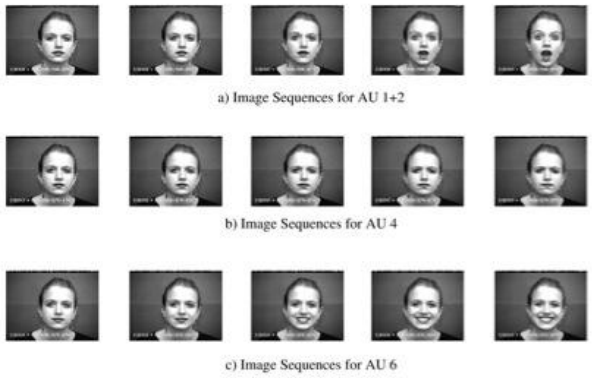

Fig 3: Image sequences for AU 1+2, 4 and 6 from Cohn-Kanade database. Note that the image sequences also depict other AUs acting simultaneously.

In contrast to static images, an image sequence is formed using all or a part of frames present in the video clip. Fig. 3 shows a sequence of five video frames depicting different AU and AU combinations. These sequences also depict other AUs acting simultaneously. However, the presence of other AUs is ignored when looking for the target AU. Video samples consisting of large number of frames were sampled at different intervals for representing change in the appearance of an AU. Frame selection is significant in providing a better understanding of AU activation. However, there is no research that suggests an optimal number of frames that can be used for depicting an AU. A general rule of thumb is, the frames that can be used depends on availability and activation of an AU or AU combination under consideration. Video sequences are used with RNNs as they are able to extract the contextual information from these samples. Support vector machines lack this ability and thus use single static images. The number of image sequences collected and used for our experiments is given in Table 1. The number of static images used with SVMs will be proportional to the figures shown in Table 1, as SVMs use only the first and last frames. Because few image sequence samples and RNNs require sufficient training data we performed subject independent three fold cross validation as opposed to widely used ten fold cross validation used in FER literature.

Table 1. Number of samples (frame sequences) collected from CK database for each AU

AUs	No. Of Samples	AUs	No. Of Samples
AU1	214	AU15	104
AU2	152	AU17	154
AU4	106	AU20	96
AU5	134	AU25	236
AU6	134	AU27	134
AU7	74		

4.2 Preprocessing

The effectiveness of feature extraction techniques such as Gabor filters is highly sensitive to the pre-processing steps implemented on the data. For FER, in general, the ideal requirements of a preprocessing technique involve normalizing the intensity, shape and size of the images present in the database. In this work, pre-processing step involves: 1) facial feature point detection such as eye centers; 2) face normalization using eye coordinates and 3) face detection and cropping. For step 1, we locate the eye coordinates and use them to rotate the face to line up eye centers. In case of image sequences eye coordinates located in the first frames are used for all the subsequent frames. This preserves head movements in the image sequences, which is absent in SVM based model. The face is then detected using MPISearch [15] and cropped to a size of 64 * 64 pixels. Cropped face region is further segmented into upper and lower face regions depending on the AU to be classified. Segmentation into lower and upper face region aids in reducing the dimensionality of the input data and is based on the knowledge that activation of upper face AUs results in very little or no appearance changes in the lower face region and vice versa [10].

4.3 Feature Extraction

In this work, we applied Gabor filters to extract features from static images and sequence of video frames. A set of five frequencies and eight orientations were chosen as in [38]. The filtering of each image using a 5 * 8 Gabor jet leads to a set of 40 Gabor coefficients for every point in the image. Filtering a cropped image of size 64*32 with a 5*8 Gabor jet results in 64*32*40 Gabor coefficients, which incur high computation costs during processing. In order to reduce the dimensionality of the data, the output responses of the 40 Gabor filters was down sampled by a factor of 16 and normalize to unit length as carried out as in [9].

4.4 FACS AU Recognition using SVMs

A single SVM was trained to detect the presence/absence of each AU. Presence of an AU was detected irrespective of its occurrence in combination with other AUs. In this work, no attempt was made to account for non-additive AU combinations that occur in the Cohn-Kanade database. An AU was regarded to be active if the output from the trained classifier was equal or greater than zero. The data for each AU was divided into disjoint subsets for training and testing and three-fold cross validation was performed. The recognition rate in each fold was calculated as the number of samples correctly classified by total number of samples. The final recognition rate for an AU was the average over the three folds. Individual recognition and false alarm rates for upper and lower face AUs are given in Table 2. Our SVM based model achieved an average recognition rate of 79.47% with a false alarm rate of 9.22% for 11 FACS AUs.

Table 2. Recognition results for six upper face and five lower face AUs using SVM

AUs	Recog. Rate(%)	False Alarm Rate(%)
AU1	85.23	9.08
AU2	89.49	6.72
AU4	77.72	6.84
AU5	79.99	7.62
AU6	76.03	7.24
AU7	73.59	11.87
AU15	70.12	15.24
AU17	69.47	11.73
AU20	81.38	6.39
AU25	88.40	10.02
AU27	84.29	7.59
Average	**79.47**	**9.22**

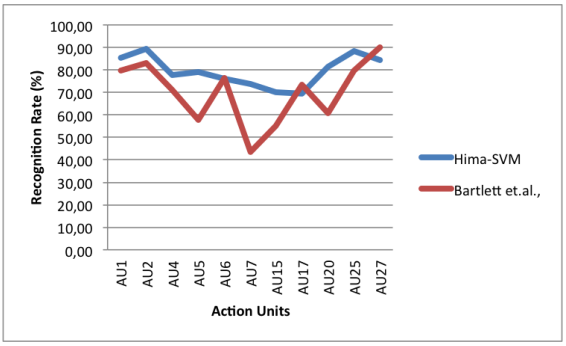

Fig 4: Comparison between our SVM and Bartlett et al., SVM based FACS AU recognition models

Fig. 4 shows the performance of our model in comparison with Bartlett et al, [4] for recognition of FACS AUs in novel subjects. From Fig. 4 it is clear that our SVM model when compared to one in [4] gave similar performance for some AUs such as AU6, AU 17 and AU25. A difference of 2-3% is natural owing to different configurations of the classifier. It is however, worthy to note that our SVM based model was able to outperform for AUs such as 5, 7 and 20. We attribute this to the segmentation of facial region into upper and lower face regions in our work. Segmentation of facial regions was absent in [4]. Recognition of AU7 for example does not need facial feature information from lower half region. Our model was limited to use only one sample from each subject, had

equal number of positive and negative samples in contrast to more than one sample collected from each subject, along with the use of larger number of negative samples used in [4]. We also performed fewer preprocessing steps in comparison to other well-known models.

4.5 FACS AU Recognition using RNN

In this section, we discuss RNNs as a classification model in conjunction with image sequences for the recognition of 11 upper and lower face FACS AUs. Extracted features from each frame of the sequence are fed to the RNN classifier at each time step. For example, features extracted frame 0 is fed at time t=0, frame 1 at time t=1 and so on. Our RNN based classifier unfolded in time is as shown in Fig. 5.

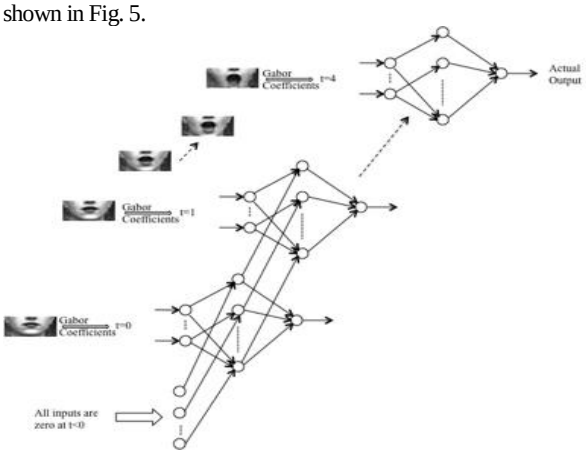

Fig 5: Elman RNN based classifier unfolded in time.

A single RNN was trained to detect the presence of a single AU. The number of input units was equal to the number of Gabor coefficients (i.e. 5120). The number of context units was equal to number of input units. Determining the optimal number of hidden neurons is done using trail and error, as there is no algorithm for the same. The objective is to find the lowest number of hidden nodes that produce the lowest possible error in the test data. Too few hidden nodes will produce a network where the relations between the variables in the input data are not fully learned. On the other hand, too many hidden nodes will over-fit the training data producing poor results when presented with the test data. In the experiments where two network models with different number of hidden nodes gives similar performance with no significant difference, the network model with fewer number of hidden nodes is considered. In our experiments hidden nodes in the range between 1 to 20 were used. The performance of our model is shown in Fig. 6. From the experimental results it is observed that 15 hidden nodes are optimal. However, it should be noted that the number of hidden units which was found to be best for our model may be just one of the local minima. Finding the global minima for any model is difficult in many real world scenarios. The general rule of thumb is to find local minima and see if the recognition model is performing well enough for a given application.

A single output unit was used to notify the presence/absence of an AU. The value from the output node was only considered once all the frames in the sequence were read. Values at the intermediate stages were made equal to zero, as we are only interested in the activation of an AU from its absence to its presence at its peak intensity. The model was trained to output a "1" if the target AU was present and a "0" if the target AU was absent. A threshold value of 0.5 or greater was taken as the target AU present and a value less than 0.5 was taken as target AU absent. The network weights were set

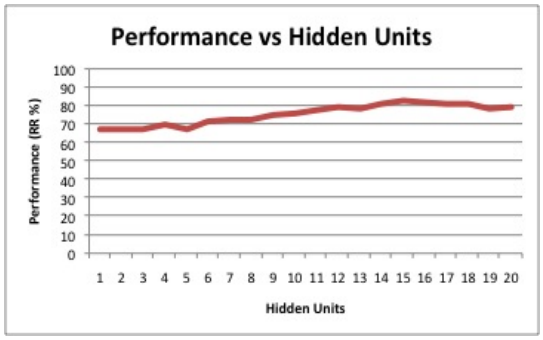

Fig 6: Experiments for finding optimal hidden units

to random initial values in the range of [-1,1]. The maximum number of epochs, where an epoch corresponds to the presentation of all the training samples to the network, was set to 500 and a learning rate of 0.01 was used. The learning process continued until the pre-defined system error was reached or the limit on maximum epochs was elapsed. The number of samples used for testing along with the number of samples recognized correctly and the number of false alarms are given in Table. 3.

Table 3. Total number of samples used for testing (NS), number of samples recognized correctly, positive rate (PR) and the false alarms (FA)

AUs	NS	PR	FA
AU1	74	66	37
AU2	50	46	2
AU4	34	27	3
AU5	44	37	2
AU6	46	35	3
AU7	26	19	3
AU15	36	25	4
AU17	52	37	5
AU20	32	26	1
AU25	36	32	1
AU27	46	40	3

As the number of samples used in each fold is slightly different, they don't exactly map on to the average AU recognition. The average individual recognition and false alarm rates for both upper face and lower face AUs using RNNs as classifiers are given in Table 4.

An average baseline recognition rate of 82.57% with a false alarm rate of 7.61% was achieved for the 11 FACS AUs. The average recognition and false alarm rate achieved using SVMs was 79.47% and 9.22%, respectively. The percentage increase in the recognition rate was over 3% and this increase supports our use of RNNs for FACS AU recognition. The RNN based

model using temporal information provided by the image sequences, clearly carries the advantage of better classification with a lesser false alarm rate. This supports our hypothesis that RNN based FACS AU recognition using image sequences performs better than the use of SVMs and single static images.

Table 4. Recognition results for six upper face and five lower face AUs using RNNs

AUs	Recog. Rate(%)	False Alarm Rate(%)
AU1	89.96	5.01
AU2	92.66	4.20
AU4	81.50	7.54
AU5	83.59	6.00
AU6	77.56	8.20
AU7	75.80	9.18
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	**82.57**	**7.67**

In performance analysis, statistical significance plays an important role. There was a considerable increase in the performance in terms of recognition rate, yet it was not statistically significant. One reason for this would be the small number of samples available for each AU. Expect for a couple of AUs, all others had a total of 80 or less samples for training and testing. This indicates the need for a larger FACS AU annotated database.

The generalization ability of an RNN depends on the balance between information available in the training examples and the complexity of the networks [6]. A very complex network for generalizing a training set containing little information will lead to over-fitting of the data where as a simple network for generalizing a training set containing complex relations will under-fit the data. Both these scenarios will lead to bad generalization [26]. A successful way of avoiding the above mentioned problem is to limit the growth of the weights through some kind of weigh decay [26]. This mechanism will prevent the weights from growing too large. A parameter l, known as "decay constant" determines the penalizing effect on large weights. Weight decay introduced by Werbos [41] decreases the weights while training them through back propagation.

In an attempt to improve the performance of our model we employed weight decay. Three different decay constants are evaluated for obtaining optimal performance. The average recognition rate and false alarm rate for the six upper face and five lower face action units using three different decay constants are given in Table 5. It is concluded that a decay constant of 0.0001 works better in terms of recognition and false alarm rates for both upper and lower face AUs. The individual recognition and false alarm rate for the upper and lower face AUs with a decay constant of 0.0001 are given in Table 6.

Table 5. Recognizing upper and lower face AUs using different DC's

Decay constant	Upper Face		Lower Face	
	RR(%)	FAR(%)	RR(%)	FAR(%)
0.0001	85.84	6.41	84.91	6.07
0.001	78.23	14.48	81.20	9.96
0.01	No Convergence		No Convergence	

Table 6. Upper and Lower face AU recognition using RNNs and weight decay

AUs	Recognition Rate (%)	False Alarm 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
AU15	78.13	8.89
AU17	74.64	14.23
AU20	89.90	0.0
AU25	92.8	4.2
AU27	91.1	3.04
Average	**85.38**	**6.24**

From the results, weight decay in general improved the overall performance of our FACS AU recognition model. Using weight decay the recognition model achieved an average recognition rate of 85.38% with a false alarm rate of 6.24%.

A comparative analysis with other recognition approaches found in the literature would also help validate our results. Fig. 7 depicts the average recognition rates by our SVM based model, RNN based model and some of the best performing models found in the literature such as Bartlett et al, SVM based model [3] and Tian et al, neural network based model [38]. The unavailability of individual recognition and false

alarm rates my other researchers makes it difficult to do a constructive analysis. Our RNN based model clearly outperformed our and Bartlett et al, SVM based models. In comparison to Tian et al's model, our models performance was better for AUs 2,5 and 7 where recognition of lower face AUs was well below.

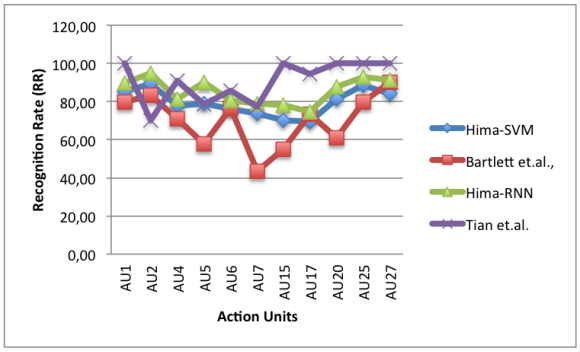

Fig 7: Comparison between different methods.

Comparing our model's performance with Yan et al.'s work [44], as shown in Fig. 8 we conclude that on average our RNN based model performs similar to that [44]. In particular RNN based model performance was better for AUs 1,2,5 and 25. For most of other AUs there was a difference of 2-3%, which can be attributed to different datasets used for testing.

The future work would involve using geometric feature techniques that will reduce dimensionality of the data as huge dimensionality of the data may have an impact on the generalization capability of the model. The classification of samples with out-of-plane head movements would also be explored.

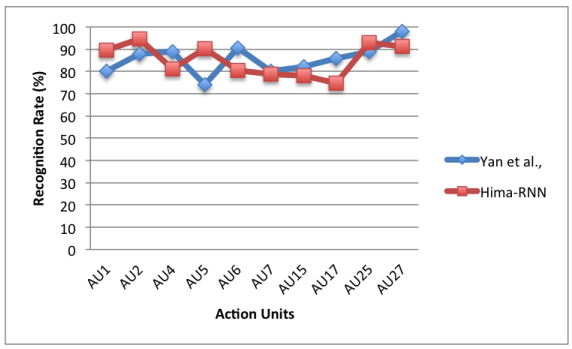

Fig 8: Comparison between RNN based and Yen et al., DBN based models

5. CONCLUSION

Results indicated that the use of RNNs along with image sequences improve the recognition of 11 FACS AUs under review when compared to the use of SVMs with single static images. The increase in the performance using optimized RNNs was nearly 6% when compared to the use of SVMs with a complementary drop in the false alarm rate. This indicates that the use of image sequences does provide a better recognition of an AU by the classification model. Present work was limited to slight in-line and no out-of-plane head movements. In our view, the real advantage of using RNNs will be more pronounce in the above scenarios. As RNNs are capable of learning long sequences of frames (±15), it would help the classification model to recognize the presence of AUs with some missing/occluded data.

6. ACKNOWLEDGMENTS

The author thanks School of Computer Science, University of Western Cape.

7. REFERENCES

[1] Ahmad, A.M., Ismail, S., and Samaon, D.F. 2004 Recurrent Neural Network with Backpropagation through Time for Speech Recognition. International Symposium on Communications and Information Technologies 2004.

[2] Ali, G., Feyzullah, T., Kader, E., and Serdar, C. Signature Verification Performance of Elman's Recurrent Neural Network. Technology, vol. 7, pp. 541-547, 2004.

[3] Bartlett, M.S., Littlewort, G., Braathen, B., Sejnowski, T.J., and Movellan, J.R. A Prototype for Automatic Recognition of Spontaneous Facial Actions. Advances in Neural Information Processing Systems, vol. 15, MIT Press-S. Becker and K Obermayer(eds.), 2003.

[4] Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J. Machine Learning Methods for Fully Automatic Recognition of Facial Expression and Facial Actions. Proceedings of the IEEE Conference on Systems, Man and Cybernetics, Netherlands, 2004.

[5] Bartlett, M.S., Littlewort, G.C., Frank, M.G., Lainscsek, C., Fasel, I.R., and Movellan, J.R. Automatic Recognition of Facial Actions in Spontaneous Expressions. Journal of Multimedia, vol. 19, no. 6, 2006.

[6] Baum, E.B., and Haussler, D. What Size Nets Get Valid Generalization. Neural Computation, vol. 1, pp. 151-160, 1989.

[7] Black, M.J., and Yacoob, Y. Tracking and Recognizing Rigid and Non-Rigid Facial Motions using Local Parametric Models of Image Motion. Proceedings of Fifth International Conference on Computer Vision, pp. 374-381, 1995.

[8] Boser, B.E., Guyon, I.M., and Vapnik, V.N. A Training Algorithm for Optimal Marginal Classifiers. D. Haussler Editor, 5th Annual ACM Workshop on COLT, ACM Press, pp. 144-152, 1992.

[9] Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., and Sejnoeski, T.J. Classifying Facial Actions. IEEE Transactions on Patterns Analysis and Machine Intelligence, vol. 21, issue 10, pp. 974-989, Oct. 1999.

[10] Ekman, P., and Friesen, W. The Facial Action Coding System: A Technique For the Measurement of Facial Movement. Consulting Psychologists Press, San Francisco, CA, 1978.

[11] Ekman, P., Friesen, W., and Hager, J.C. Facial Action Coding System (FACS). A Human Face, Salt Lake City, 2002.

[12] Elman, J. Finding Structure in Time. Cognitive Science, vol. 14, pp. 179-211, 1990.

[13] Essa, I.A., and Pentland, A.P. Coding, Analysis, Interpretation and Recognition of Facial Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 757-763, July 1997.

[14] Fadi, D., and Franck, D. Facial Expression Recognition in continuous Videos using Linear Discriminant Analysis. Proceedings of IAPR Conference on Machine Vision Applications, pp. 277-280, May 2005.

[15] Fasel, I., Dahl, R., Hershey, J., Fortenberry, B., Susskind, J., and Movellan, J.R. Machine Perception Toolbox. Machine Perception Laboratory, University of California San Diego.

[16] Graves, A., Mayer, C., Wimmer, M., Schmidhuber, J., and Radig, B. Facial Expression Recognition With Recurrent Neural Networks. Proceedings of the International Workshop on Cognition for Technical Systems, Germany, 2008.

[17] Hai Tao, Chen, H., and Huang, T. Analysis and Compression of Facial Animation Parameters Set (FAPs). IEEE First Workshop on Multimedia Signal Processing, Princeton, USA, pp. 245-250, June 1997.

[18] Hoque, M.E., McDuff, D.J., and Picard, R.W. Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles. IEEE Transactions on Affective Computing, vol. 3, issue 3, 2012.

[19] Ira, C., Nicu, S., Ashutosh, G., Lawrence, S.C. and Thomas, S.H. Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. Computer Vision an Image Understanding, pp. 160-187, 2003.

[20] Jacob, W., and Christian, W. O. Haar Features for FACS AU Recognition. Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition, 2006.

[21] Jordan, M.I. Attractor Dynamics and Parallelism in a Connectionist Sequential Machine. Proc. of the Ninth Annual Conference of the Cognitive Science Society, Lawrance Erlbaum, pp. 531-546, 1986.

[22] Kanade, T., Cohn, J., and Tian, Y. Comprehensive Database for Facial Expression Analysis. Proceedings of 4th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46-53, 2000.

[23] Kobayashi, H., and Hara, F. Dynamic Recognition of Basic Facial Expressions by Discrete-time Recurrent Neural Network. Proceedings of the International Joint Conference on Neural Networks, pp. 155-158, 1993.

[24] Kondratenko, V.V. and Kuperin, Yu. A. Using Recurrent Neural Networks To Forecasting of Forex. Disordered Systems and Neural Networks, April 2003.

[25] Kotsia, I. and Pitas, I. Facial Expression Recognition in Image Sequences using Geometric Deformation Features and Support Vector Machines. IEEE Transactions on Image Processing, vol. 16, no. 1, pp. 172- 187, 2007.

[26] Krogh, A., and Hertz, J.A. A Simple Weight Decay Can Improve Generalization. Advances in Neural Information Processing Systems 4, J. E Moody, S J Hanson and R P Lippmann eds. Morgan Kauffmann Publishers, San Mateo CA, pp. 950-957, 1995.

[27] Lien, J. Automatic Recognition of Facial Expressions using Hidden Markov Models and Estimation of Expression Intensity. PhD dissertation, Carnegie Mellon University, Pittsburg, PA, 1998.

[28] Lien, J.J., Kanade, T., Cohn, J., and Li, C. Automated Facial Expressions Based on FACS Action Units. Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 390-395, April 1998.

[29] Ligang, Z., and Dian, T. Facial Expression Recognition using Facial Movement Features. IEEE Transactions on Affective Computing, 2011.

[30] Littlewort, G., Bartlett, M.S., Fasel, I., Susskind, J., and Movellan, J. Dynamics of Facial Expression Extracted Automatically from Video. IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Face Processing in Video, vol. 5, pp. 80, June 2004.

[31] Lyons, M.J., Budynek, J., Plante, A., and Akamatsu, S. Classifying Facial Attributes using a 2D Gabor Wavelet Representation and Discriminant Analysis. Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition, Grenoble France, IEEE Computer Society, pp. 202-207, 2000.

[32] Oliver, N., Pentland, A., and Berard, F. LAFTER: A Real-time Lips and Face Tracker with Facial Expression Recognition. Pattern Recognition, vol. 33, no. 8, pp. 1369-1382, 2000.

[33] Otsuka, T., and Ohya, J. Spotting Segments Displaying Facial Expression from Image Sequences using HMM. In IEEE Proceedings of the Second International Conference on Automatic Face and Gesture Recognition (FG98), Nara, Japan, 1998, pp. 442-447, 1998.

[34] Petar S. A., and Aggelos K. K. Automatic Facial Expression Recognition Using Facial Animation Parameters and Multistream HMMs. In IEEE Transactions on Information Forensics and Security, vol 1, No. 1, pp. 3-11, March, 2006.

[35] Philipp M., and Rana E. K. Real Time Facial Expression Recognition in Video Using Support Vector Machines. In Proceedings of the 5th International Conference on Multimodal Interfaces, ICMI03, pp. 258-264, 2003.

[36] Rosenblum M., Yacoob Y., and Davis L. Human Expression Recognition from Motion using a Radial Basis Function Network Architecture. In IEEE Trans. Neural Networks, Vol 7, No. 5, pp. 1121-1138, 1996.

[37] Smith E., Bartlett M.S., and Movellan J.R. Computer Recognition of Facial Actions: A Study of Co-articulation Effects. In Proceedings of the 8th Annual Joint Symposium on Neural Computation, 2001.

[38] Tian Y., Kanade T., and Cohn J. F. Recognizing Action Units for Facial Expression Analysis. In IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 23, No. 2, pp. 97-115, 2001.

[39] Tian Y., Kanade T., and Cohn J. F. Evaluation of Gabor Wavelet Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity. In Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 229-234, May, 2002.

[40] Theekapun C., Tokia S., and Hase H. Facial Expression Recognition from a Partial Face Image by Using Displacement Vector. In the Proceedings of 5th International Conference on Electrical/Electronics, Computer, telecommunications and Information Technology, ECTI-CON, pp.441-444, 2008

[41] Werbos P. Backpropagation: Past and Future. In Proceedings of the IEEE International Conference on Neural Networks, IEEE Press, pp. 343-353, 1988.

[42] Werbos J. Paul. Back Propagation Through Time: What it Does and How to do it. In Proceedings of the IEEE, Vol 78, No. 10, pp. 1550-1560, October, 1990.

[43] Williams R.J., and Zisper D. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. In Neural Computation, Vol 1, No. 2, pp. 270-280, 1989.

[44] Tong, Y., Chen, J., and Ji, Q. A Unified Probabilistic Framework for Spontaneous Facial Activity Modeling and Understanding," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32, no. 2, pp. 258-273,2010.

[45] Tong, Y., Liao, W., and Ji, Q. Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 10, pp. 1683-1699, 2007.

[46] Zhang Z. Feature Based Facial Expression Recognition: Sensitivity Analysis and Experiments With a Multi Layer Perception. Technical Re- port 3354, INRIA Sophia Antopolis, 1998.

[47] Zhang Z., Lyons M., Schuster M., and Akamatsu S. Comparison Between Geometry Based and Gabor Wavelets Based Facial Expression Recognition Using Multi Layer Perception. In Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara Japan, IEEE Computer Society, pp. 454-459, 1998.