AuthNet: A Deep Learning Based Authentication Mechanism Using Temporal Facial Feature Movements (Student Abstract)

Mohit Raghavendra[1*], **Pravan Omprakash** [2*], **Mukesh B R**[1]
[1] Department of Information Technology
[2] Department of Metallurgical Engineering
National Institute of Technology Karnataka, Surathkal, India, 575025
{mohithmitra, pravanop, mukeshbr1999}@gmail.com

Abstract

Deep learning algorithms are widely used to extend modern biometric authentication mechanisms in resource-constrained environments like smartphones, providing ease-of-use and user comfort, while maintaining a non-invasive nature. In this paper, an alternative is proposed, that uses both facial recognition and the unique movements of that particular face while uttering a password. The proposed model is language independent, the password doesn't necessarily need to be a set of meaningful words or numbers, and also, is a contact-less system. When evaluated on the standard MIRACL-VC1 dataset, the proposed model achieved a testing accuracy of 98.1%, underscoring its effectiveness.

Introduction

Facial authentication systems allow only authorised users to access restricted systems, by verifying user's identity from the person's digital image or a video frame. They have gained widespread adoption over the past few years primarily due to their contact-less/non-invasive nature and ease of use. Another major milestone that further fueled the development of facial recognition systems was the advent of highly accurate deep learning models such as FaceNet (Schroff, Kalenichenko, and Philbin 2015) and Baidu (Liu et al. 2015), that have surpassed human performance. Despite such advancements, face recognition systems have been shown to fail under perturbations and bad lighting, and have often been bypassed by imposters using a photograph of the authorized user to gain access. In view of these limitations, more dynamic models that are trained on lip movement patterns of the user as they utter a particular word were introduced. These are much more secure when compared to basic face recognition systems, while also being immune to noisy environments. Such lip movement based password systems (Liu and Cheung 2012) are however dependent on the language and also to lighting conditions, and have thus far failed to achieve the high accuracy requirements of authentication systems. In this paper, we propose a language-agnostic password based authentication model that captures the temporal facial movement patterns of a person uttering a particular password. It is not hindered by language bias or restrictions as the model is not dependent on the actual word, but is trained on the videos of word utterance. The proposed model was benchmarked on the publicly available standard dataset, MIRACL-VC1 (Rekik, Ben-Hamadou, and Mahdi 2014), and a manually compiled dataset consisting of videos from various smartphone cameras taken under varying conditions, the specifics of which are discussed in subsequent sections.

Proposed Approach

For the experimental validation of the proposed methodology, the standard MIRACL-VC1 dataset was used. Each training and testing sample in the dataset is a sequence of images of the person uttering a password. The images are first passed through a Haar cascade face detector to crop out the face regions, and the sequence is padded with white images and resized to ensure uniformity while feeding it into VGGFace (Parkhi, Vedaldi, and Zisserman 2015). The manually compiled dataset consists of videos that are first segmented into images using a constant frame rate, after which padding and resizing is applied. Each image is passed through the pre-trained VGGFace model. This is done for each word and for each person and reshaped to make it suitable for feeding into the LSTM layer. A network of 4 LSTM layers is employed, each consisting of 20 timesteps and an output sigmoid layer for predicting probabilities. Once trained, this model can determine if the input test video contains the correct person-word combination or not to authorise user. The training is repeated for 5 different speakers, each speaking 7 different words. To avoid presenting results for only one person-word combination and instead provide more comprehensive results, the model was trained in an iterative manner and tested on all possible combinations. The training set provides 35 person-word (5 speakers X 7 words combinations, by assigning each word as a password. The collated dataset consists of videos, each of 2 seconds duration. The videos were obtained from different smartphone cameras. They were captured under different lighting conditions, and the languages used were different (English and Hindi).

Results and Discussion

To quantify AuthNet's performance, a cross validation dataset was prepared, that accounted for the various imposter cases available for the person-word combination,

* denotes equal contribution.

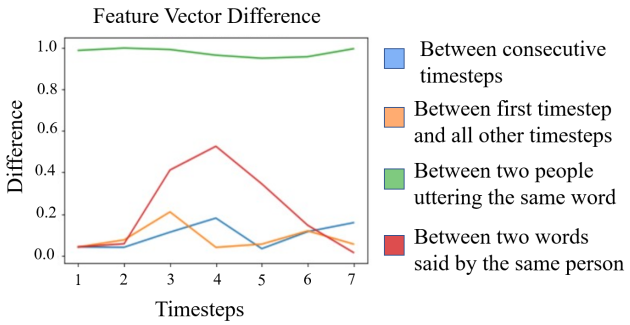

Figure 1: Variations in Features Extracted using VGGFace

Models	Accuracy
AuthNet *(Proposed)*	**0.981**
FaceNet+Gergen *et al.* (Gergen et al. 2016)	0.864
FaceNet+LipNet (Shillingford et al. 2019)	0.951

Table 1: Benchmarking AuthNet against State-of-the-art

namely, the authorised user saying a different word, an unauthorised user saying the correct password and an unauthorised user saying a different word. The different words spoken by the authorised user were also chosen such that they are completely new and hence the capability of the model to generalize to new person-word combinations is thoroughly tested. The feature vectors variations extracted from VGGFace highlight the maximum difference between two people uttering the same word (Fig 1). There is also a significant difference between two words being uttered by the same person, which enables the user to change his password as per his choice. These differences also substantiate the uniqueness of the way a person utters a word or a phrase, compared to how other people may pronounce it. The VGGFace model clearly extracts the required temporal difference and provides a good representation of the variation of facial features across timesteps. This underscores the fact that the proposed model is capable of handling imposters uttering the same password, thus allowing it to be an open password system.

The performance of the proposed model with other hybrid pipelines are demonstrated in Table 1. The performance of the proposed model with other metrics are shown in Table 2. The individual specificity errors for the different imposter cases are as follows. The system denies the attacks from a different person around 98% of the time, while it correctly identifies a wrong password being spoken 94% of the time. This indicates the model's ability to act as an excellent defense mechanism against attacks by malicious entities.

Conclusion and Future Work

In this paper, AuthNet, a temporal facial movement based authentication mechanism using 2-level CNN-RNN deep neural models was presented. It is an end-to-end model that is inherently language and domain agnostic and is a more robust mechanism for authentication that can be used by

Metric	MIRACL-VC1	Our Dataset
Sensitivity	0.998	0.987
Specificity	0.969	0.976
Accuracy	0.981	0.980
AUC Score	0.990	0.989
Equal Error Rate	0.023	0.037

Table 2: Performance of proposed AuthNet model

smart devices. When evaluated on the standard MIRACL-VC1 dataset, AuthNet achieved an accuracy of 98.1%, underscoring its effectiveness. The results obtained on a collated dataset gave good results even when trained with only 10 positive video samples, thereby demonstrating the real-world application of this system. It can function as an open password system, as the model effectively classified the same password being spoken by a different person and has no language barrier, since the model is not aware of the language in which the password/phrase is being spoken. This is an effective improvement of face recognition systems since it is also a contact-less biometric system, which is becoming increasingly important.

Acknowledgements

The authors would like to thank Dr. Sowmya Kamath S, Department of Information Technology, NITK Surathkal for her guidance.

References

Gergen, S.; Zeiler, S.; Hussen Abdelaziz, A.; Nickel, R.; and Kolossa, D. 2016. Dynamic Stream Weighting for Turbo-Decoding-Based Audiovisual ASR. 2135–2139. doi: 10.21437/Interspeech.2016-166.

Liu, J.; Deng, Y.; Bai, T.; and Huang, C. 2015. Targeting Ultimate Accuracy: Face Recognition via Deep Embedding. *ArXiv* abs/1506.07310.

Liu, X.; and Cheung, Y.-m. 2012. A multi-boosted HMM approach to lip password based speaker verification. 2197–2200. doi:10.1109/ICASSP.2012.6288349.

Parkhi, O. M.; Vedaldi, A.; and Zisserman, A. 2015. Deep Face Recognition. In Xianghua Xie, M. W. J.; and Tam, G. K. L., eds., *BMVC*, 41.1–41.12. BMVA Press. doi:10.5244/C.29.41.

Rekik, A.; Ben-Hamadou, A.; and Mahdi, W. 2014. A New Visual Speech Recognition Approach for RGB-D Cameras. 21–28. Cham: Springer International Publishing.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. 815–823. doi:10.1109/CVPR.2015.7298682.

Shillingford, B.; Assael, Y. M.; Hoffman, M. W.; Paine, T.; Hughes, C.; Prabhu, U.; Liao, H.; Sak, H.; Rao, K.; Bennett, L.; Mulville, M.; Denil, M.; Coppin, B.; Laurie, B.; Senior, A. W.; and de Freitas, N. 2019. Large-Scale Visual Speech Recognition. In *INTERSPEECH*, 4135–4139. URL https://doi.org/10.21437/Interspeech.2019-1669.