Facial Micro-expression Analysis via a High Speed Structured Light Sensing System

Yuping Ye[1], Zhan Song[2,3], and Juan Zhao[2]
[1] University of Chinese Academy of Sciences, Beijing, China
[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[3] The Chinese University of Hong Kong, Hong Kong, China
Email: {yp.ye, zhan.song, juan.zhao}@siat.ac.cn

***Abstract*—Facial micro-expressions play a pivotal role in human non-verbal emotional expression, so it can be used in many fields such as criminal interrogation, clinical diagnosis and animation. Conventional means are usually based on 2D image analysis means, which has shown its disadvantage in real applications. In this paper, a 3D-based facial micro-expression analysis method is firstly proposed. The 3D acquisition equipment based on high-speed structured light is established to capture dense and accurate 3D facial shapes with a maximum speed of 300Hz. Facial micro-expression detection method is put forward to extract the onset and offset of micro-expression sequence. Finally, a 4D descriptor is introduced to describe the timing characteristic of facial micro-expression. Experiments on real human faces are used to verify feasibility of the proposed system and method.**

***Index Terms*—facial micro-expression, 3D facial expression, structured light, 3D reconstruction**

I. Introduction

Facial micro-expression refers to the subtle and fast facial motions which exposes a person's real feelings and emotions in psychology. Facial expression analysis has been a classic research topic in computer vision domain, and most existing methods are implemented based on 2D image analysis means. With the recent development of deep learning technology, many facial micro-expression methods based on deep learning have been proposed [1], [2]. Recently, 3D reconstruction methods like multiple view reconstruction and Time of Flight (ToF) techniques also have been introduced for the facial analysis applications [3]-[5].

In comparison with traditional facial expression analysis technique, facial micro-expression analysis has attracted fewer attentions. In [6], E. A. Haggard and K. S. Isaacs first discovered the existence of micro-expression, and then the facial micro-expression was treated as a vital psychological phenomenon and studied in social psychology [7], [8] domain. Facial micro-expression has numerous applications in many fields. In criminal interrogation, police can utilize micro-expression to make judgement about criminal behavior. With the aids of micro-expression, doctors can understand the autistic patients' underlying emotions. The character without the facial micro-expression in animation and video game looks unrealistic. The major challenges in facial micro-expression analysis involve their very short duration (1/3-1/25 second [8]) and low intensity. Researchers usually treat micro-expression as a dynamic process instead of a static, while limited number of frames can be achieved via the ordinary camera. A micro-expression only appears within 1-3 frames in a video sequence. Even these frames can be captured by the camera, since the micro-expression only appears as a very tiny change of face image, that's very difficult to identify them from 2D images directly. That's also the motivation for us to study this topic in 3D space.

To our best knowledge, this paper first applied the High Speed structured Light Sensing (HSLS) method for the analysis of facial micro-expression identification problem. Some 3D techniques have been adopted for classical facial expression analysis [9]-[12]. Huang, H., *et al*. [10] reconstructed a high-fidelity 3D facial information by combining motion capture data with the minimal set of face scans in a blendshape interpolation framework. Sun, Y. and L. Yin [11] proposed a spatial-temporal expression analysis approach based on 3D dynamic geometric facial model sequences. Generally, existing 3D facial expression analysis methods can be classified into three categories according to input raw sources, i.e., image from high-speed camera [12]-[14], 3D static model [5], [10], 3D dynamic model with low precision or slow imaging speed [11], [15], [16].

Research works related to micro-expression all utilize the high-speed camera to capture the image sequences. In [14], the paper introduced a temporal interpolation model together with a spontaneous facial micro-expression corpus to accurately recognize very short expression. Wen-Jing Yan *et al.* [12] created a corpus of micro-expression with temporal resolution (200 fps) and spatial resolution (about 280×340 pixels on facial area). S. Polikovsky [13] proposed a 3D gradient histogram descriptor to detect and measure the micro-expression in image frames. In [17], a dynamic texture-based approach was introduced for the recognition of facial action units and their temporal models. In [18], the authors extended the static spatial descriptor Local Phase Quantization (LPQ) to a dynamic one named Three Orthogonal Planes (LPQ-TOP). Compared with the descriptor in image sequences, the descriptors of 3D model are nearly all

Manuscript received June 2, 2020; revised December 21, 2020.

doi: 10.18178/joig.9.1.15-19

static. In [19], performance of some classical descriptors are evaluated on some baseline 3D models. Recently, deep learning as a popular tool have been introduced into 3D static descriptor as show in [20]. Descriptors based on deep learning may lose details, and thus not suitable to depict the facial micro-expression.

The organization of this paper is as follow. High speed 3D face imaging system is introduced in Section II. And the proposed facial micro-expression detection method is illustrated in Section III. Experimental results are provided and discussed in Section IV. Conclusion and potential future works are offered in Section V.

II. High Speed 3D Face Imaging via Structured Light Means

Subject to the low resolution and precision of commercial 3D sensors like Kinect and RealSense, these devices cannot be used for accurate 3D face model acquisition. Although high accuracy can be obtained by 3D scanner like laser or structured light-based devices, but their 3D imaging speed is low, and thus not suitable for dynamic 3D acquisition. To realize high accuracy and real-time 3D imaging of dynamic face, a binary encoded structured light means was applied in our work as described in this section.

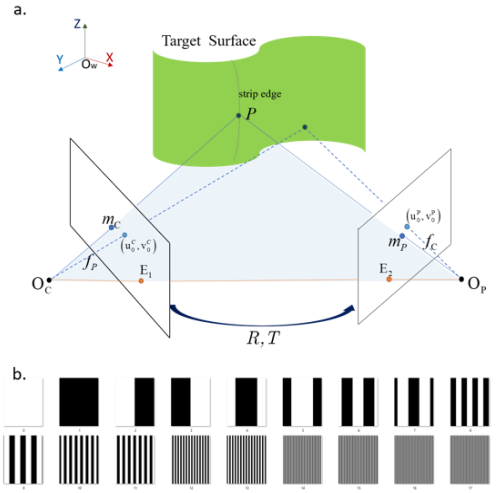

Figure 1. a) Geometric model of the structured light system which consists of a high-speed camera and a DLP projector; b) the encoded binary strip patterns.

Our high-speed acquisition equipment is built on our previous work [21]. We employed a unique codeword which is a combination of a global Gray code value and a local strip-edge code value for assigning the strip-edge point as shown by Fig. 1. Based on the perspective projection model, the structured light system can be described as:

$$s\begin{bmatrix} m^{C/P} \\ 1 \end{bmatrix} = \begin{bmatrix} f_u^{C/P} & \gamma^{C/P} & u_0^{C/P} \\ 0 & f_v^{C/P} & v_0^{C/P} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} I_3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} M^{C/P} \\ 1 \end{bmatrix}$$

$$\Leftrightarrow s\begin{bmatrix} m^{C/P} \\ 1 \end{bmatrix} = K_{C/P}\begin{bmatrix} R^{C/P} & T^{C/P} \end{bmatrix} \begin{bmatrix} M^{C/P} \\ 1 \end{bmatrix} \quad (1)$$

where the superscript C/P denote whether it is projector's or camera's parameters, m and M indicate the 2D captured image coordinate and 3D coordinate with respect to the world reference frame. K indicates the camera's or projector's intrinsic parameters matric. The 3D position in camera and projector coordinate systems can be expressed as:

$$\begin{bmatrix} M^P \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathrm{M}^C \\ 1 \end{bmatrix} \quad (2)$$

As Fig. 1 shows, $O_C O_P$ is epipolar line, E_1 and E_2 are epipolar point, using the epipolar constraint:

$$m_C^T \left(K_P^{-T} T K_C^{-1}\right) m_P = 0 \quad (3)$$

We can calculate the depth value by

$$Z_C = \frac{\left(Rm_C m_P\right)\left(m_C T\right) - \left\|m_C\right\|^2 \left(Rm_C T\right)}{\left\|Rm_C\right\|^2 \left\|m_P\right\|^2 - \left(Rm_C m_P\right)^2} \quad (4)$$

While the structured light system was established, a calibration method as described in our previous work [22] was used for system calibration and parameter optimization.

III. Facial Micro-expression Detection

As mentioned above, facial micro-expressions always show up along with short duration and low intensity. Existing facial micro-expression detection methods are usually implemented with high-speed camera. For the lack of depth information, image-based methods are difficult to detect facial micro-expression frames from the sequence. However, based on the established high-speed 3D imaging setup, 3D face model can be captured with a frame rate up to 300 Hz. That makes the detection of facial micro-expression in 3D space possible.

A. 3D Facial Micro-expression Regions Detection

Facial micro-expression detection algorithm aims at extracting the continuous frames where the expression occurred. In our algorithm, we first registered adjacent 3D frames, then calculated their difference in 3D space. Calculating of whole 3D facial difference is very time-consumed and meaningless. There are two common tactics for facial region selection, one is to divide the frame simply by facial landmarks [11], the other is based on a Facial Action Coding System (FACS) [23]. FACS decomposes facial expressions into 46 components named Action Unit (AU) based on face muscle structure and movement. We divided the facial region into 12 parts according to [13]. Because the structure light system has been calibrated, 3D coordinates of image features can be easily retrieved from the textured 3D face model. As Fig. 2 shows, the 3D face model can be divided into 12 parts according to the 2D selection. And the facial micro-expression algorithm can be described as follows.

Algorithm: Assume Ω be the 3D face sequence with texture, $\Omega = \{(I_1, M_1), (I_2, M_2) \cdots (I_n, M_n)\}$, where $M_i \supset \{R_j^1, R_j^2, \cdots R_j^{12}\}$. By detecting the model regions from texture images, we use $\mathrm{Re}\, gErr(R_i^k, R_j^k)$ to represent the distance disparity of the *k*-th region after registration of M_i and M_j via the algorithm in [24]. In general, facial micro-expressions are usually starts from neutral state and back to neutral state. According to [25], we can define a function $DetFE(I_j) \in$ {*neutral*, *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*} to represent the facial expression based on the texture image I_i. In our algorithm, $\Theta = \{1,2,3,4,7,8,9,10,12\}$ represent the selected region index. In details, the algorithm can perform as following steps.

- Extract the potential model sequences Λ which may include the facial expression. $\Lambda = \{(I_i, M_i), (I_{i+1}, M_{i+1}) \cdots (I_j, M_j)\}$ is the potential model sequences while $DetFE(I_k) = Neutral$, $k \in \{i, i+1, \cdots j\}$;
- Remove static and flashy sequences. If all the adjacent 3D face frames subject to $\sum_{k\in\Theta} \mathrm{Re}\, gErr(R_i^k, R_{i+\Delta}^k) < \varepsilon$, where $\Delta > 10$, it will be removed. If the sequence lasts less than 1/25 sec, it will be also removed;
- Detect the micro-expression start frame **s** and it end frame **e**. Calculate all the adjacent registration error $\sum_{k\in\Theta} \mathrm{Re}\, gErr(R_i^k, R_{i+\Delta}^k) \quad i \in \{1, \cdots n\}$. Finding **s** and **e** which satisfy following condition:

$$\begin{cases} \forall j : \sum_{k\in\Theta} \mathrm{Re}\, gErr(R_i^k, R_{i+\Delta}^k) < \varepsilon, j \in \{a, \cdots s\} \\ \forall i : \sum_{k\in\Theta} \mathrm{Re}\, gErr(R_i^k, R_{i+\Delta}^k) > \varepsilon, i \in \{s, \cdots e\} \\ \forall j : \sum_{k\in\Theta} \mathrm{Re}\, gErr(R_i^k, R_{i+\Delta}^k) < \varepsilon, j \in \{e, \cdots z\} \\ a - s > N; e - s > M; z - \mathrm{e} > N \end{cases} \quad (5)$$

where *M* is often five times that of *N*, and *M* indicates the number of interval of 1/25s.

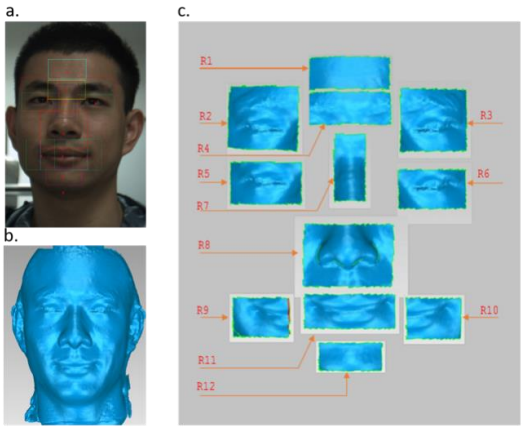

Figure 2. a) Selected facial regions on the 2D texture image; b) sample 3D face frame; c) facial regions on the 3D face model.

B. 4D Facial Micro-expression Descriptor

Temporal characteristic of facial micro-expression is necessary to be described for the behavior and psychological analysis. To make sure the temporal-spatial coherence across model sequences, we treated the time axis as the fourth dimension. Inspired by [13], a facial descriptor is proposed to measure the temporal feature of subtle and fast deformation of 3D face model.

With method in Section IIIA, facial micro-expression regions can be detected from the 3D facial sequence. The motion descriptor was put forward to depict the detected sequences. Firstly, we utilize the rigid transform from the adjacent model to register each 3D facial frame with the start frame. Secondly, we crop the two models into several regions based on the facial selections in Section IIIA. Generally, the deformation of facial micro-expression is very tiny. To depict the facial micro-expression, we divided the deformation amplitude into several range. Each facet's normal can be obtained by calculating the average of normal of triangle vertexes. As a result, we can figure out the volume which deformation amplitude lies in the specified range as shown by Fig. 3.

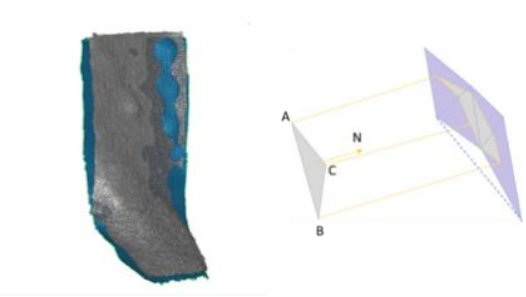

Figure 3. Left: Two meshes cropped by region 7. Right: Geometrical model to calculate the volume in different amplitude range.

To give a general 3D descriptor for the facial micro-expressions, a unique coordinate system is proposed in this work. The z-axis of the coordinate system points to the normal of Frankfort horizontal plane, the y-axis points from the left canthus to right, the origin is the barycenter of the triangle composed with left and right canthus and chin corner. We placed a sphere on the coordinate system origin, and label each pieces of the sphere by longitude and latitude. Then we can label each triangle normal into a specified sphere block. Compared with the start frame s, we can describe the 3D model change using several histograms, where each bin responds to the volume of normal within the sphere blocks. To give a visual description of amplitude deformation lies in specified range, the volume value can be indicated by the sphere block radius as shown by Fig. 4.

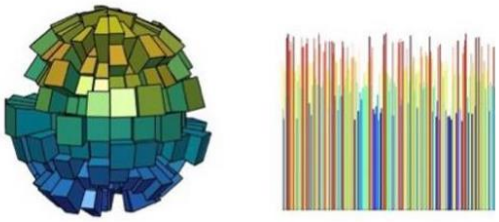

Figure 4. The visual descriptor and histogram of specified range based on the defined sphere coordinate system.

IV. Experimental Results and Discussion

The established structured light system is composed by one high-speed camera (Photron FASTCAM SA1.1, with the resolution of 1024*1024 pixels and a maximum frame rate of 5400 fps), and one high-speed DLP projector (TI-DLP6500, with the resolution of 1920*1080 pixels) as shown by Fig. 5. Since the adopted structured light patterns are all binary strips, the projection speed of the DLP device can be fully utilized (maximum frame rates up to 11574 Hz while output binary patterns). By synchronizing the projector and camera, a maximum scanning rate up to 300 Hz can be obtained by our 3D structured light system, and each 3D frame contains maximum 1M points with a depth error less than 0.05 mm.

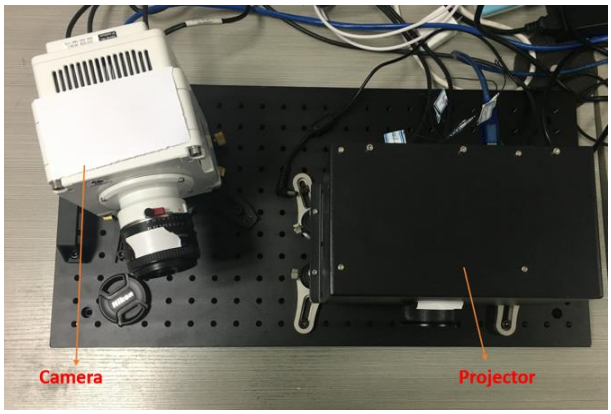

Figure 5. The constructed high-speed 3D scanning system can realize a maximum scanning speed of 300 Hz, while the image resolution is 1M pixels.

Scanning area of the system is about 500*500 mm, which is enough to cover the shoulder and facial areas of the subject. In the experiment, the actor was requested to perform several facial micro-expressions with the guidance. Each capture lasts about 4s, and more than 24,000 images can be captured. With the offline 3D reconstruction, more than 1,200 3D frames can be calculated. Fig. 6 shows the facial micro-expression detection results in face region 8 & 9.

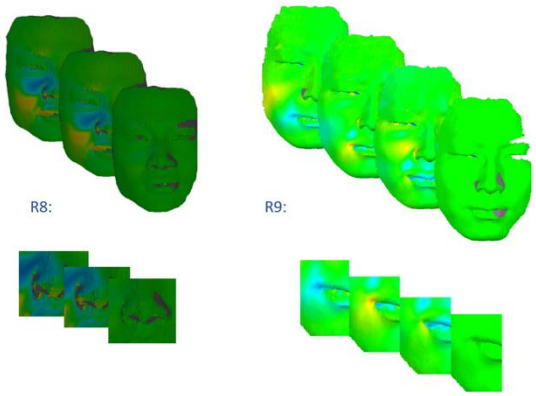

Figure 6. Detected facial micro-expression regions in different facial regions.

To verify feasibility of the proposed 4D descriptor, face region 9 in the Fig. 6 is used for the experiment. By the algorithm described in Section IIIB, the temporal feature of region 9 in Fig. 6 was calculated as shown by Fig. 7. In Fig. 7, the y-axis represents the range of amplitude of the surface along with the normal. The x-axis indicates the 3D frame index. Each histogram and sphere display the change volume and direction. Sudden change of the histogram indicates the change of facial model in specific face regions where potential facial micro-expression occurred.

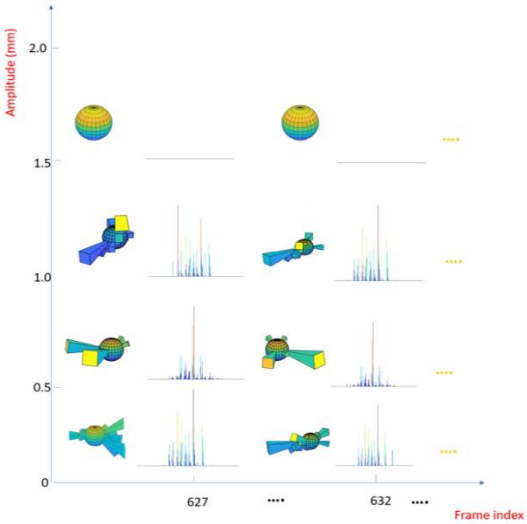

Figure 7. Visual samples of the 4D descriptor based on face region 9 in Fig. 6.

V. Conclusions and Future Work

This paper presents a 3D model-based facial micro-expression analysis approach. To obtain accurate dynamic 3D face models, a high-speed 3D imaging device was established based on structured light sensing principle. With the obtained 3D face models, the face region is first divided into some regions according to the FACS method in 2D image space. Then, a facial micro-expression detection method is introduced. Moreover, to describe the micro-expression regions more generally, a 4D descriptor is proposed based a spherical coordinating system. Experimental results on real human face showed that, the proposed 3D facial micro-expression detection method and 4D descriptor can work well in different face regions. Future work can address how to fuse the 2D color information and 3D information together to realize more accurate and robust micro-expression detection.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Zhan Song and Juan Zhao conducted the research. Yuping Ye wrote the paper and analyzed the data. All authors had approved the final version.

Acknowledgment

This work was supported in part by the Science and Technology Key Project of Guangdong Province, China

(2019B010149002), and Shenzhen Science Plan (KQJSCX20170731165108047).

REFERENCES

[1] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548-1568, Aug. 2016.

[2] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259-275, Jan. 2003.

[3] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "Multiple nose region matching for 3D face recognition under varying facial expression," *IEEE Trans Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1695-1700, Oct 2006.

[4] H. Soyel and H. Demirel, "Facial expression recognition using 3D facial feature distances," in *Proc. International Conference Image Analysis and Recognition*, 2007, pp. 831-838.

[5] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D facial expression recognition based on primitive surface feature distribution," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 1399-1406.

[6] E. A. Haggard and K. S. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Methods of Research in Psychotherapy*, Springer, 1966, pp. 154-165.

[7] C. M. Hurley, "Do you see what I see? Learning to detect micro expressions of emotion," *Motivation and Emotion*, vol. 36, no. 3, pp. 371-381, 2012.

[8] W. J. Yan, Q. Wu, J. Liang, Y. H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217-230, 2013.

[9] G. Sandbach, S. Zafeiriou, M. Pantic, and L. J. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image and Vision Computing*, vol. 30, no. 10, pp. 683-697, Oct. 2012.

[10] H. Huang, J. Chai, X. Tong, and H. T. Wu, "Leveraging motion capture and 3D scanning for high-fidelity facial performance acquisition," *ACM Transactions on Graphics*, vol. 30, no. 4, p. 74, 2011.

[11] Y. Sun and L. Yin, "Facial expression recognition based on 3D dynamic range model sequences," in *Proc. European Conference on Computer Vision*, 2008, pp. 58-71.

[12] W. J. Yan, *et al.*, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS One*, vol. 9, no. 1, p. e86041, 2014.

[13] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expression detection in hi-speed video based on Facial Action Coding System (FACS)," *Ieice Transactions on Information and Systems*, vol. E96d, no. 1, pp. 81-92, Jan. 2013.

[14] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. IEEE International Conference on Computer Vision*, 2011, pp. 1449-1456.

[15] G. Sandbach, S. Zafeiriou, M. Pantic, D. J. I. Rueckert, and V. Computing, "Recognition of 3D facial expression dynamics," *Image and Vision Computing*, vol. 30, no. 10, pp. 762-773, Oct. 2012.

[16] X. Zhang, *et al.*, "A high-resolution spontaneous 3d dynamic facial expression database," in *Proc. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1-6.

[17] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1940-1954, Nov. 2010.

[18] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 314-321.

[19] L. A. Alexandre, "3D descriptors for object and category recognition: A comparative evaluation," in *Proc. IEEE International Conf. on Intelligent Robotic Systems - IROS*, Vilamoura, Portugal, October 2012, pp. 1-6.

[20] Y. Fang, *et al.*, "3d deep shape descriptor," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2319-2328.

[21] Z. Song, R. Chung, and X. T. Zhang, "An accurate and robust strip-edge-based structured light means for shiny surface micromeasurement in 3-D," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 3, pp. 1023-1032, Mar. 2013.

[22] Y. Ye and Z. Song, "A practical means for the optimization of structured light system calibration parameters," in *Proc. IEEE International Conference on Image Processing*, 2016, pp. 1190-1194.

[23] P. Ekman and W. V. Friesen, "Facial action coding system: A technique for the measurement of facial action," *Palo Alto*, 1978.

[24] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Proc. Third International Conference on 3-D Digital Imaging and Modeling*, 2001, pp. 145-152.

[25] S. Alizadeh and A. Fazel, "Convolutional neural networks for facial expression recognition," *Computer Vision and Pattern Recognition*, arXiv:1704.06756, 2017.

Yuping Ye received his B.S. degree from Wuhan University, Wuhan, China, in 2013; and M.S. degree in computer science from University of Chinese Academy of Sciences, Beijing, China, in 2016. He is currently a Ph.D. student in mechanical engineering at University of Chinese Academy of Sciences, Beijing, China. His research interests include computer vision and computer graphics.

Zhan Song received his Ph.D. in mechanical and automation engineering from The Chinese University of Hong Kong, Hong Kong, in 2008. He is currently a Professor at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His research interests include structured light-based sensing and vision-based human-computer interaction.

Juan Zhao received the Ph.D. degree in mechanical from Heriot-Watt University, UK, in 2014. She is currently an associate Professor at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. Her research interests include structured light-based sensing and optical metrology.