FACIAL ANALYSIS IN THE WILD WITH LSTM NETWORKS

Sarasi Kankanamge, Clinton Fookes, Sridha Sridharan

Image and Video Research Lab, Queensland University of Technology
2 George Street, GPO Box 2434, Brisbane, QLD 4001, Australia

ABSTRACT

The promise of computer vision systems to efficiently and accurately recognize faces and facial variations in naturally occurring circumstances still remains elusive. In this paper we present two separate systems for face analysis, both of which use Long Short Term Memory (LSTM) Networks: unconstrained video-based face verification (FaceVideoModel) and spontaneous facial expression recognition (ExpModel). Since LSTM models have influential ability to capture sequential patterns, our results prove such LSTM models have significant advantages over other proposed models in the state-of-the-art for facial analysis in the wild. On the recently introduced Youtube Faces database our FaceModel achieves an accuracy of 98.70% for face verification with a value of 99.94% for the Area Under Curve (AUC) and 1.2% Equal Error Rate (EER) which is the best performance on this database compared to other recently proposed methods. Experimental results achieved through the proposed ExpModel on the challenging FER2013 dataset, including the CK+ database, also demonstrate the effectiveness of our deep model for facial expression recognition.

Index Terms— LSTM, FaceVideoModel, ExpModel

1. INTRODUCTION

The last two decades have seen an escalating interest in methods for automating the coding of face and facial expression in the wild. Such systems will have numerous applications in a wide range of fields including business; national security; robotics; consumer applications; education; mental and physical health; and automotive applications. Enabling machines to accurately and instantaneously recognise faces and facial expressions in the wild data is the key for heralding a long awaited new era in artificial intelligence (AI) where machines (robots/computers/mobile devices) interact, anticipate and plan seamlessly with humans. Yet, despite this keen interest, the reality is that the promise of computer vision systems to efficiently and accurately recognize faces and facial expressions in naturally occurring circumstances still remains elusive [1].

Recently, models based on deep learning, in particular deep Convolutional Neural Network (CNN) have been proposed which yield remarkable performance in object and image classification tasks. These deep CNN architectures have the capacity to leverage the performance of face and facial expression recognition on data acquired in the wild. However, as pointed out in an extensive review [2], current CNN approaches have critical deficiencies: models have relied on large amounts of training data requiring an enormous number of parameters to be learnt.

The aim of our work is twofold. Firstly, we aim to use deep learning techniques to improve face recognition and facial expression recognition accuracy on wild data. Secondly, we also aim to reduce the number of parameters of a deep model, the number of training samples required, as well as the required training time. To achieve this, we propose to use a Long Short Term Memory (LSTM) architecture instead of deep CNNs for facial analysis. LSTMs are flexible models to handle a variable-length sequential data in computer vision applications with lower computation cost. Furthermore, they are a powerful tool for facial analysis with fundamental explanations of their ability to capture sequential patterns. Moreover, major studies [3] have claimed that LSTMs are more effective than conventional CNNs for several classification tasks.

Targeting to improve the classification performance of facial analysis in the wild, we propose a separate deep network learning model by incorporating LSTM networks. We make the following significant contributions in this paper: (1) The development of a LSTM model for video-based face verification in the wild that achieves verification accuracy that outperforms state-of-the-art results on the recently introduced challenging face video database (Youtube faces) [4, 5]; (2) The development of a combined deep CNN model and LSTM model architecture to obtain improved spontaneous expression performance demonstrated on the challenging FER2013 [6] facial expression dataset.

2. RELATED WORKS

As pointed out in an extensive review, almost all previous research on face analysis in the wild has a critical deficiency. This includes that training and testing algorithms have relied on datasets that consist of a small number of subjects, either posed or scripted facial expressions and behaviours performed by an actor, or natural facial behaviours but of

a single person observed in a constrained and artificial set-up. This deficiency is a major reason previous results have transferred poorly to real applications including intelligent surveillance use. Recent research based on the use of deep learning techniques represent a departure from this deficiency and has opened up exciting opportunities to achieve higher levels of performance on the significantly difficult tasks of face verification in the wild and facial expression recognition in the wild.

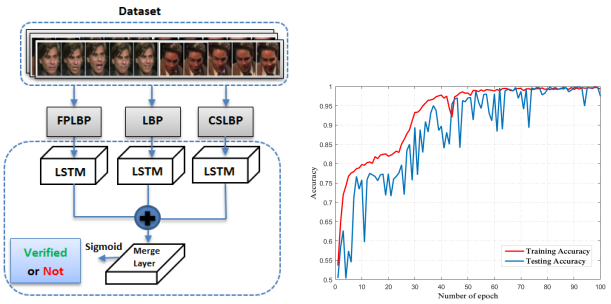
The fundamental thrust of recent studies is that databases deployed are far larger and more diverse representing real-life data. Further, deep learning architectures such as convolutional networks [2], and LSTM networks [3] have shown remarkable performance when coupled with a tradition classifier such as Support Vector Machine (SVM) [7] and/or hypercolumn method [2, 8]. The deep inspired classifiers and features have revolutionized computer vision and speech processing as we know it.

Sun *et al.* [9] introduced a series of DeepID frameworks and obtained higher performance on the challenging and recently introduced Labeled Faces in the Wild (LFW) and Youtube Faces (YTF) [10] in the Wild databases. Their system has the capability to predict about 10,000 face identities. Then, the effective joint deep face identification and verification method is proposed by [11]. Later, Schroff *et al.* [4] proposed a FaceNet model for face verification, recognition, and clustering based on a L2 normalization which is monitored by the triplet loss minimization. Their system achieved the best accuracy on LFW (99.63%) database and the remarkable performance on YTF (95.12%) database, in comparison to the state-of-the-art results. Recently, Parkhi *et al.* [5] proposed a new triplet loss function for learning a face embedding and beat human-level performance achieving a 97.3% verification accuracy rate on YTF dataset.

In parallel, the computer vision community is working to understand how a facial expression recognition system attains high performance whilst introducing several approaches particularly based on the use of deep concepts. The work for spontaneous facial expression recognition using deep learning techniques have been extended by several major evaluation studies. In 2013, Goodfellow *et al.* [12] released a FER2013 database as a solution to the data related problem of facial expression recognition. This database consists of 28,709 training images. The most recent systems [6], first pre-trained their model on a larger FER2013 training set and fine-tuned on other wild datasets with a smaller samples. Recent spontaneous expression models [6, 13, 14] can be considered as a black box, coupled with input dataset (input feature set), deep architecture, classification function with different configurations.

However, almost all existing deep CNN models for facial analysis applications are computationally expensive due to the number of parameters and training time required. While some approaches have shown the promise of LSTMs for facial analysis, they have not yet fully uncovered their ability to exploit sequential information for facial analysis in the wild. In this paper we demonstrate that ability through both video-based face verification in the wild, and facial expression recognition in the wild.

3. THE PROPOSED FACEVIDEOMODEL



(a) Outline of the FaceVideoModel.

(b) Training and testing accuracy curves on YTF.

Fig. 1: FaceVideoModel.

Our goal in face verification is to verify two individuals represented by a pair of video sequences which is known as a video-to-video face matching problem. The input to our method is sequence occurrences of people in wild conditions. In video based face verification we should exploit the temporal identity information that is present. Our work proposes a novel system to capture the spacial-temporal identity information for face verification using multiple LSTM models with several feature descriptors to represent this information of each individual face in the wild.

Temporal information is critical for classification problem and performance of the model. LSTM networks [3] are very powerful forms of Recurrent Neural Network (RNN) architecture to address many sequential problems in computer vision. LSTM network has capability to learn long-term dependencies. Standard RNNs contain a chain of repeating modules with a simple single layer (i.e. simple tanh activation layer). In contrast, the repeating modules of LSTMs consist of four interacting layers.

3.1. Multiple LSTM model overview

Solution to the pair matching problem of image sequences in the wild is very important for demanding applications where subject cooperation and invariance of the imaging conditions cannot be guaranteed. This is a challenging and open issue for current state-of-the-art face recognition systems. We propose a novel direction to address for this distortion using multiple LSTM models analyzing the video face image variations. Here, we consider a sequential inputs and fixed outputs learning strategy with video inputs and a single label (i.e. verified or not verified). An overview of the entire framework is illustrated in Fig.1(a).

We start our approach by extracting several feature descriptors per frame in the video data. Then we directly con-

struct a separate LSTM network for each feature descriptor type. The input to each LSTM network is a combination of matched pairs and mismatched pairs. Finally, these K vectors are concatenated using a merge layer concept, and a fully connected network is trained on top of the concatenated representations. In our case, we use three feature descriptors including Local Binary Patterns (LBP), Center-Symmetric LBP (CSLBP) [10] and Four-Patch LBP (FPLBP) [15] which are already provided with YTF database [10].

4. THE PROPOSED EXPMODEL

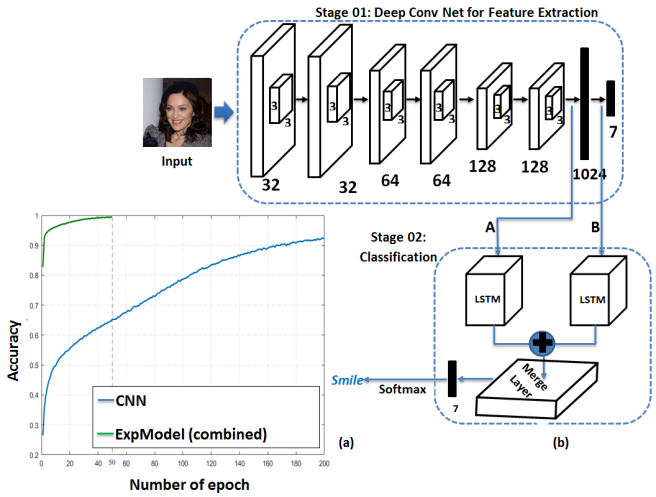


Fig. 2: (a) Training accuracies. (b) Outline of the ExpModel.

Our model for facial expression recognition in the wild (ExpModel) consists of two stages as illustrated in Fig.2(b). It is true that LSTMs are originally introduced for sequence learning. Observing this, we presented a combined CNN (stage 01) and LSTM (stage 02) learning based approach for image-based facial expression recognition in the wild, in order to demonstrate LSTMs utility for expressive image labelling. We analysed the sequential coherency of a single image as explained below: (1) We used the channel-wise feature maps pattern of the last convolutional layer as it keeps longer range dependencies of an image without losing global coherence. (2) Since a non-overlapping input pattern combines the local correlation of the feature unit while maintaining global coherence of the image as sequential data, the last dense feature vector is divided into non-overlapping windows.

4.1. Two-Stage Model Overview

Our CNN network contains six convolutional layers, three max pooling layers (after second, fourth, and sixth convolutional layers), and two dense layers. We train our CNN model with stochastic gradient descent (SGD) optimizer with 0.02 learning rate. Typical CNN models use the output of the last layer as a feature representation. Furthermore, major evaluation studies [8] have made it abundantly clear that hypercolumn-based representation makes significant contribution to improve the object recognition performance. However, this hypercolumn-based systems are computationally expensive compared with typical CNN models. Hence, in order to use the advantages of both concepts, we extract the last convolutional layer features and the dense layer features from our trained CNN model as last layer outputs convey more significant information than initial layer outputs.

As a second stage, we construct a separate LSTM network for each feature set. LSTM network with the last convolutional layer feature set is defined as $LSTM_{conv}$ and LSTM network with the dense layer feature set is defined as $LSTM_{dense}$. In order to obtain sequential inputs and fixed output, the input configuration of each LSTM network is implemented as below: The input shape for $LSTM_{conv}$ (A) is equal to (128,36) and the input shape for $LSTM_{dense}$ (B) is equal to (32,32). Finally, two LSTMs are concatenated using a merge layer.

5. EXPERIMENTAL RESULTS

5.1. FaceVideoModel Evaluation

For face verification experiments, we followed the image restricted protocol to evaluate our model with 5000 video pairs equally to 10 independent splits, with 2500 intra-personal pairs and 2500 inter-personal pairs. We conducted the 10-fold cross validation protocol. Fig.1(b) shows the accuracy curves on YTF training and testing dataset. As the figure illustrates, the model converges after about 100 epochs. Finally, we achieved a classification accuracy (with standard deviations) of $98.70\% \pm 0.5002$ which is a new record performance on YTF dataset. We obtained 99.94% value for the Area Under Curve (AUC). We achieved 1.2% value for the Equal Error Rate (EER) which is the best value on YTF dataset compared to the state-of-the-art performance.

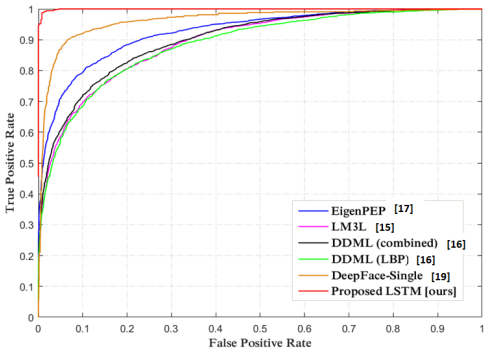


Fig. 3: ROC comparison with the state-of-the-art face verification methods on YTF.

We compared our model with the benchmark face verification approaches. Fig.3 represents the ROC comparison of our model with some of the benchmark methods of the state-of-the-art face verification methods as mentioned in above. Our approach obtains the best results on YTF. As shown in Table 1, outcomes (face verification accuracy with standard deviation, AUC, and EER) of our method is thoroughly

ranked against existing state-of-the-art in benchmarking exercises. Furthermore, we run an experiment with the AlexNet feature descriptor with a single LSTM classification model, in order to demonstrate the contribution of the deep feature to the final classification value. It achieves 93.2% verification accuracy which is relatively less than [4, 5] methods, but this value shows that AlexNet Feature descriptor makes significant contribution to improve the verification performance. In Table 1, * indicates the video-to-video based face verification methods in the wild on YTF. Other methods are the image-to-video based models in the wild which are trained on image face database and tested on YTF.

Method	Accuracy±SE	AUC	EER
LM3L [16] *	81.3±1.2	89.3	19.7
DDML (LBP) [17] *	81.3±1.6	88.7	19.7
DDML (combined) [17] *	82.3±1.5	90.1	18.5
EigenPEP [18] *	84.8±1.4	92.6	15.5
MMMF Fusion [19] *	-	93.9	12.6
DeepFace-Single [20]	91.4±1.1	96.3	8.6
AlexNet+LSTMs (ours) *	93.2±0.6136	-	-
FaceNet [4]	95.12±0.39	-	-
Embedding Learning [5]	97.3	-	-
FaceVideoModel (ours) *	**98.7±0.5002**	**99.94**	**1.2**

Table 1: Comparison of state-of-the-art face verification methods on YTF.

More broadly, our results demonstrate that our LSTM architecture is a powerful tool for sequential face verification in the wild, as previous methods such as LM3L [16], DDML (combined) [17] have achieved relatively poorer accuracies with combined LBP, CSLBP, and FPLBP hand-crafted features.

5.2. ExpModel Evaluation

We first run experiments on the FER training, validation, and final testing sets. During the feature learning process, we configured a CNN network as explained in Section 4. As the Fig.2(a) illustrates, the CNN model converges after about 200 epochs. During the classification process, we configured two LSTM models with the merge layer as shown in Fig.2(b). The model converges after about 50 epochs as shown in Fig.2(a). Our system finally achieved 0.7 accuracy with validation set and 0.715 accuracy with final test set which is significantly effective performance against the baseline methods [6, 7]. Table 2 represents the performance comparison of our proposed model with multiple network learning (MNL) [6] and DLSVM [7] approaches. MNL method obtained the best results on FER2013 wild database which are equal and closer to our results and DLSVM method achieved remarkable performance by replacing softmax function with linear support vector machine function. Our method shows effective results using two LSTMs with a merge layer. Our results show such LSTM models provide significant accuracies by feeding the last convolutional features and the dense features as sequential pattern inputs.

Approach	Validation	Test
DLSVM [7]	0.694	0.712
MNL [6]	≈**0.7**	≈**0.72**
CNN [our]	0.650	-
ExpModel (dense output) [ours]	0.683	-
ExpModel (last conv output) [ours]	0.667	-
ExpModel (combined) [ours]	**0.700**	0.715

Table 2: performance comparison with the benchmarking approaches on FER2013.

Next, we tested the proposed model on CK+ dataset (cross-database validation) and our approach yields an average accuracy of 60.6%. As shown in Fig.4, the prediction accuracy for smile expression is in the higher range and prediction accuracies for disgust and fear expressions are in the lower range.This performance indicates that our method can yield effective results on the controlled databases even when our training dataset is diverse and wild.

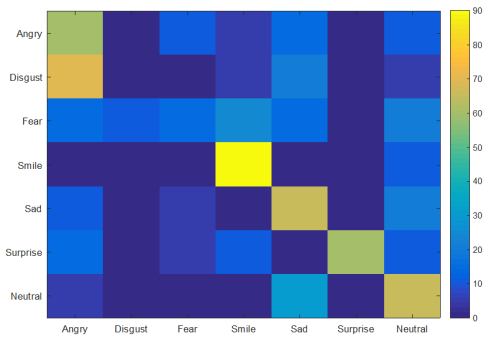


Fig. 4: Confusion matrix of our framework on CK+.

6. CONCLUSION

We have shown the power of LSTM Networks to exploit sequential information for facial analysis in the wild. We have demonstrated this ability through both video-based face verification in the wild as well as spontaneous facial expression recognition. Our proposed models performance is very competitive on challenging wild databases as LSTMs have significant capability to capture sequential information. FaceVideoModel achieved 98.70% face verification on YTF database which is the best performance in the benchmarking exercises. Moreover, ExpModel reported effective performance for spontaneous facial expression recognition on the FER2013, and our ExpModel can yield good results on controlled CK+ database even with more diverse wild training set. Hence the proposed systems have the potential value within the computer vision community for more effectively managing unconstrained facial analysis applications: video-based face verification in the wild and spontaneous facial expression recognition.

7. REFERENCES

[1] Gary B Huang, Honglak Lee, and Erik Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2518–2525.

[2] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *arXiv preprint arXiv:1603.01249*, 2016.

[3] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[4] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[5] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015, vol. 1, p. 6.

[6] Zhiding Yu and Cha Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 435–442.

[7] Yichuan Tang, "Deep learning using support vector machines," *CoRR, abs/1306.0239*, vol. 2, 2013.

[8] Chenchen Zhu, Yutong Zheng, Khoa Luu, T Hoang Ngan Le, Chandrasekhar Bhagavatula, and Marios Savvides, "Weakly supervised facial analysis with dense hyper-column features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 25–33.

[9] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun, "Bayesian face revisited: A joint formulation," in *European Conference on Computer Vision*. Springer, 2012, pp. 566–579.

[10] Lior Wolf, Tal Hassner, and Itay Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 529–534.

[11] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

[12] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al., "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[13] Ali Mollahosseini, David Chan, and Mohammad H Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[14] Xianlin Peng, Zhaoqiang Xia, Lei Li, and Xiaoyi Feng, "Towards facial expression recognition in the wild: A new database and deep recognition system," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 93–99.

[15] Lior Wolf, Tal Hassner, and Yaniv Taigman, "Descriptor based methods in the wild," in *Workshop on faces in'real-life'images: Detection, alignment, and recognition*, 2008.

[16] Junlin Hu, Jiwen Lu, Junsong Yuan, and Yap-Peng Tan, "Large margin multi-metric learning for face and kinship verification in the wild," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 252–267.

[17] Junlin Hu, Jiwen Lu, and Yap-Peng Tan, "Discriminative deep metric learning for face verification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1875–1882.

[18] Haoxiang Li, Gang Hua, Xiaohui Shen, Zhe Lin, and Jonathan Brandt, "Eigen-pep for video face recognition," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 17–33.

[19] Lacey Best-Rowden, Brendan Klare, Joshua Klontz, and Anil K Jain, "Video-to-video face matching: Establishing a baseline for unconstrained face recognition," in *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*. IEEE, 2013, pp. 1–8.

[20] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.