H-012

Facial Emotion Recognition System by Using Depth Sensor

チャンタパン　ナッタワット† 　内村　圭一† 　里中　孝美‡ 　牧岡　毅‡
Nattawat Chanthaphan, Keiichi Uchimura, Tamaki Satonaka and Tsuyoshi Makioka

1. Introduction

In our previous works, [1, 2], we introduced a novel approach for extracting facial features from the moving facial skeleton (skeleton based approach) for facial emotions recognition by using depth camera.

Contemporary approaches (texture or 3d based) still have some limitations in dealing with some variations involving head pose, head orientation, distance from camera, light condition, skin tone and so forth. To overcome the limitations, we have proposed the novel approach [1, 2] by employing the Structured Streaming Skeleton (SSS) method, which is introduced in the human gesture recognition by Zhao et al. [3]. They indicated possible solutions to solve problems of the following intra-class variations.

(1)Viewpoint variation: This variation describes the relation between human body and viewpoint of the camera.

(2) Anthropometry variation: This variation is related to the difference between human body sizes which do not affect the human movement.

(3) Execution rate variation: This variation indicates the problem with different frame rate of the camera or the moving speed of human.

(4) Personal style variation: This variation is about the difference of human performing their action differently.

In our work [1], we was facing the problem for lack of dataset (five-people dataset). Therefore, we have added more dataset in [2], and it has shown a promising result. In this paper, we would like to add more emphasis on comparing the Structured Streaming Skeleton (SSS) feature, stream feature and state-of-the-art approach [4] by 15-people dataset.

2. Related works

As we have mentioned various variations in the introduction, image pre-processing is conducted prior to the feature extraction phase. Zhu et al. [5] showed that RGB camera-based approach consumed very expensive computation time. New sensory device was released in 2010 named Microsoft Kinect V2. We could reduce our effort on pre-processing by using Microsoft Kinect. The Kinect V2 is a recently-developed depth sensor and can directly provide the 3D point cloud sequence for generating the motion stream of human face. Figure 1 illustrates the facial skeleton (wire-frame) generated by using Kinect HD face API.

Mao et al. [4] described that human faces were 3D object, 2D images were insufficient to represent the geometrical feature. Therefore, they decided to use two sorts of feature generated by Kinect. Multi-pose was requiring in their database because their approach was still unable to deal with viewpoint variation and anthropometry variation.

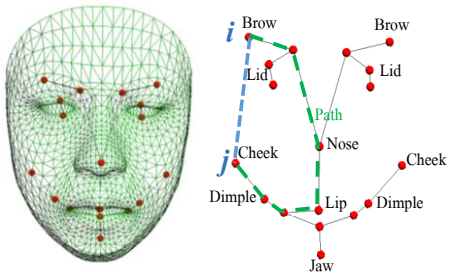

Figure 1　Facial skeleton consists of 18 feature points. The blue dotted line indicates direct distance between points *i* and *j*. The green dotted line indicates path distance between points *i* and *j*

Zhao et al. in [3] introduced a novel approach called Structured Streaming Skeleton (SSS). Their approach succeeded on handling all those intra-class variations by utilizing the streams generated from moving body skeleton this might be applicable to facial expressions recognition - if we treat them as facial gestures. So, we have decided to adapt SSS feature extraction approach to our approach in order to distinguish human facial emotion rather than human body gesture.

3. Proposed approach

Our approach is based on SSS feature extraction as described in the related work. Therefore, the framework of our system could be shown as in Figure 2.

In data stream generation phase, FACS of Ekman et al. [6] is employed to select the proper vertices used for generating the motion streams (Red vertices in Figure 1). The motion streams will be calculated using Equations (1) and (2)

$$S_{ij}(t) = \frac{E(p_i(t), p_j(t))}{Path_{ij}} \quad (1)$$

$$Path_{ij} = \sum_{m=1}^{\#Node\_between\_point\_ij-1} E\left(p_{L_m}, p_{L_{m+1}}\right) \quad (2)$$

Where $S_{ij}(t)$ stands for normalized distance between point *i* and *j* at frame *t*, $p_i(t)$ and $p_j(t)$ stand for coordinate (*x*,*y* and *z*) of points, L_m stands for sorted point indices list of particular $Path_{ij}$ indexed by *m*, $E(p_i(t), p_j(t))$ is Euclidean distance between points, $Path_{ij}(t)$ is path distance between points.

After receiving the streams, now we are ready to feed them into SSS based feature extraction.

Prior to doing feature extraction, we must generate template dictionary for feature extraction following instructions in [1, 2] After receiving template dictionary which is generated from the data streams, we will use it to extract the feature vectors from the stream. Therefore, all streams will be scanned again to calculate

† Graduate School of Science and Technology, Kumamoto University.
‡ Electronic Systems Technology and Information Systems Technology, Kumamoto Prefectural College of Technology.

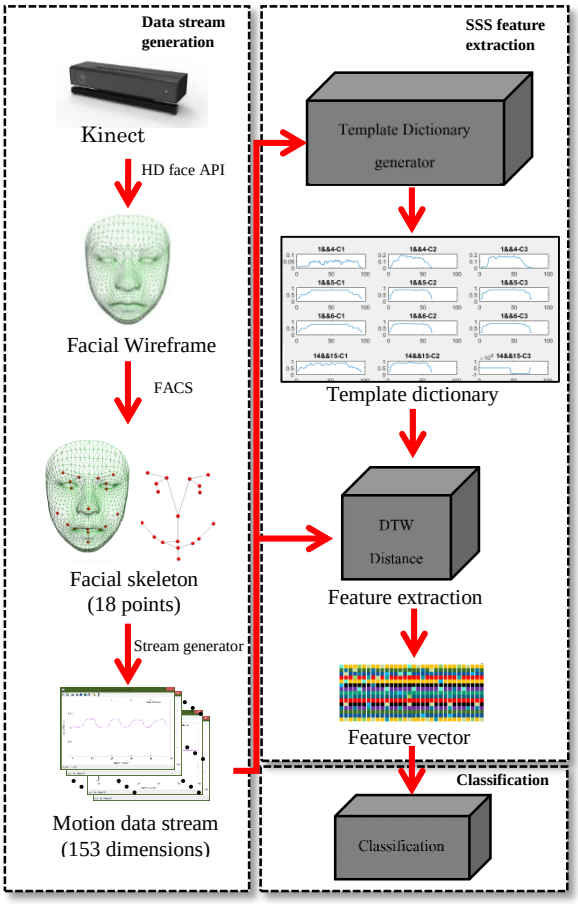

Figure 2　Overall of frame work of proposed approach

Dynamic Time Warping (DTW) distance starting from current frame with each sequence inside template dictionary.

4. Experiment

Fifteen people participated in dataset creation, eight emotions per each, each emotion of each person has 325 frames, and ten samples were selected from 325 frames of each emotion.

In this experiment, we conducted the qualitative experiment by using two kinds of feature vectors: the SSS feature vector and the simple stream feature vector. The SSS feature vector is prepared by using the method described in [1, 2] and it consists of 765 attributes per one vector. The stream feature vector consists of 153 attributes of Euclidean distance between the particular pairs on the facial skeleton. To the best of our knowledge, we conducted fair comparisons between our approach and the state- of-the-art approach [4] by considering different factors such as dataset, number of classes and so forth.

Table 1 Accuracy comparison

Approach↓	Accuracy (%)
State-of-the-art approach [4]	80.57
SSS feature (K-NN)	81.52
SSS feature (SVM)	66.76
Stream feature (K-NN)	71.78
Stream feature (SVM)	41.8

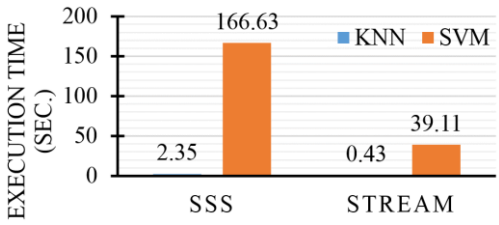

Figure 3　Execution time

Table 1 shows the result of our approach (Stream and SSS based) and the state-of-the-art approach. The present approach using the SSS feature attained accurate rate of 81.52% which was comparable to that of the as state-of-the-art approach and it could reduce the effects of intra-class variations in the system compared with the state-of-the-art approach.

Figure 3 shows overall execution time, K-NN algorithm using the SSS feature and the stream one required execution times of 2.35 and 0.43 seconds, respectively. The ratio of 2.35 to 0.43 becomes 5.60. The SVM using the SSS feature and the stream one spent these of 166.63 and 39.11 seconds, respectively. The ratio of 166.63 to 39.11 becomes 4.86. The execution time using the stream features are approximately 5.23 times faster than that using the SSS features.

5. Conclusion

We propose the novel approach utilizing a Kinect sensor to extract the facial expression features from movement of the facial skeleton model. We conducted the quantitative comparisons between our method and the state-of-the-art approach [4]. In the quantitative experiments, we achieved all the goals as we mentioned in the introduction of this paper. The proposed approach had reduced the effects of intra-class variations in the human facial emotion recognition system. It could be concluded that our approach has achieved superiority over previously reported approaches by overcoming the intra-class variations. For the future work, we consider that is necessary to study the performance dependence on other parameters and implement real-time classification.

References

[1] N. Chanthaphan, K. Uchimura, T. Satonaka and T. Makioka, “Facial emotion recognition based on facial motion stream generated by Kinect”, Proc. of 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Bangkok, pp.117-124 (2015)

[2] N. Chanthaphan, K. Uchimura, T. Satonaka and T. Makioka, “New feature extraction method for facial emotion recognition by using Kinect”, Proc. of The Korea-Japan joint workshop on Frontiers of Computer Vision (FCV), Takayama, pp.200-205 (2016)

[3] X. Zhao, X. Li, C. Pang, Q. Z. Sheng, S. Wang and M. Ye ,“Structured streaming skeleton - a new feature for online human gesture recognition”, ACM Trans. Multimedia Comput. Commun. Appl., 11, 1, pp. 1-18 (2014)

[4] Q. R. Mao, X. Y. Pan, Y. Z. Zhan and X. J. Shen, “Using Kinect for real-time emotion recognition via facial expressions”, Frontiers of Information Technology & Electronic Engineering, vol.16, no.4, pp.272-282 (2015)

[5] X. Zhu and D. Ramanan, “Face detection, pose estimation, and land mark localization in the wild”, Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), pp. 2879-2886 (2012)

[6] P. Ekman and W. Friesen, “Measuring facial movement”, Environmental psychology and nonverbal behavior, pp.56-75 (1976)