https://doi.org/10.2352/ISSN.2470-1173.2021.6.IRIACV-318

Concurrent Two-Factor Identify Verification Using Facial Identify and Facial Actions

Zheng Sun, Dah-Jye Lee; Brigham Young University, Provo, UT 84602, USA
Dong Zhang, Xiao Li; Sun Yat-sen University, Guangzhou, Guangdong 510006, China

Abstract

Identity verification is ubiquitous in daily life. Its applications range from unlocking mobile device to accessing online account, boarding airplane or other types of transportation, recording times of arrival and leaving work, controlling access to a restricted area, facility, or vault, and many more. The traditional and the most popular identity verification is password authentication but with many challenges. Human biometric identifiers like fingerprint, retina scan, and 2D or 3D facial features have become popular alternatives. Some applications use two-factor or multi-factor authentication to increase system security, e.g., password and login code sent to a mobile device. All these identity verification methods have their challenges ranging from forgotten or stolen password to unaware or unintentional authentication and complexity and high costs. This paper presents a promising alternative that could be an improvement to the existing identity verification methods. This improved identity verification is a two-factor approach that concurrently analyzes facial features and unique facial actions. The user's facial features and facial actions must both match what have been stored in the system in order to pass identity verification. This two-factor verification requires only the frontal view of the face and authenticates facial features and facial actions concurrently. It generates an embedding of facial features and facial action in a short video for matching. We name this method Current Two-Factor Identity Verification (C2FIV). Two frameworks that use recurrent neural networks to learn the representation of facial features and actions. One uses an auto-encoder, and the other one uses metric learning. Experimental result shows that the metric learning model performs reliably with an average precision of 98.8%.

Keywords - Facial Action Analysis, Access Control, Identity Verification, Deep Learning, Concurrent Two-factor Authentication.

Introduction

Identity verification is ubiquitous in our daily life. Based on its applications, it could be categorized into authentication and authorization. Authentication is a process that controls the access of the data or system by verifying the identity of the user. Similarly, authorization is a mechanism that grants the right to access a resource or facility or to perform certain tasks. Often, authentication requires a user to log in to a system by using some form of credentials which, at a minimum, consist of a user ID and password. Authorization requires a user to swipe their ID card or other hardware devices on the reader, and the system compares the user's identity to a preapproved user list. Both processes are essential to effective security.

The traditional and most popular form of identity verification uses password or personal identification number. It has evolved into two-factor or multi-factor authentication to provide a more secure mechanism to verify that the user is who he or she claims to be. Usually, it combines a piece of secret information, such as a password, with a device the user possesses, such as a smartphone, a code card, or a physical key that must be connected to the verification system. Human biometric identifiers like fingerprint, retina scan, and 2D or 3D facial features have become popular alternatives for identity verification. For example, two-factor authorization using biometric identifiers could include an ID card or password accompanying with fingerprint or facial features.

All the aforementioned identity verification methods, biometric identifiers or physical device, have their weaknesses. For example, the password could be lost, stolen, or phished. Two-dimensional facial recognition could be compromised by a 2D photo or image. Fingerprint could be copied using sticker or special rubber glove. 3D facial recognition could be tricked by fabricated 3D rubber face mask. All these existing and popular biometric-based identity verification methods could be compromised if the user is unaware, sleeping, or even unconscious. An additional level of protection could improve the effectiveness of many types of security systems. In recent years, research on better defending systems against hackers or perpetrators never ceases [1]. Our research aims to increase the level of protection with a single camera and to build a more reliable identity verification system.

This work was motivated by the biggest deficiency of the existing biometric based identify verification methods in which the security could be compromised when the user is unaware or unconscious. Systems using fingerprint, retina scan, and 2D or 3D facial model for identity verification all could be breached while the authorized user is sleeping, unconscious, or even deceased. Just like entering the password in a public place, some may argue that facial action could be revealed to or recorded by a potential hacker if identity verification is done in public. This scenario can only be successful if the attention is deliberate and carefully planned. Unlike entering the password, the recording conditions such as camera angle and distance increase the difficulty for duplicating facial action. Even if mimicking the exact facial action is possible, matching the correct facial action and the exact facial features at the same time post a huge challenge for hackers or perpetrators. Most identity verification applications such as controlling access to a restricted area, facility, or vault will not face this challenge.

The proposed identity verification system is designed to address the weaknesses of the existing biometric approaches and to

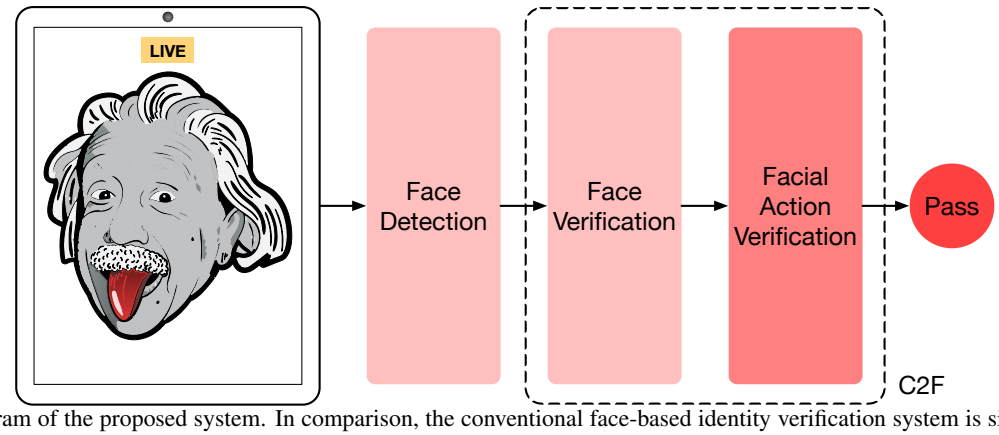


Figure 1: Diagram of the proposed system. In comparison, the conventional face-based identity verification system is single factor. The proposed system employs facial action as the second factor. Access to the system is granted only if both factors are verified.

provide more effective security. This unique algorithm combines facial features and facial actions to perform verification. The algorithm verifies identity by using video input from the camera to perform facial feature recognition and facial action analysis concurrently to increase security. The user must be present in front of the camera and consciously express a unique facial action. Using facial features concurrently with facial actions provides better security than using facial features alone. This unique approach is shown in Figure 1.

The system uses a single camera to acquire the video. It does not require special hardware to obtain a 3D face model or other signatures. In the registration stage, the user shows the frontal face in the camera view and record a short video of a unique facial action or the lip movement from reading a secret phrase. This recorded video is input to the computer for extracting facial features and features of the facial action. The extracted features are then stored in the system for identity verification. In the identity verification stage, the features extracted from the input video is compared to the stored features. The verification is successful if the two sets of features are similar. This method is considered two-factor because it tests both facial features and facial actions for identity verification. We name this new method Recurrent Tw-Factor Identity Verificaiton or C2FIV.

In this work, our focus is on developing a reliable algorithm for facial action analysis. We implemented two dynamic facial action analysis frameworks that use recurrent neural networks to learn the representation of facial actions. One uses an auto-encoder, and the other one uses metric learning. Our result shows that our metric learning model performs reliably with an average precision of 98.8%.

Our contributions can be summarized as follows:

- The development of an identity verification algorithm that combines the inherence factor (facial features) and the knowledge factor (facial action features) to increase the level of security.
- Two machine learning models, one unsupervised and one supervised, were designed to learn the representation of facial action.
- Our metric learning model performs with high precision and demonstrates the feasibility of our innovative idea.

Background

In recent years, the data-driven machine learning approaches, like deep neural networks (DNN), have shown tremendous success in computer vision applications. The deep convolutional neural networks (CNNs) based models have outperformed the traditional methods in almost every visual computing task, especially for image classification[2] and object detection. It has also been confirmed that CNNs work very well for face-related tasks including face recognition[3, 4] and facial expression recognition[5].

The pipeline of the proposed system is designed for facial action verification. Different from facial expressions, customized facial action may not have a key-frame that can outline itself. All frames in the video must be involved in order to represent the facial movement. We use recurrent neural networks (RNNs) to model dynamic facial action. RNN is considered a powerful tool to model sequential data. It produces promising results in many machine learning tasks such as automatic speech recognition[6] and video translation[7].

Related Works

We performed a thorough literature review and patent search. Xie[8] proposed a similar user authentication system utilizing dynamic facial actions. However, this work used Kinect, a depth-sensing camera. Our system only requires a single camera and employs neural networks to process the facial action video without depth information. Some works[9, 10] attempted to enhance identity verification using facial expressions, which is barely a subset of general facial actions. Also, their systems only handled static image input.

Method

The first step of our facial action verification pipeline is detecting the facial landmarks in each frame. These landmarks are normalized and packed into a sequence whose length is equal to the number of frames. The second step is feeding this sequence to RNN. In this work, we developed two models based on RNN. One is the auto-encoder model, which is an unsupervised learning method and can work without labels. The other is a metric learning model. It requires identical tags of samples but can cluster similar samples and distinguish the different pieces.

In both models, the state of RNN cells is used to produce

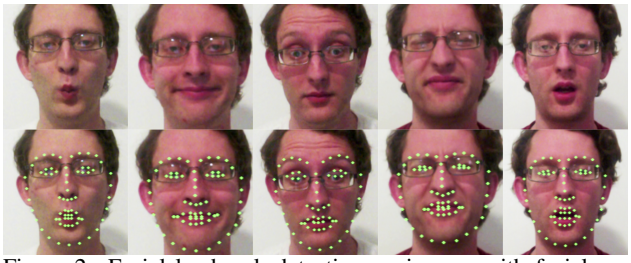


Figure 2: Facial landmark detection on images with facial action. The top five images are the original frames. The second row shows the detected facial landmarks of these frames. The landmark detector works well with face with distinct facial actions.

the embedding for identity verification. The embedding of the user's facial features and facial action at the registration stage is stored on a server. When the authentication process is running, the computer compares the newly generated embedding to the saved one. The verification is passed if the dissimilarity is under a preset threshold.

Facial Landmark Detection

The facial landmark detector used in the system is an ensemble of regression trees[11]. The face region must be detected before landmarks detection. Our face detector uses the classic Histogram of Oriented Gradients (HOG) features[12]. Some recent research[13, 14] showed CNN-based face detectors outperform those using hand-designed features. However, our system is specific for identity verification, in which the head pose is mostly stable, and the image quality is more than adequate. Our test in the lab shows that the traditional detector works perfectly for our needs. Also, some camera modules already have this detector integrated into the processor. The face detection can be done in real time with built-in hardware, and it is power efficient. Once the face region is detected, the landmark detector starts with the face region's centered mean shape. Then the shape is updated at each level of the cascade. We are aware of some CNN-based facial landmark detectors, like FAN[15], which performs well with the non-frontal view and noisy background. For the same reason above, the result from regression trees, shown in Figure 2, is accurate for our application. Interestingly, for some images with intense facial action in our experiments, the ensemble of regression trees worked better than neural networks.

Recurrent Neural Networks

The state of RNN cells is called the hidden state. The hidden state is updated using both the input data at time t and the hidden state at time t-1. The refresh process of plain RNN is the most native form, unable to handle data with long-term dependencies[16]. The facial landmarks in our videos contain long-range contextual information. Therefore, we upgrade the cell with a gated recurrent unit (GRU)[17]. The GRU cellcontains a reset gate and an update gate. The reset gate determines how much of the old memory to forget, while the update gate decides how much of old memory to retain.

Sequence-to-Sequence Autoencoder

We first treat our facial action representation task as a dimension reduction problem. The auto-encoders can transform

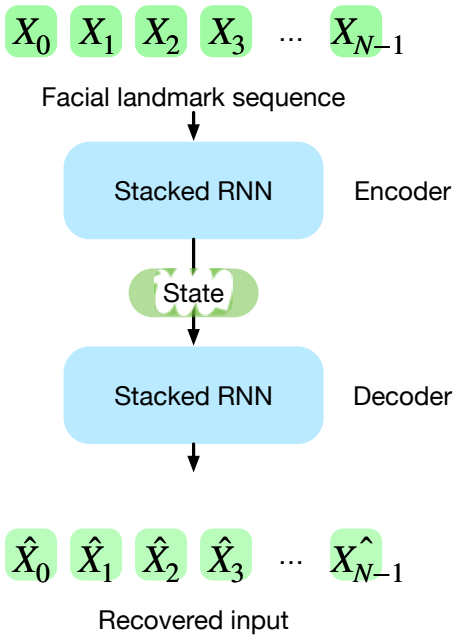


Figure 3: Auto-encoder model. The output sequence (recovered input) is not exactly the same as the input sequence. Then we use the difference as the loss to update the model parameters.

the high-dimensional input data into a lower-dimensional vector, which preserves the input's identity information. Then the decoder can recover the original data from this vector. Some research works have applied auto-encoders to sequential data, and the results are promising. We use auto-encoder on our face landmark sequences to obtain the facial action representation.

Figure 3 shows our auto-encoder model. This model includes two parts, encoder and decoder. Both of them use stacked RNN as the backbone. The hidden state of RNNs in the encoder works as the intermediate state (green block). The RNNs in the decoder use it to initialize their hidden layers. The output of the decoder has the same shape as the input sequence. The difference between the input and output determines the loss.

The encoder and decoder in our design both use a two-layer RNN as the backbone. The intermediate vector becomes the representation of facial action.

Metric Learning

The other model we developed is shown in Figure 4. It is an RNN based metric learning model that can generate an embedding for each facial landmark sequence. This model shares the same structure as the encoder in our auto-encoder model. The difference is that it employs a contrastive loss as opposed to learning the embedding in an unsupervised manner. We train this model with positive and negative pairs. Each positive pair for training includes two sequences of the same facial action from the same person. Each negative pair has two sequences from either the same person but different facial actions, different people but the same facial action, or both are different. This training process's goal using contrastive loss is to obtain a higher score for positive pairs and a low score for negative pairs, which is how the system

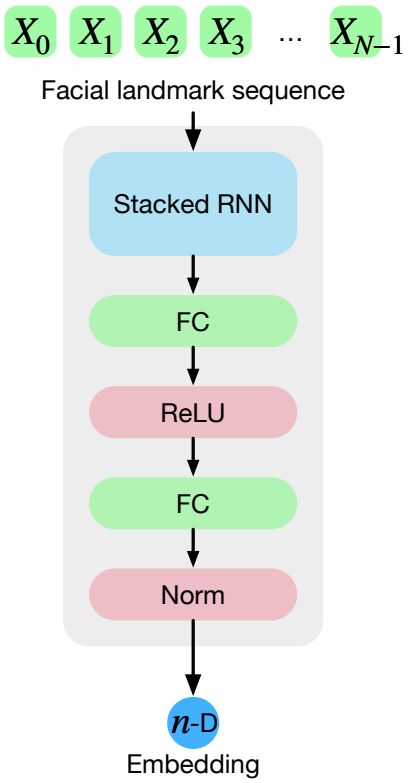
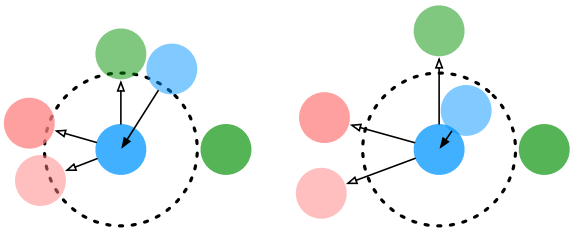


Figure 4: Metric learning model. We apply the L_2 normalization to the output to normalize the length of the n-D representation vector.

is supposed to report when comparing the live video input to the recorded reference video for identity verification.

This model employs contrastive loss function [18], which is a common criterion designed for metric learning. It is similar to the binary cross-entropy (BCE) for binary classification tasks. Different from BCE, the contrastive loss is determined by the input positive and negative sample pairs. When a pair of two samples from the same category (positive), their representation vectors should be similar. For samples from a pair of two different categories (negative), the representation vectors are expected to be dissimilar. The similarity is usually measured using Euclidean distance or cosine similarity.

As shown in Figure 5a, the light blue, and dark blue circles represent the same facial action, and circles of other colors represent different facial actions. Their distances to the reference facial



(a) Initial (b) Final

Figure 5: Contrastive loss.

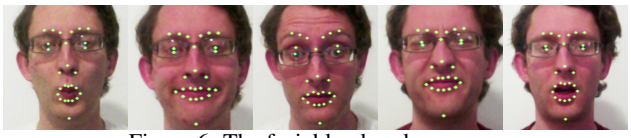


Figure 6: The facial landmarks we use.

action (dark blue) do not represent their similarity to the reference facial action in the initial training stage. After iterations of training, the same facial action (light blue) will be drawn toward the reference. All other facial actions will be pushed away from the reference, and their distances will increase. In other words, this training process using the contrastive loss will provide better separation if facial actions are different. If a facial action (dark green) distance is already longer than a given threshold (dashed circle), this training process will ignore it and not update its representation vector. The formula of the contrastive loss function we use is,

$$\mathscr{L}_{W,b} = (1-y) \times {D_{W,b}}^2 + y \times max(0,\ m - D_{W,b})^2 \qquad (1)$$

y is the pair label. $D_{W,b}$ is the distance between two embedding vectors generated by the model with parameter set (W,b). m is the distance margin that denotes the minimum acceptable distance of vectors from the negative pair.

Experiments

We used the implementation of the face detector algorithm and facial landmarks detection algorithm developed by dlib[19]. This facial landmark detector generate2 72 landmarks for each face image. However, the proposed two models are difficult to converge because of the many redundant and noisy points. We removed some noisy landmarks such as the points on the chin. We also combine redundant landmarks, such as the points around the eyes. The landmarks of active facial muscles were not removed as shown in Figure 6. The proposed models can converge easier with these important landmark points and obtained better accuracy.

As the proposed method is data-driven, we examined some benchmark facial expression video datasets. There were not enough positive samples that one subject makes the same facial action multiple times. Using these data sets would not provide a sufficient number of positive sample pairs for training and testing. Researchers created these datasets for modeling facial expressions for recognition, which is a subset of general facial actions. They are not suitabl for our identity verification algorithm. To verify the feasibility of the proposed method, we created a dataset specifically for our experiments. Our dataset contains ten subjects, and each subject repeats ten predefined facial actions 40 to 50 times. We collected a total of 4351 video clips. Figure 7 shows some samples of our dataset. We down-sampled them to 5 frames per second (FPS) for viewing, while the original framerate of all samples in the dataset is 15 FPS.

We implemented our two neural network models using PyTorch[20].

Results and Discussion

It is harder to form positive pairs than negative pairs because the subject had to repeat the same facial action multiple times to generate positive pairs. The numbers of binary-class pairs are very imbalanced. Therefore, we used Precision-Recall (PR) to

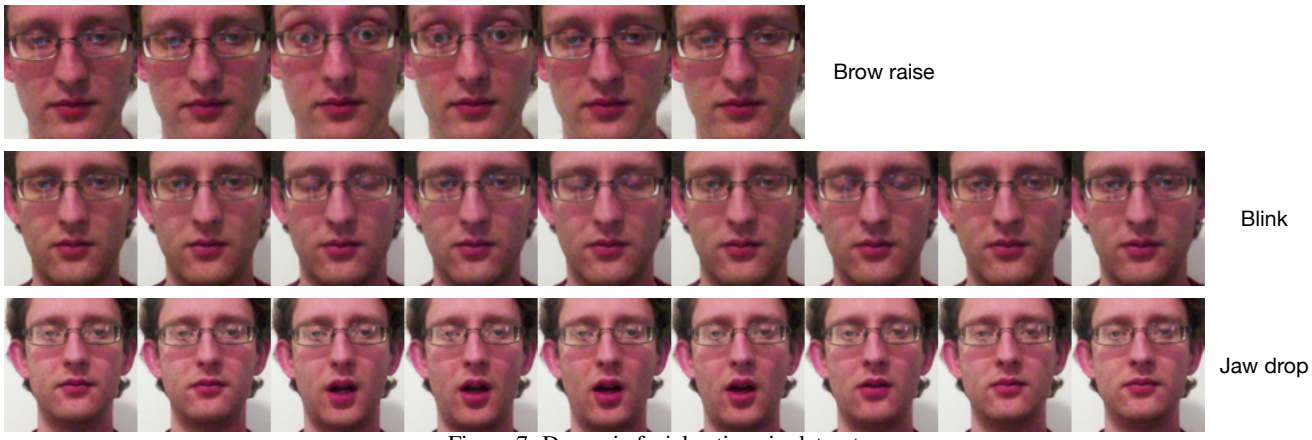


Figure 7: Dynamic facial actions in dataset.

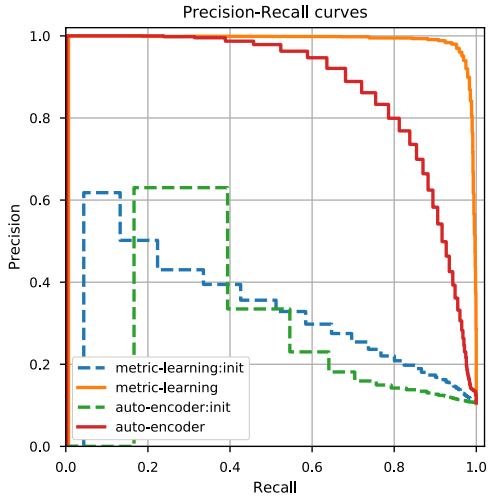


Figure 8: PR curves. The two solid curves (red and orange) indicate the final performances of two proposed models, while the dashed curves show their initial power before training. The metric learning model is up-and-coming.

evaluate the performance of our two trained models. Figure 8 shows four PR curves. The two solid curves denote two models' final performance, while the dashed curves show where the models start. Both models were able to learn the representation of facial actions. The metric learning model had the best PR curve, whose average precision (AP) was 0.988. For facial authentication, the number of false-positive cases should be minimal. Precision (P) is more critical than Recall (R) for systems with a high-security level.

Our C2FIV performs identity verification by analyzing facial features and facial actions concurrently. It is considered a two-factor verification because both facial features and facial actions are used for verification to increase the security. Only the same action from the same person can pass the verification. The same facial action but from different people will fail the verification because facial features are embedded in the features extracted by our network. The wrong facial action whether from the same person or not will surely fail.

We noticed that, in the final curve of the metric learning model (orange), when P started dropping slightly below 100%, R was still close to 88%. In other words, the proposed method verifies identity with 100% accuracy, and no false facial actions can pass. The only minor issue is that the user may have to repeat the verification process approximately every one out of ten tries. These numbers are more than acceptable for identity verification systems that require a very high level of security.

Of course, for different applications, this performance can be adjusted by using a different threshold. For example, for getting access to a restricted area, the precision should be 100% even though the user may sometimes have to try the authentication process more than once. For low-level security entrance control or tracking workers' presence or work hours, precision could be lower, but the user only needs to authenticate once (high recall).

Conclusions

We proposed a novel two-factor identity verification system. It can verify the user's identity using both facial features and facial actions. We also implemented two prototypes. One is based on auto-encoder and the other one uses deep metric learning. They both use RNN as the neural network backbone. We created a small dataset that contains some common facial actions. The experiment result shows that both architectures can learn the representation of facial features and facial actions concurrently. The deep metric learning with contrastive loss provides better accuracy.

Future Works

The algorithm works well with the frontal view, the head pose variation cannot be ignored. Usually, the user's head pose is near perfect during the registration process. The identity verification process may have to deal with head pose variations. A robust identity verification system should handle a slight deviation from the frontal view and warn the user if their head pose is not acceptable for verification.

Currently, only facial landmarks are used to model facial actions. Bypassing landmarks detection and using the raw video shoudl improve its performance at least for analyziing facial features. CNN has some variants[21] that can learn both spatial and temporal information from video. In the future, we will explore the feasibility of using facial action video as the input to generate its embedding in one stage without landmark points detection.

Facial action video includes facial features and facial action information. People could learn and mimic the facial action to pass the verification. Our next goal is to extract more detail facial features from the input video to improve the verification performance. Multiple-level metric learning could be an ideal method for this approach.

References

[1] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE transactions on information forensics and security*, 11(10):2268–2283, 2016.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[3] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12, 2015.

[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*, pages 815–823, 2015.

[5] 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*, pages 435–442, 2015.

[6] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[7] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[8] Pengqing Xie. Facial movement based human user authentication. *Graduate Theses and Dissertations*, 14267, 2014.

[9] Felix Chow and Arvin Wai Kai Tang. Method and system of identification and authentication using facial expression, April 11 2017. US Patent 9,619,723.

[10] Delina Beh Mei Yin, Amalia-Amelia Mukhlas, Rita Zaharah Wan Chik, Abu Talib Othman, and Shariman Omar. A proposed approach for biometric-based authentication using of face and facial expression recognition. In *2018 IEEE 3rd International Conference on Communication and Information Systems (ICCIS)*, pages 28–33. IEEE, 2018.

[11] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[12] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[13] Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24, 2015.

[14] Huaizu Jiang and Erik Learned-Miller. Face detection with the faster r-cnn. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 650–657. IEEE, 2017.

[15] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[16] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013.

[17] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[18] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[19] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[21] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

Author Biography

Zheng Sun received his Bachelor of Engineering degree from Sun Yat-sen University in 2017. He is pursuing a Ph.D. degree in Electrical & Computer Engineering at Brigham Young University. His work has focused on computer vision and machine learning.

Dr. D. J. Lee received his Ph.D. degree in electrical engineering from Texas Tech University in 1990 and MBA degree from Shenandoah University in 1999. He served in the machine vision industry for eleven years before joining Brigham Young University faculty in 2001. He is currently a professor and the director of the Robotic Vision Laboratory in the Electrical and Computer Engineering Department at BYU. He co-founded Smart Vision Works, Inc. in 2012. His research includes vision systems and devices with artificial intelligence, high-performance visual computing, real-time robotic vision, and visual inspection automation applications.

Dr. Dong Zhang received his B.S.E.E. and M. S. degrees from Nanjing University, China, in 1999 and 2003, respectively, and Ph.D. degree from Sun Yat-sen University, China, in 2009. He is currently an associate professor in the school of Electronics and Information Technology, Sun Yat-sen University. His research interests include image processing, pattern recognition and information hiding.

Xiao Li received his B.S. degree and M.S. degree from Sun Yat-sen University, China, in 2018 and 2020, respectively. He is currently a research assistant at the School of Electronics and Information Engineering, Sun Yat-sen University. His research interests include head pose estimation, eye-tracking, and computer vision.

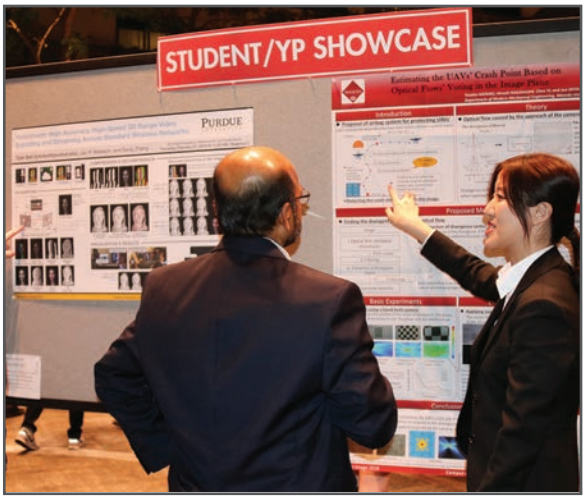
JOIN US AT THE NEXT EI!

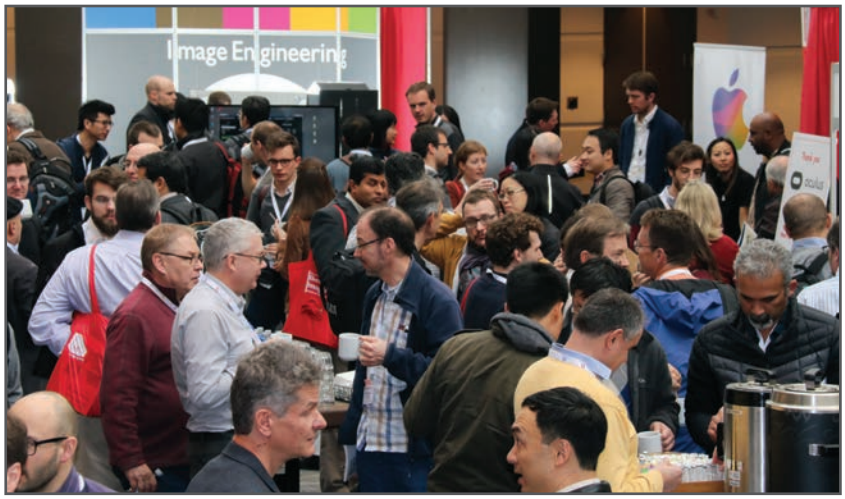
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org