Machine-Learning Based Fitness Behavior Recognition from Camera and Sensor Modalities

Chih-Chieh Fang
Graduate Institute of Dance Theory
Taipei National University of Art
Taipei, Taiwan
m10446017@gmail.com

Ting-Chen Mou
Dept. Communication Engineering
National Central University
Taoyuan, Taiwan
mou15361226@gmail.com

Shih-Wei Sun
Department of New Media Art
Taipei National University of Art
Taipei, Taiwan
swsun@newmedia.tnua.edu.tw

Pao-Chi Chang
Dept. Communication Engineering
National Central University
Taoyuan, Taiwan
pcchang@ce.ncu.edu.tw

***Abstract*—We implemented a prototype system to recognize fitness behaviors using the skeleton information from the camera modality and the accelerometer/gyro sensor values. In addition, by fusing the camera modality and the sensor modality, the recognition accuracy of the complex fitness behaviors can be improved.**

Keywords—behavior recognition, camera, sensor, multi-modal, fusion

I. Introduction

Behavior recognition is widely used in many human-computer interface (HCI) applications, e.g., Kinect camera from Xbox One [1], and Labo for Nitendo Switch [2]. The above-mentioned devices utilize the RGB-D camera [1], accelerometer, and gyro sensors [2]. In addition, to recognize human actions, Chen et al. [3] proposed a sensor fusion approach based on depth and inertial sensors to train multiple action classifiers. On the other hand, Tapia et al. [4] used the wireless accelerometer and a heart rate monitor to recognize the physical fitness activity in real-time. However, to precisely recognize the fitness behavior is still a challenging issue. In this paper, we propose a machine learning-based fitness behavior recognition system to be operated in real-time.

II. Proposed Method

As shown in Fig. 1, the raw data of the camera-based modality and the sensor-based modality are captured from a Kinect camera [1] and a x-OSC [5] device, correspondingly. The raw data needs to be segmented, low-pass filtered, and the feature vectors are extracted from the corresponding modalities. Next, the features are concatenated and fed into a machine learning based classifiers, as shown in Fig. 1.

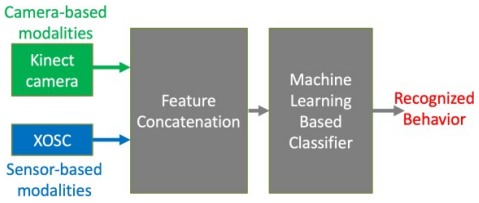

Fig. 1. Machine learning architecture from camera-based and sensor-based modalities.

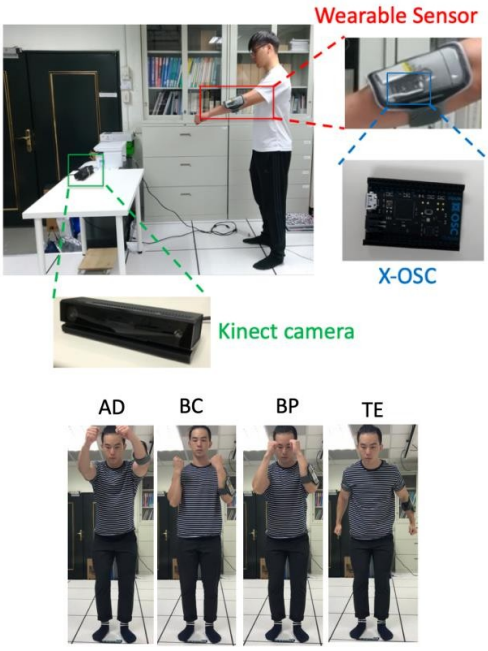

Fig. 2. The environment of the mounted camera and the worn sensor device; the fitness behaviors: anterior deltoid (AD), biceps curl (BC), back pullover (BP), and triceps extension (TE).

The testing environment of the proposed system is shown by the upper part of Fig. 2. The Kinect camera is mounted in front of a user, and the x-OSC device is worn on the arm of a user. We should notice that, the real-time sensor values of accelerometer and gyro sensor are sent from x-OSC through a Wifi connection. As shown in the lower part of Fig. 2, the target fitness behaviors for recognition in the implemented prototype are: anterior deltoid (AD), biceps curl (BC), back pullover (BP), and triceps extension (TE). The raw data captured from Kinect camera is shown in Fig. 3, and the sensor data captured from x-OSC is shown in Fig. 4. For example, the 3D position data of the wrist joint (from Kinect camera) on the left hand is depicted in Fig. 3. In addition, S1 to S5 are the data captured from subject 1 to subject 5. It is obvious that the signal distribution for the same behavior (e.g. anterior deltoid) from different users (S1 to S5) have similar waveform. On the other hand, the corresponding inertial signals from accelerometer and gyro sensor are depicted in Fig. 4 (a) and Fig. 4 (b),

978-1-5386-9269-1/18/$31.00 ©2018 IEEE
DOI 10.1109/AIVR.2018.00058

correspondingly. Similarly, the signal waveforms of the same behavior acted by different users still have similar manners.

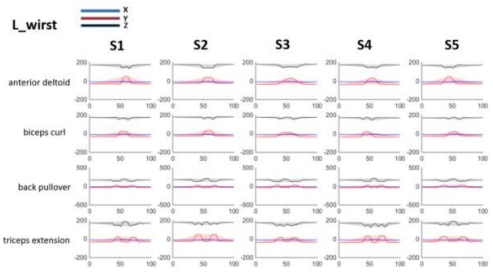

Fig. 3. The features extracted from the Kinect camera acted by 5 users from four different behaviors: anterior deltoid, biceps curl, back pullover, and triceps extension.

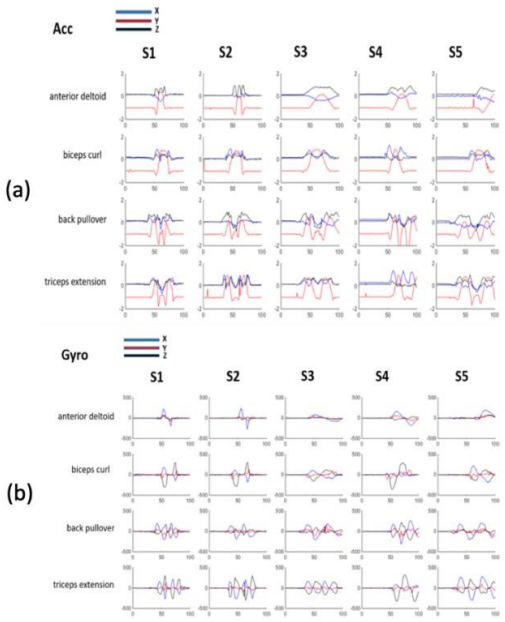

Fig. 4. Row data captured from sensor signals acted by 5 users from four different behaviors: anterior deltoid, biceps curl, back pullover, and triceps extension, (a) the accelerometer signals, and (b) the gyro sensor signals.

Once the raw data of the camera modality and the sensor modality can be obtained, the data in all modalities are divided into four time-equal regions, and the features are extracted by calculating the mean, maximum, minimum, and standard deviation of each region. Next, the features extracted from multiple modalities are concatenated as feature vectors. Furthermore, a conventional support vector machine (SVM) algorithm is adopted to train the classifiers in the proposed machine learning prototype for behavior recognition.

In the preliminary study, 7 volunteers (6 males and 1 female) are invited to act 4 different fitness behaviors. Each behavior is acted for 10 times from the same person. In the experimental results, a random chosen 6 person's data are used for training and 1 person's data is used for testing.

TABLE I. BEHAVIOR RECOGNITION RESULTS

(a) Camera only Modality

Predict Behavior \ Actual Behavior	AD	BC	BP	TE
AD	95.7%	2.9%	4.3%	0%
BC	0%	90%	0%	2.9%
BP	4.3%	1.4%	94.3%	11.4%
TE	0%	5.7%	1.4%	85.7%

(b) Sensor only Modality

Predict Behavior \ Actual Behavior	AD	BC	BP	TE
AD	54.3%	15.7%	28.6%	1.4%
BC	8.6%	**47.1%**	5.7%	21.4%
BP	37.1%	22.9%	60%	11.4%
TE	0%	14.3%	5.7%	**65.7%**

(c) Feature Fusion (Camera + Sensor)

Predict Behavior \ Actual Behavior	AD	BC	BP	TE
AD	94.3%	1.4%	17.1%	0%
BC	0%	**97.1%**	0%	10%
BP	5.7%	0%	82.9%	1.4%
TE	0%	1.4%	0%	**88.6%**

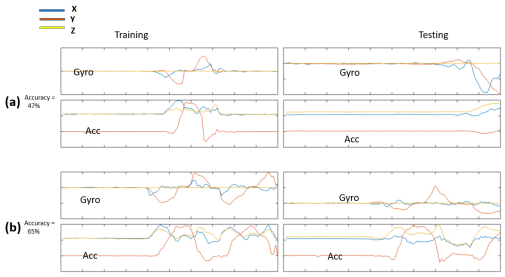

Fig. 5. The sensor raw data of accelerometer and gyro: (a) BC behavior, and (b) TE behavior.

III. Experimental Results and Discussion

The confusion matrices shown in Table I represent the accuracy results of camera-only, sensor-only, and feature fusion (camera + sensor) results. As shown in Table I (b), the sensor-only results has limited behavior recognition capability. However, when the features are fused from the camera modality to the sensor modality, the accuracy is improved, as shown in Table I (c). For example, as shown by the boldface digits in BC and TE behaviors in Table I (b) and (c), the accuracy results are improved with about 7% and 3% from the camera-only modality. Compared with the sensor-only results, the accuracy is improved more than 50% and 20% of BC and TE behaviors, respectively.

On the other hand, a representative raw sensor data (the bold face 47.1% in Table I (b)) of BC behavior is depicted in Fig. 5 (a). In this example, the sensor pattern is quite different from the training data to the testing data. This is the main reason to cause the low behavior recognition accuracy. However, another representative raw sensor data (the bold face 65.7% in Table I (b)) of the TE behavior has similar patterns (Fig. 5(b)), resulting higher accuracy. Nevertheless, by fusing the features extracted from the camera modality and the sensor modality, the final fitness behavior recognition accuracy can be improved.

References

[1] https://www.microsoft.com/en-us/download/details.aspx?id=44561

[2] https://labo.nintendo.com/

[3] C. Chen, R. Jafari, and N. Kehtarnavaz, "A Real-Time Human Action Recognition System Using Depth and Inertial Sensor Fusion", IEEE Sensors Journal, Vol. 16, No. 3, pp. 773 – 781, 2016.

[4] E.M. Tapia et al., "Real-Time Recognition of Physical Activities and Their Intensities Using Wireless Accelerometers and a Heart Rate Monitor", IEEE Intl. Symp. Wearable Computers, 2007

[5] http://x-io.co.uk/x-osc/