https://doi.org/10.2352/ISSN.2470-1173.2019.8.IMAWM-401

Dense prediction for micro-expression spotting based on deep sequence model

Thuong-Khanh Tran, Quang-Nhat Vo, Xiaopeng Hong, Guoying Zhao*;
Center for Machine Vision and Signal Analysis, University of Oulu; Oulu, Finland.
**** Corresponding author***

Abstract

Micro-expression (ME) analysis has been becoming an attractive topic recently. Nevertheless, the studies of ME mostly focus on the recognition task while spotting task is rarely touched. While micro-expression recognition methods have obtained the promising results by applying deep learning techniques, the performance of the ME spotting task still needs to be largely improved. Most of the approaches still rely upon traditional techniques such as distance measurement between handcrafted features of frames which are not robust enough in detecting ME locations correctly. In this paper, we propose a novel method for ME spotting based on a deep sequence model. Our framework consists of two main steps: 1) From each position of video, we extract a spatial-temporal feature that can discriminate MEs among extrinsic movements. 2) We propose to use a LSTM network that can utilize both local and global correlation of the extracted feature to predict the score of the ME apex frame. The experiments on two publicly databases of ME spotting demonstrate the effectiveness of our proposed method.

Introduction

Micro-expressions are brief and involuntary facial emotions which are occurred when people are trying to conceal their feelings [1]. The research from Ekman [2] shows that MEs play an important role in psychology which helps people understanding spontaneous emotions. Analyzing the suppressed and concealed emotions can help us building potential applications in diverse areas, e.g., lie detection system for law enforcement, abnormal emotions analysis of psychotherapy, etc. Therefore, MEs have attracted lots of attention from various fields such as computer vision and psychology.

In the field of computer vision, ME analysis has been divided into two major tasks: spotting and recognition. The first one is locating the positions of MEs in videos and the second one is classifying the type of emotions. Although there are many studies involved in ME analysis by using computer vision, building a practical system for analyzing ME is still far away due to several issues. One reason is the lack of research in ME spotting. Indeed, most of the ME-related studies focus on the recognition task, while spotting is seldom touched [3, 4, 5, 6]. In a practical system of ME analysis, the positions of MEs should be located precisely in the sequence before further emotion interpretation or recognition.

ME spotting studies are facing several problems caused by the subtle movements such as the small changes on the face due to illumination effects, slight head movement, etc. These changes are similar to MEs and can lead to false-alarms with existing methods. For overcoming these issues, robust tools from machine learning are promising to discriminate MEs from extrinsic movements on the face. Among machine learning methods, deep learning is emerging as a powerful framework which outperformed traditional handcrafted features in many challenging topics of computer vision. While many deep learning-based MEs recognition methods have appeared in the literature [6, 7, 8, 9], the application of deep models in ME spotting is rarely considered.

To fulfill this gap, we propose a novel approach for locating ME in long videos based on the combination of a spatial-temporal descriptor and a deep sequence model. The spatial-temporal feature can discriminate MEs among extrinsic movements. In addition, Long-short term memory (LSTM) network is capable of recognizing and synthesizing both local and global temporal correlation of extracted features [9]. The proposed method is different with previous works which only consider the local correlation. Furthermore, LSTM showed that it can be adapted to a variable length of the video sequence, i.e. both very long and short sequence. To our best knowledge, this is the first time, deep sequence model approaches the ME spotting problems.

The rest of this paper is organized by following sections. In the next section, we briefly summarize the ME spotting methods in previous studies. The third section describes the proposed method. Then, our experiments and results are reported in the fourth section. Finally, we conclude our work in the last section.

Related Work

In the development of ME studies, several researchers realized the importance of detecting ME location when processing with long videos. Certainly, a real ME analysis system needs to locate ME positions exactly in long videos first, before any recognition steps can be applied. Therefore, ME spotting has been recently becoming an attractive topic in ME analysis. There are previous studies of ME spotting which are categorized into two groups: measurement-based methods and classification-based methods [6].

Firstly, we briefly go through several studies of the first group, measurement-based methods. In [10], Moilanen et al. spot MEs by using Chi-Square distance of Local Binary Pattern (LBP) in fixed-length scanning windows . This method is utilized to provide the baseline results in the first system which combines spotting and recognition [1]. Patel et al. proposed calculating optical flow vector for small local spatial regions, then using heuristics algorithm to remove the non MEs [11]. Wang et al. suggested a method named Main Directional Maximal Differences which utilizes the magnitude of maximal difference in the main direction of optical flow [14]. Generally, almost all these methods focus on

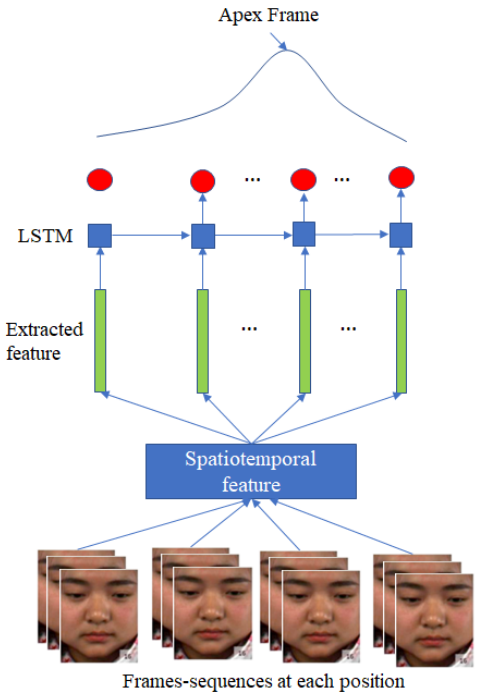


Figure 1. *Overview of ME spotting based on LSTM. First, we extract spatial-temporal feature on each positions of video, then sequence of feature vectors is inputted to LSTM to predict the score of apex frame position in video sequence.*

finding differences between non-micro and micro frames. They often calculate a threshold to eliminate false alarms caused by: for example, head movement or illumination effect. However, these methods are still facing with false alarms caused by, e.g., illumination effects, eye blinking, etc.

In order to discriminate MEs which just include small changes on face, several methods belonging to the second category were introduced. In [12], the first attempt utilizing machine learning in ME spotting was introduced. In this research, author employs Adaboost to predict whether the probability of a duration of frames belonging to a ME or not. Then, random walk functions were used to refine and integrate the output from Adaboost to return the final result. Recently, for providing the benchmark to standardize the evaluation in ME spotting studies, Tran et al. [13] proposed using multi-scale sliding-window based method for detecting MEs. This method tackles ME spotting as a binary classification problem based on a window sliding across positions and scales of video sequence. Although studies from [12, 13] take advantages from machine learning, the performances are still not good enough. This is because the traditional methods are not robust enough to handle the subtle movements of MEs.

Recently, deep learning techniques have been very popular in the research community. Many research showed that deep learning outperformed handcrafted features or traditional methods in many computer vision tasks, such as face recognition and pedestrian detection. ME recognition has also obtained promising results with deep learning methods [15], [8], [9], [7] . Therefore, it is reasonable to employ deep learning techniques in ME spotting.

Proposed Method

The overview of our method is illustrated on Fig. 1. There are two main steps in this framework: the first one is the extraction of the feature on each position of a video sequence, the second one is construction of deep learning network to predict the score of apex frames in short-clip of video sequence. Additionally, our method also contains the pre-processing step to carry out face-alignment and the post-processing step to process the final output. The following sub-sections will describe each step in our framework.

Pre-propocessing

At the beginning of our system, we perform the Pre-processing step which carries out face-alignment for image sequences. This step is necessary to minimize the variance of face sizes, differences of face shapes across video samples. The process is as follow: (1) face area is located in each frame by using Viola-Jones and KTL tracking method [17, 18], (2) face-alignment is conducted by utilizing Dlib toolbox [23] to locate 5 land-marks point on each frame of the video sequence. Then, we apply face-registration by using Local Weighted Mean [20]. (3) face area is cropped by utilizing the defined rectangles from specific landmarks points.

Spatial-Temporal Feature Extraction

After face alignment, we extract facial features on each position of the video sequence. There are many choices: single-frame featurs, spatial-temporal feature and deep feature. The use of spatial-temporal feature can help us obtaining both spatial and temporal information in each temporal position. Because micro-expressions are fast changes on the face, they often occur on short consecutive frames. Therefore, processing consecutive frames to calculate the score of apex frame of MEs is more reasonable than processing single frames. Although deep features have been investigated in many research, most of pre-train networks are only applied for single images. By that reason we do not explore deep features in this research.

With the aim of using spatial-temporal feature, three descriptor from the work of Li et al. are selected [1]:

- Local Binary Pattern for Three Orthogonal Planes (LBP-TOP): this feature, which is extended from LBP for spatial-temporal description, was introduced by Zhao and Pietikinen [21]. It's widely utilized in facial-expression analysis and ME recognition.
- Histogram of Oriented Gradient for Three Orthogonal Planes (HOG-TOP). This feature is extended from HOG to 3D to calculate oriented gradients on three orthogonal planes for modeling the dynamic texture in video sequence.
- Histogram of Image Gradient Orientation for Three Orthogonal Planes (HIGO-TOP). Histogram of Image Gradient Orientation (HIGO) is the degraded variant of HOG. It ignores the magnitude and counts the responses of histogram bins.

To construct the sequence of spatial-temporal feature vectors, we slide a scanning-window across all positions of a video. In each position, for example at frame i, we extract features in the sequence of frame i to $(i+L)$. On Figure 2, we illustrate the

feature extraction and the weighting socre of each feature. By using this strategy, we consider each position is a candidate of ME. In next step, the sequence of consecutive features is inputted to a deep sequence model to predict the score of MEs in video.

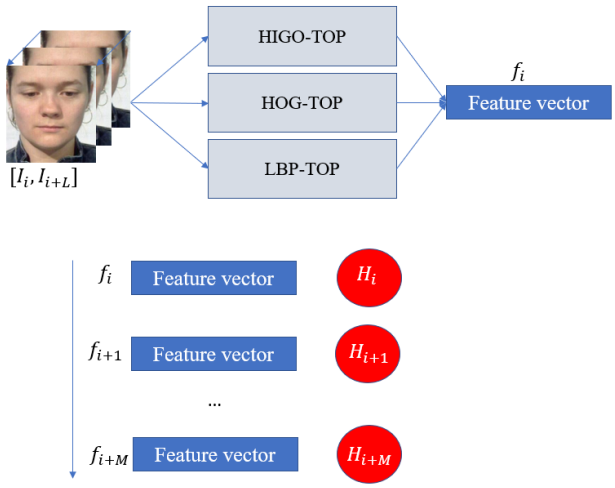


Figure 2. *Illustration of ME location prediction in one scanning-window. First, we extract spatial-temporal feature on each position i^{th} of video sequence. Each feature will take information of L-consecutive frames ($[I_j, I_t]$ means the frames from j to t). After extracting features, we slide a scanning window across video to create a sequence of features for the LSTM model. In one scanning-window, we predict the score of ME (H_i) on each position. The ground-truth (H_i) is calculated by Eq. 1 and Eq. 2.*

Long-short term memory for ME spotting

In this section, we describe the proposed LSTM network structure and our idea for predicting the score of ME.

Long short-term memory (LSTM) network is a special kind of Recurrent Neural Network (RNN). It was introduced by Hochreiter [22], and was refined and popularized in many research works, especially in sequence learning. The power of the LSTM network in learning and modelling sequence data is useful for estimating the position of ME in the video sequence.

Our idea for using LSTM in ME spotting is to slide a scanning-window across the video sequence. Each scanning-window contains M spatial-temporal feature positions. Our constructed network input M spatial-temporal features and predicts M values representing the score of MEs on each temporal position. The feature positions (i^*) that have the high score are considered as the ME samples. The specific apex frame is determined by the middle location $\frac{i^*+(i^*+L)}{2}$ of feature.

For constructing the ground-truth of the input feature sequence, which is the score of ME apex position inside each scanning-window, we propose two methods. The first method is based on the number of frames in the scanning-window that located in the range of onset and offset. The second method is based on the normalized distance between apex frame and center frame of scanning-window.

In Eq. 1, it's the formula of the first method to label the score of ME. We calculate the overlap rate between ME frames and feature scanning-window.

$$
\begin{aligned}
OverlapScore_i &= \frac{1}{L}\sum_{j=i}^{i+L}\delta_j \\
\delta_j &= \begin{cases} 1, & \text{if } j \in [I^{onset}, I^{offset}] \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
\tag{1}
$$

where $OverlapScore_i$ is the overlap weight of ME in feature i^{th}, L is the length of consecutive frames using to extract one spatial-temporal feature. I^{onset} and I^{offset} are the indexes of onset and offset frame, respectively. We set $Overlap_i$ to 0 if the value of $Overlap_i$ is less than 0.5.

In Eq. 2, it presents the second method to provide the weight for each feature position. As Eq. 1, we consider one feature as ME sample, the center frame in L consecutive frames is determined as apex frame. By this idea, we propose the normalized distance between ground truth apex and center frame of one spatial-temporal feature.

$$
ApexScore_i = \begin{cases} 1 - \frac{|I^{apex} - I_i^{middle}|}{L}, & \text{if } I^{apex} \in [I_i, I_{i+L}] \\ 0, & \text{otherwise} \end{cases}
\tag{2}
$$

where $ApexScore_i$ is the normalized distance score of feature i^{th}. I^{apex} is the index of ground truth apex frame, I_i is the index of frames belonging to feature i^{th}, I_i^{middle} is the frame index of middle position in feature i^{th}. The value of $Apex_i$ is set to 0 if $Apex_i < 0.5$. .

The architecture of LSTM is presented on Figure 3. Keras framework is utilized to build this model. The network is constructed by one LSTM layer with K units (in our experiment, K is set to 20) and one dense layer. The output of our LSTM model is the score of ME position (apex frame) for each temporal position. In our experiment, we tried two LSTM layers, however the performance of two layers is worst than 1 LSTM layer. The poor performance may be caused by overfitting issue. Therefore, we only report the network with one LSTM layer and one dense layer.

After obtaining predicted score from LSTM model for each scanning-window, we have to handle the overlap issue. Since each position can contain multiple detected apex positions, we utilize non-maximal suppression to merge the multiple detected apex frames between overlap scanning-windows. For removing false-positive, several heuristics rule are used to determine the positions for ME apex. This threshold is defined as the average values of maximum value and non-zero minimum value in one video sequence.

Experiment

Experimental Setup

In this sub-section, we briefly introduce two publicly datasets which are utilized for conducting experiments. The details of implementation and setup are also reported.

The first dataset we use is SMIC-VIS-E. This dataset has 76 video sequences with the frame size 640×480 pixels recorded at 25fps. It consists of 71 micro-expression videos and five non-micro videos. People in the video are shown emotional clips and are instructed to hide their feelings. These videos are annotated

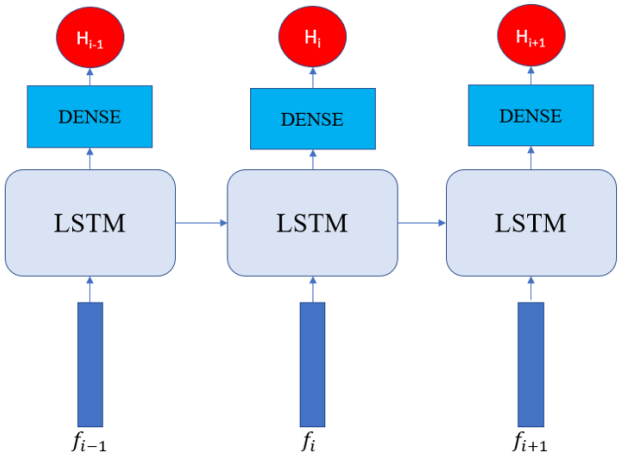


Figure 3. *The architecture of the deep sequence model in our method. It inputs a scanning-window containing M feature vectors and returns M scores of apex frame. It consists of two layers: LSTM (20 units) and Dense layer (20 input units computing one output). X_i is the input feature, H_i is the output, i from 1 to M.*

with onset and offset frames. In [13], the first benchmark of ME spotting was provided on this dataset. In implementation setup, we set $L = 9$ since it is the length of the most ME videos in SMIC-VIS-E, M is set from 30 to 40.

The second dataset is CASME2. There are two sections in CASME2: A and B. We only use section A which is built for ME spotting task. This section A consist of 7 subjects with 95 video sequences. The average of ME size is 23. The frame size of CASME2 is 640×480 pixels recorded at 30fps. In our implementation, L is set to 23 and M is set from 30 to 35.

To evaluate and compare our methods, we suggest using Leave-one-subject-out test setup, and calculating F1-score value to compare different methods. In one video sequence, we count the final detected apex and missed apex (counted as false negative). The detected location are then compared with ground truth. If spotted apex is located inside onset and offset ($SpottedApex \in [Onset, Offset]$), this is one true-positive, otherwise it is one false-positive. We calculate the average of F1-score values for the final results. On our techniques (the combinations between spatial-temporal features with two ground truth weighting methods), we trained with maximum of 150 epoch. Then we select the epoch having the best results to report on each techniques. To compare with other studies, we follow their results such as MDMD [6] or we re-implemented their method such as [10].

Results

Experimental results are reported in Tables 1 and 2. The first column describes the corresponding method. For example, $HIGO - TOP\ Overlap\ LSTM$ means the proposed method with spatial-temporal HIGO-TOP features and using Overlap method in labeling. $HIGO - TOP\ Apex\ LSTM$ is the proposed method with HIGO-TOP features and using Apex method in labeling. *LBP ChiSquare* is the method from [10], which was re-implemented by ourself. *MDMS* is the method Main Direction Maximal Difference analysis, which was reported on [6] and [14].

On table 1, we report the result when evaluating on SMIC-VIS-E dataset. In these results, our proposed methods achieved the promising results. If we use HOG-TOP combining with LSTM and Overlap weighting, we obtained the best result by F1-score 0.62. The second one is combining HIGO-TOP feature when using with LSTM. On these experiment, using ground truth "Overlap Weighting" is better than "Apex Distance".

Table 1. Experiment results on SMIC-VIS-E dataset.

Method	F1-Score
LBP ChiSquare [10]	0.29
HOG-TOP Apex LSTM	0.38
HIGO-TOP Apex LSTM	0.4
LBP-TOP Overlap LSTM	0.41
HIGO-TOP Overlap LSTM	0.5
HOG-TOP Overlap LSTM	**0.62**

On Table 2, we report the results when evaluating on CASME2 dataset. Following the results, our methods outperformed the previous methods of [10] and [6]. The method $HIGO - TOP\ Overlap\ LSTM$ obtained the best F1-score: 0.86, and the second is $HOG - TOP\ Overlap\ LSTM$ with F1-score 0.84.

Table 2. Experiment results on CASME2 dataset.

Method	F1-Score
LBP ChiSquare [10]	0.32
MDMD [6]	0.38
LBP-TOP Apex LSTM	0.69
HIGO-TOP Apex LSTM	0.71
HOG-TOP Apex LSTM	0.77
LBP-TOP Overlap LSTM	0.80
HOG-TOP Overlap LSTM	0.84
HIGO-TOP Overlap LSTM	**0.86**

Conclusion

In this paper, we introduced a novel approach for ME spotting based on the spatial-temporal features and the deep sequence model. To our best knowledge, this is the first work in the literature that explores the application of deep learning techniques for ME spotting. In the experimental results, we showed that the proposed method performs promisingly for ME spotting.

For future work, we will do experiment with more datasets of ME spotting. We have been building a benchmark for ME spotting to standardize the comparison between ME spotting methods and provide the baseline results. The first version of this work has been published on [13], [24]. Moreover, the current proposed framework only utilize three spatial-temporal features, there are still rooms for us to explore more spatial-temporal features, deep features in this framework [8, 16].

Acknowledgment

This work was supported by the Academy of Finland, Tekes Fidipro program (Grant No. 1849/31/2015), Business Finland project (Grant No. 3116/31/2017), and Infotech Oulu.

References

[1] LI, Xiaobai, et al. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Transactions on Affective Computing, 2017

[2] Ekman, P., OSullivan, M., Frank, M. G. (1999). A Few Can Catch a Liar. Psychological Science, 10(3), 263266

[3] WU, Qi; SHEN, Xunbing; FU, Xiaolan. The machine knows what you are hiding: an automatic micro-expression recognition system. In: Affective Computing and Intelligent Interaction. Springer, Berlin, Heidelberg, 2011. p. 152-162.

[4] PFISTER, Tomas, et al. Recognising spontaneous facial micro-expressions. In: Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011. p. 1449-1456.

[5] HUANG, Xiaohua, et al. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In: Proceedings of the IEEE international conference on computer vision workshops. 2015. p. 1-9.

[6] OH, Yee-Hui, et al. A Survey of Automatic Facial Micro-expression Analysis: Databases, Methods and Challenges. Front. Psychol., 10 July 2018

[7] LI, Yante; HUANG, Xiaohua; ZHAO, Guoying. Can Micro-Expression be Recognized Based on Single Apex Frame?. In: 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018. p. 3094-3098.

[8] PATEL, Devangini; HONG, Xiaopeng; ZHAO, Guoying. Selective deep features for micro-expression recognition. In: Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016. p. 2258-2263.

[9] KHOR, Huai-Qian, et al. Enriched Long-term Recurrent Convolutional Network for Facial Micro-Expression Recognition. In: Automatic Face Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on. IEEE, 2018. p. 667-674.

[10] MOILANEN, Antti; ZHAO, Guoying; PIETIKINEN, Matti. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In: Pattern Recognition (ICPR), 2014 22nd International Conference on. IEEE, 2014. p. 1722-1727.

[11] PATEL, Devangini; ZHAO, Guoying; PIETIKINEN, Matti. Spatiotemporal integration of optical flow vectors for micro-expression detection. In: International conference on advanced concepts for intelligent vision systems. Springer, Cham, 2015. p. 369-380.

[12] XIA, Zhaoqiang, et al. Spontaneous micro-expression spotting via geometric deformation modeling. Computer Vision and Image Understanding, 2016, 147: 87-94.

[13] TRAN, Thuong-Khanh; HONG, Xiaopeng; ZHAO, Guoying. Sliding Window Based Micro-expression Spotting: A Benchmark. In: International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, Cham, 2017. p. 542-553.

[14] WANG, Su-Jing; WU, Shuhang; FU, Xiaolan. A main directional maximal difference analysis for spotting micro-expressions. In: Asian Conference on Computer Vision. Springer, Cham, 2016. p. 449-461.

[15] PENG, Min, et al. Dual temporal scale convolutional neural network for micro-expression recognition. Frontiers in psychology, 2017, 8: 1745

[16] HUANG, Xiaohua, et al. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing, 2016, 175: 564-578.

[17] VIOLA, Paul; JONES, Michael. Rapid object detection using a boosted cascade of simple features. CVPR 2001.

[18] TOMASI, Carlo; KANADE, Takeo. Detection and tracking of point features. 1991.

[19] COOTES, Timothy F., et al. Active shape models-their training and application. Computer vision and image understanding, 1995, 61.1: 38-59.

[20] GOSHTASBY, Ardeshir. Image registration by local approximation methods. Image and Vision Computing, 1988

[21] ZHAO, Guoying; PIETIKAINEN, Matti. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence, 2007, 29.6: 915-928.

[22] HOCHREITER, Sepp; SCHMIDHUBER, Jrgen. Long short-term memory. Neural computation, 1997, 9.8: 1735-1780.

[23] KING, Davis. Dlib c++ library. Access on: http://dlib. net, 2015

[24] HONG, Xiaopeng; TRAN Thuong-Khanh; ZHAO Guoying. Micro-Expression Spotting: A Benchmark. https://arxiv.org/abs/1710.02820

Author Biography

Thuong-Khanh Tran received his B.S in Mathematics and Computer Science from University of Science -VNUCHM, in 2010; and his M.S degree in Electronics Engineering from Chonnam National University, Republic of Korea, in 2015. Since 2016, he has been doing Ph.D. in the Center for Machine Vision and Signal Analysis, University of Oulu, Finland, under supervision of Professor Guoying Zhao. His study interests are emotion analysis, computer vision, pattern recognition.

Quang-Nhat Vo received his B.S. degree in Information Technology from the University of Science-VNUHCM, Vietnam, in 2010, and his M.S. and Ph.D. degree in Electronics and Computer Engineering from Chonnam National University, Republic of Korea, in 2017. He is currently a Postdoc researcher at Center for Machine Vision and Signal Analysis, University of Oulu, Finland. His study interests are multimedia and image processing, facial expression analysis, and pattern recognition.

Xiaopeng Hong received his Ph.D. degree in computer application and technology from Harbin Institute of Technology, P. R. China, in 2010. He is a Docent with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland, where he has been a scientist researcher since 2011. Dr. Hong has published over 30 articles in mainstream journals and conferences such as IEEE T-PAMI, T-IP, CVPR and ACM UbiComp. His current research interests include multi-modal learning, affective computing, intelligent medical examination, and human-computer interaction, etc. His research has been reported by global media including MIT Technology Review and Daily Mail.

Guoying Zhao (IEEE Senior member) received the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 2005. She is currently a Professor with the Center for Machine Vision and Signal Analysis, University of Oulu. She has authored or co-authored more than 190 papers in journals and conferences. She was co-publicity chair for FG2018, and is associate editor for several journals. Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, body gesture analysis, and person identification.

Figure 4. *IS&T logo.*

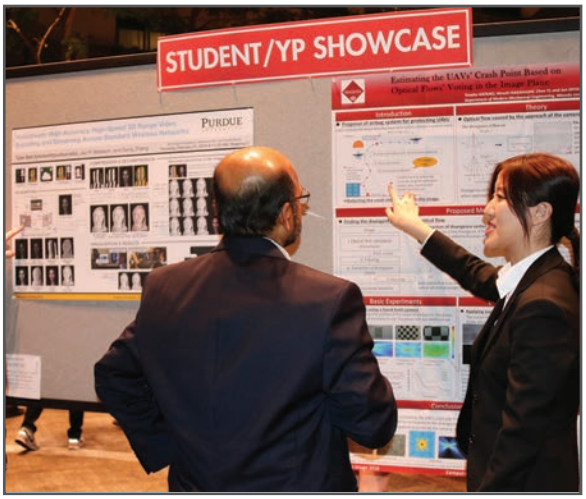
JOIN US AT THE NEXT EI!

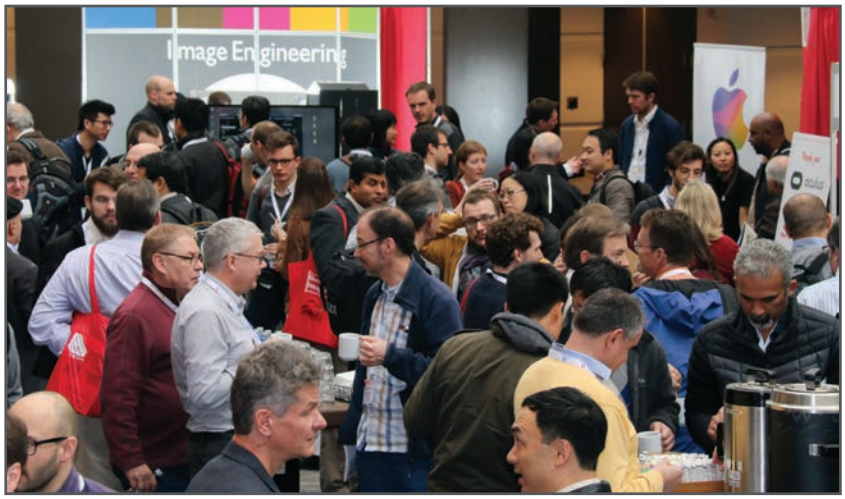
IS&T International Symposium on

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