World Wide Web (2019) 22:1343–1358
https://doi.org/10.1007/s11280-018-0567-0

Self-feeding frequency estimation and eating action recognition from skeletal representation using Kinect

Qianhui Men[1] · Howard Leung[1] · Yang Yang[2]

Received: 13 December 2017 / Revised: 29 March 2018 / Accepted: 8 April 2018
Published online: 21 April 2018

Abstract Under healthcare research, eating activity detection and recognition have been studied for many years. Most of the previous approaches rely on body worn sensors for eating behavior detection. However, measurement errors from these sensors will largely reduce the tracking accuracy in estimating velocity and acceleration. To avoid this problem, we utilize Microsoft Kinect to capture skeleton motions of eating and drinking behaviors. In this paper we introduce a moving average method to remove the noise in relative distances of the captured joint positions so that eating activity can be segmented into feeding and non-feeding frames. In order to identify different eating patterns, eating and drinking behavior recognition is performed based on the features extracted from the resulting feeding periods. The experiments are evaluated on our collected eating and drinking action dataset, and we also change the distance between Kinect and subject to test the robustness of our approach. The results achieve better detection and recognition performance compared with other approaches. This pioneer work of our eating action behavior analysis can lead to many potential applications such as the development of a Web system to facilitate people to share and search their eating and drinking actions as well as carrying out intelligent analysis to provide suggestions.

This article belongs to the Topical Collection: *Special Issue on Social Media and Interactive Technologies*
Guest Editors: Timothy K. Shih, Lin Hui, Somchoke Ruengittinun, and Qing Li

✉ Qianhui Men
qianhumen2-c@my.cityu.edu.hk

Howard Leung
howard@cityu.edu.hk

Yang Yang
yyoung@ujs.edu.cn

[1] Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

[2] School of Computer Science and Communication Engineering, Jiangsu University, Jiangsu, China

Keywords Skeleton motions · Microsoft Kinect · Self-feeding detection · Eating action recognition

1 Introduction

Overweight has become a world-wide growing problem for people. One major cause of such problem is due to a person's fast eating pace which is considered as an unhealthy eating behavior [24], and this poor habit forms unconsciously most of the time. Eating too fast will likely lead to digestive problems such as obesity and malnutrition. It is essential to generate a monitoring system to evaluate eating behavior and detect abnormal eating patterns. In this paper, upper-body skeleton motions of eating and drinking behaviors are captured using Kinect. With the development of Kinect, human action detection and recognition technologies have been utilized in many life aspects such as gaming and healthcare applications. For example, Kinect-based physical therapy system can instruct patients to accurately perform rehabilitative actions [1, 4, 22]. Kinect technology is also applied in fall detection and prevention of healthcare monitoring [3, 17, 18, 32]. Online monitoring of people's eating and drinking behaviors can be made possible as Kinect can provide real-time streams of RGB, depth and skeleton information, which can be transmitted through the Web.

Previous studies in action recognition mostly focus on the classification among actions with large inter-class variations such as "kicking", "pitching" and "sitting" [7, 8, 14, 27]. These discriminative actions are carried out by different body parts. However, eating and drinking actions are all performed by the feeding arm(s) of the upper body, therefore the classification among these similar actions may require more representative features. Besides, the current public motion datasets do not contain many variations of eating and drinking actions. As a result, we start to collect our own eating and drinking motion dataset and present here a preliminary work in classifying eating related behaviors by tracking skeleton motion. The skeleton data can provide a clue about the subject's action without the need to analyze the texture information from videos.

In this work, we assume that eating and drinking actions can be generally divided into two consecutive processes: self-feeding (active) periods and resting (inactive) periods. The hand(s) that perform(s) self-feeding is(are) feeding hand(s). During self-feeding periods, people bring food to their mouth to consume, and during resting periods, the feeding hand(s) is(are) down with almost no movements. Pace of intake is reflected by self-feeding frequency estimation. We then give insight into the recognition of eating and drinking patterns by tracking skeleton motion of the upper body. It is nontrivial to recognize people's eating patterns because of the diversity in utensils and food types. For example, people always consume food with knife and fork in western-style eating patterns. While in Chinese eating style, people prefer to use chopsticks to eat. In order to present an unbiased motion recognition, we categorize different eating and drinking patterns including eating with single hand, both hands, knife and fork, chopsticks and drinking with spoon. We capture subjects eating different types of food including fries, burgers, steaks, noodles and soup. In addition to drinking with spoon, we also recognize common drinking patterns including drinking with straw and drinking bottled water. After this pioneer work, we will extend our dataset to include more variations in food diversity for eating and drinking actions. There are many potential applications by tracking and recognizing the eating and drinking behaviors during food and beverage consumption. For example, a Web system can be developed to allow people to upload and share their eating and drinking actions and our recognition result provides an easy way for searching and gathering useful information for carrying out big data analysis in the future.

We extract the main contributions of this paper as follows.

- An eating and drinking motion dataset is collected for research.
- Skeletal measurements are introduced to represent eating and drinking actions during food and beverage consumption.
- We use a moving average K-means clustering method on the proposed skeletal action representation to estimate self-feeding frequencies.
- We perform eating and drinking action recognition based on the proposed skeletal representation and show better performance compared with other methods.

The rest of this paper is structured as follows. Section 2 reviews some previous work related to our research. Section 3 gives an introduction of our collected eating and drinking motion dataset. Section 4 describes the skeletal representation and provides a visualization of the joint pair distances during eating and drinking. The details of our proposed method in eating action segmentation and recognition are described in Sections 5, and 6 presents our experiments and results. In Section 7, we make a conclusion of our paper and illustrate potential future work.

2 Related work

Monitoring ingestive behavior has gained great interest as more and more people are getting overweight, and obesity is always one inducement for some diseases like cardiovascular disorders and diabetes [16]. Eating and drinking behaviors can be detected and monitored through many aspects including food and calorie intake, motions and gestures of arms, or self-feeding frequencies, chews and swallows.

Food recognition faces great difficulties because of variations in food shapes and diversity in food types. Yang et al. [29] identified food images in Pittsburgh fast-food image dataset [5] via multi-dimensional histograms generated by pairwise local features. Wu and Yang [28] used scale-invariant feature transform (SIFT) to describe keypoints of food images in training set and image frames in video before matching. Combined with SIFT, Zong et al. [33] proposed the Local Binary Pattern (LBP) to represent the global structure of the food image and obtained higher classification accuracy than [29] within the same dataset. After food identification, calorie and nutrition estimations are always expected to form a relatively clear blueprint of energy consumption. Yang et al. [29] also estimated calorie intake in a meal according to the fast food appearance, but the error estimation occurred partially due to the exclusion of the food portion. Kawano and Yanai [13] designed a smartphone system to recognize food and estimate calories. They segmented single food item in the ground truth bounding box and extracted bag-of-SURF features before recognition.

Wrist-worn devices and approaches have been widely studied to detect and recognize motions or gestures of arms in eating and drinking activities. With accelerometers attached to testers' both wrists, Zhang et al. [31] extracted features of Euler angles and classified repeated eating or drinking patterns through a proposed hierarchical temporal memory network. Amft et al. [2] attempted to recognize different drinking postures in nine types of containers and achieved a recognition rate above 70% on average, and the signal traces of three-dimensional acceleration and gyroscope help them to discriminate fetch and sip drinking motions. Based on a wrist-fastened inertial sensor, Thomaz et al. [26] compared gestures of food intake, i.e., eating with knife and fork, spoon or fork, and eating with hand, with other activities of daily living (ADL) (e.g., walking, chatting, waiting), and their results were evaluated on a random forest classification algorithm.

Wrist-worn and other body-worn devices like microphone and smart necklace can also be designed to detect and monitor upper-body motion during the ingestion of bites, chews and swallows. In the study of Dong et al. [6], they counted biting times by detecting the motion pattern of wrist rotation using a watch-resembled sensor. They further expanded the number of participants in order to analyze the relationship between bite patterns with age, gender and race variations [23]. Moreover, Zhang and Amft [30] utilized a smart eyeglass to reveal the hardness of food in the process of chewing banana, cucumber and carrot. Kalantarian et al. [11, 12] detected and categorized swallows into different food classes using a piezoelectric sensor placed between neck and chest. Nguyen et al. [19] proposed SwallowNet which was formed by recurrent neural network (RNN) with long short-term memory (LSTM) units, and their swallow evaluation results (average 76.07%) outperformed the random forest classifier (average 66.58%). The problem of inertial sensors is that drift will be accumulated in the integration of orientation and velocity measurements, which has become a major drawback for the state-of-the-art acceleration sensor tracking systems [31]. As a result, in this paper we propose to use Kinect to track skeleton motions in eating or drinking actions to avoid this problem since the Kinect sensor is fixed during motion capture. In food intake monitoring, Kinect was used to track the arm gestures [10] which was considered to be more convenient than wearable sensors. A major contribution in their work was to detect food intake using raw skeleton positions and main angles of upper body. In our work, we focus on joint distance features to segment and recognize different eating or drinking actions.

3 Data collection

Existing public motion datasets contain very few eating and drinking motions. As a result, we collect our own eating and drinking motion dataset using Kinect. The dataset includes five subjects with variations in upper-body sizes performing seven eating and drinking actions: *eating fries, eating burger, eating steak, eating noodles, drinking soup, drinking soft drink with straw and drinking bottled water*. Each eating or drinking process is performed four times and captured as four sequences. Food or drink and its corresponding eating or drinking utensils are well prepared on the desk in front of the fixed Kinect sensor before recording. In each sequence, the subject is told to eat one type of food or drink at his or her own pace. All the sequences are recorded within 50 seconds at 30 frames per second (FPS).

The dataset contains skeleton positions of 10 joints in human upper body as shown in Figure 1. The visualization of our dataset with screenshot instances of seven actions is illustrated in Figure 2.

4 Skeletal representation

4.1 Joint relative distance: JRD

Let $M = \{M_1, M_2, ..., M_T\}$ be an eating or drinking action in T frames. For each joint i $(1 \leq i \leq 10)$ in human upper body, we consider $J_i(t) = (x_i(t), y_i(t), z_i(t))$ as the joint

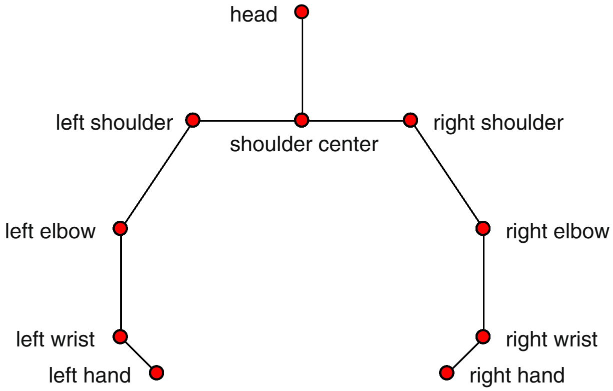


Figure 1 Skeleton joints of human upper body captured by Kinect

position at frame t ($1 \leq t \leq T$). Let p denote the joint pair index at each frame. Because there are 10 joints in our upper-body model, the total number of joint pairs is $\binom{10}{2} = 45$. For the pth ($1 \leq p \leq 45$) joint pair at the tth frame, the Joint Relative Distance (JRD) [25] is computed by the L_2-norm:

$$JRD(t, p) = d_{L_2}(J_i(t), J_j(t)). \tag{1}$$

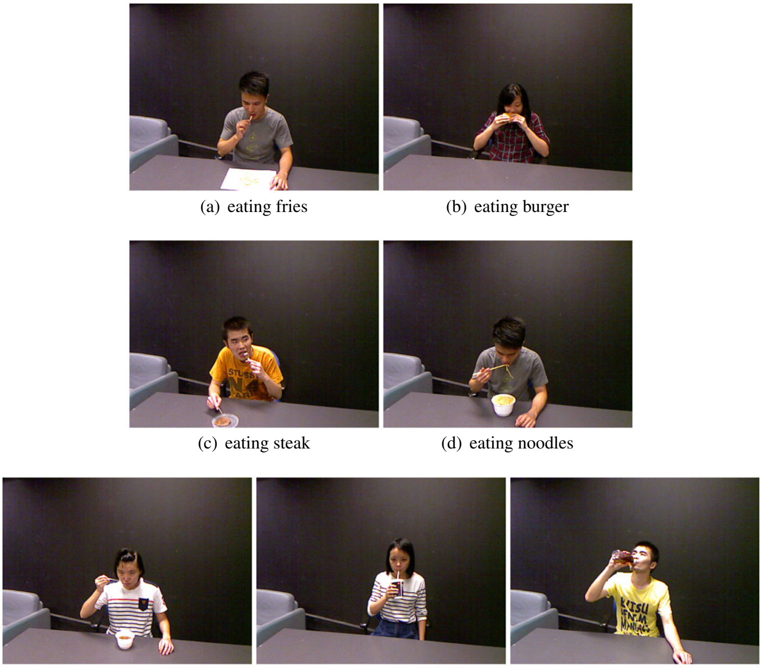

(a) eating fries
(b) eating burger
(c) eating steak
(d) eating noodles
(e) drinking soup
(f) drinking soft drink with straw
(g) drinking bottled water

Figure 2 Example frames of our eating and drinking action dataset

4.2 Visualized representation

Eating or drinking actions can be regarded as using hand(s) to put food into mouth, and shoulder center is the joint which is believed to be the closest to mouth. We present several JRDs between shoulder center joint and joints of double hands over the whole sequences in Figure 3, and this will give out an intuitive representation of self-feeding periods during a consumption process. There are seven subfigures in Figure 3 corresponding to seven different types of eating and drinking sequences in our dataset. We can draw clearly that in each subfigure, frames with smaller distances reflect that the subject is raising hand(s) up and performing self-feeding. On the other hand, frames with larger distances reflect that the subject has already put down the feeding hand(s) and resting. The persistent length of frames in the same self-feeding action can be seen as the duration time of single feeding. Compared with other types of consumption, it is clear that left hand performs feeding in having steak since the variation of left hand is more significant than right hand, and both hands are acting as feeding hands in eating burger since the variations are almost synchronous.

5 Proposed method

In this section, we present our proposed approach for analyzing eating and drinking motions captured by the Kinect sensor. The flow chart illustrated in Figure 4 provides an overview of our proposed method. The positions of 10 joints from the upper body are obtained from the Kinect sensor while capturing a person's eating or drinking motion. Based on the distance between shoulder center and each hand, the feeding hand is identified. The distance between shoulder center and the feeding hand is further analyzed to perform segmentation to divide the motion into feeding and non-feeding frames. Based on the feeding frames, on one hand we estimate the self-feeding frequency and on the other hand we perform recognition to identify the eating or drinking behavior.

5.1 Moving average joint relative distance: MAJRD

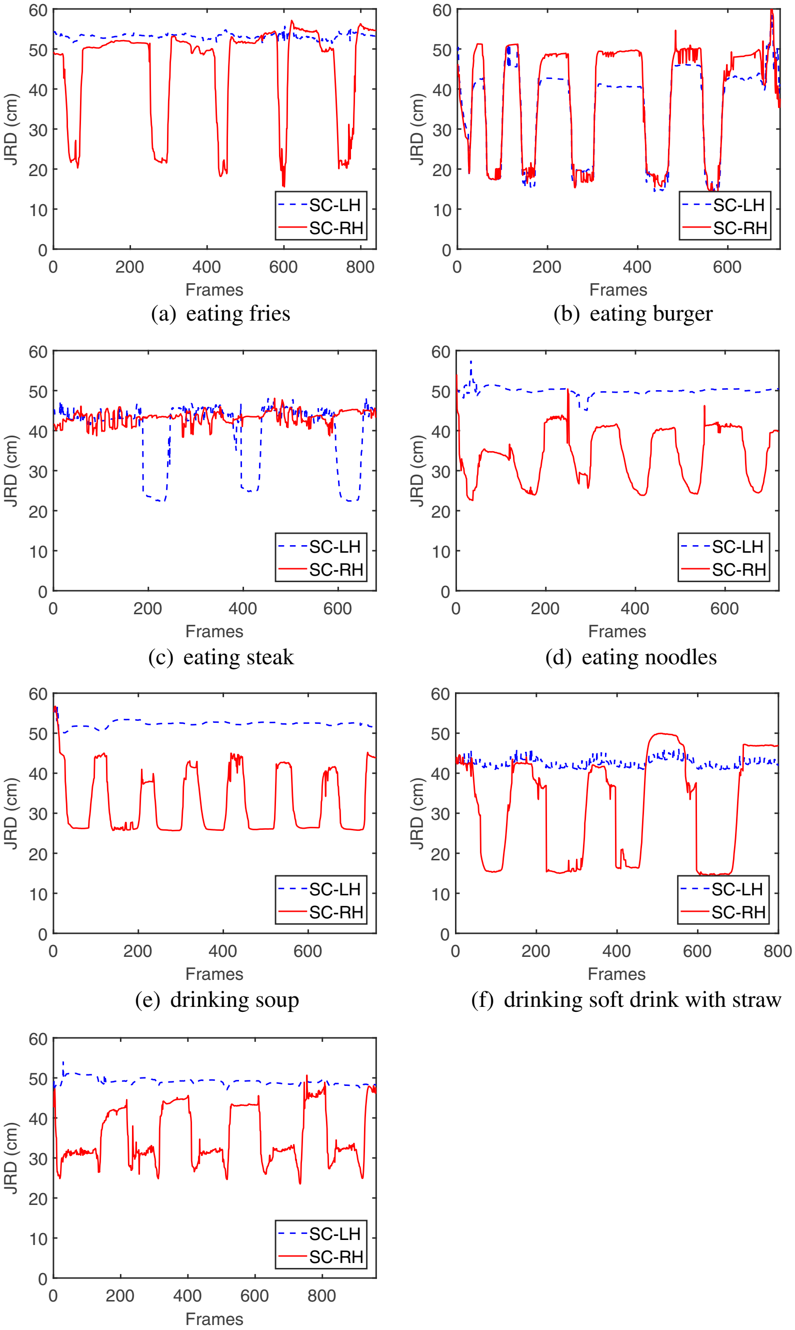
Since noise may exist in the tracked body motions from a depth camera, in order to prevent outliers, we compute the moving average of JRD, which we denote as MAJRD. MAJRD can be considered as a smoothed version of JRD over a sliding window with size a. The formal definition of MAJRD for the joint pair p at frame t is provided as follows:

$$MAJRD(t,p) = \begin{cases} \frac{1}{2t-1}\sum_{l=1}^{2t-1} JRD(l,p) & \text{if } 1 \leq t \leq \frac{a-1}{2}, \\ \frac{1}{2T-(2t-1)}\sum_{l=2t-T}^{T} JRD(l,p) & \text{if } (T+1) - \frac{(a-1)}{2} \leq t \leq T, \\ \frac{1}{a}\sum_{l=t-\frac{a-1}{2}}^{t+\frac{a-1}{2}} JRD(l,p) & otherwise. \end{cases} \tag{2}$$

Note that the window size a is a positive integer and it should also be an odd number so that we can weigh over the same number of frames on each side with respect to a given frame.

5.2 Segmentation into feeding and non-feeding frames

We first detect which arm or hand performs self-feeding. The feeding hand is determined by the one with smaller MAJRD of the two joint pairs (shoulder center to left hand, shoulder



(a) eating fries

(b) eating burger

(c) eating steak

(d) eating noodles

(e) drinking soup

(f) drinking soft drink with straw

(g) drinking bottled water

Figure 3 Example JRDs of seven categories between shoulder center and double hands. SC-LH denotes the joint pair between shoulder center and left hand. SC-RH denotes the joint pair between shoulder center and right hand

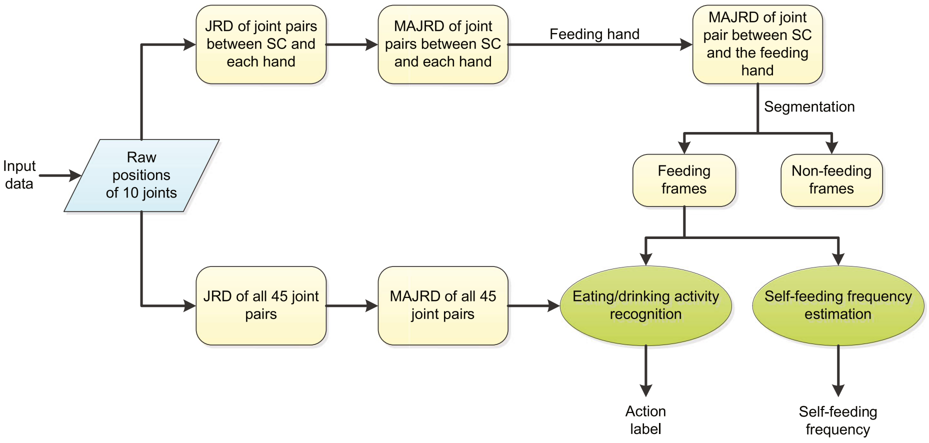

Input data
Raw positions of 10 joints
JRD of joint pairs between SC and each hand
MAJRD of joint pairs between SC and each hand
Feeding hand
MAJRD of joint pair between SC and the feeding hand
Segmentation
Feeding frames
Non-feeding frames
JRD of all 45 joint pairs
MAJRD of all 45 joint pairs
Eating/drinking activity recognition
Self-feeding frequency estimation
Action label
Self-feeding frequency

Figure 4 Flow chart of our proposed method. SC denotes shoulder center

center to right hand) averaged over whole sequence. The mean of MAJRD for a joint pair p in action M can be computed by the following equation:

$$\overline{MAJRD_M(p)} = mean_t(MAJRD_M(t, p)). \tag{3}$$

As mentioned in Section 1, eating or drinking process consists of two main partitions: self-feeding periods and resting (non-feeding) periods. These two partitions can be discriminated by examining the MAJRDs between shoulder center and the feeding hand at different frames. In a feeding period, the MAJRDs between shoulder center and the feeding hand are similar to the MAJRDs between these two same joints in other feeding periods within one sequence, which tend to be small as the feeding hand is moving toward the mouth. On the other hand, the MAJRDs during the resting periods are relatively large compared with the self-feeding periods. As a result, we apply K-means clustering on the MAJRDs between the shoulder center and the feeding hand in order to classify the frames into feeding and non-feeding.

The K-means clustering algorithm [9] is applied to segment an eating or drinking motion sequence into feeding and non-feeding frames. In order to give out a relatively stable clustering, we initialize the centroids of the k clusters with the first k peaks in the probability density function (PDF) of MAJRDs estimated over the entire sequence with T frames. In this algorithm, the MAJRD at each frame is assigned to its nearest cluster center. Each cluster center will be revised to take the average of its electorate. The clustering algorithm converges when the assignment of all the MAJRDs over T frames is unchanged. The number of clusters in this problem is set with $k = 2$. Frames in the cluster with lower mean MAJRD are considered as self-feeding periods, and frames in the other cluster are considered as non-feeding or resting periods.

5.3 Self-feeding frequency estimation

During each feeding period, the frames are classified by the method in Section 5.2 as feeding frames and they form a block. It is then followed by a block of frames classified as non-feeding frames as the subject moves the feeding hand away from the mouth. These blocks of feeding and non-feeding periods interleave with each other with similar patterns, so we can estimate the self-feeding frequency by counting the number of blocks with feeding frames. Figure 5 shows two example sequences of eating fries with different self-feeding frequencies in a clip of 500 frames, and in both streams the subjects use their left hands as feeding hands. From Figure 5 we can observe that the subject in sequence 1 feeds himself or herself 3 times, while the subject in sequence 2 self-feeds 5 times within the same time duration.

5.4 Eating and drinking behavior recognition

One objective of our work is to determine the eating and drinking behaviors of subjects by recognizing them eating or drinking with different utensils or consuming different types of food. If the entire motion sequence is used for recognition, then the recognition accuracy would be affected by the non-feeding frames as they may represent resting which provides little information about the actual eating or drinking actions. Since the motion sequence has already been segmented into feeding and non-feeding frames, the recognition can be applied by focusing on the feeding frames to increase the accuracy. We compute $\overline{\text{MAJRD}}$ for each of the 45 joint pairs from the feeding frames and form the feature vector. The feature vectors obtained from training samples in each class are used to train a support vector machine (SVM) as classifier. During recognition, given an input motion sequence that has been segmented into feeding and non-feeding frames using the method described in Section 5.2, the feature vector is obtained by computing $\overline{\text{MAJRD}}$ of the 45 joint pairs and passed to the SVM classifier to recognize the eating or drinking patterns.

6 Experiment

We perform some experiments to evaluate the performance of our proposed approach in self-feeding frequency estimation and eating/drinking action recognition.

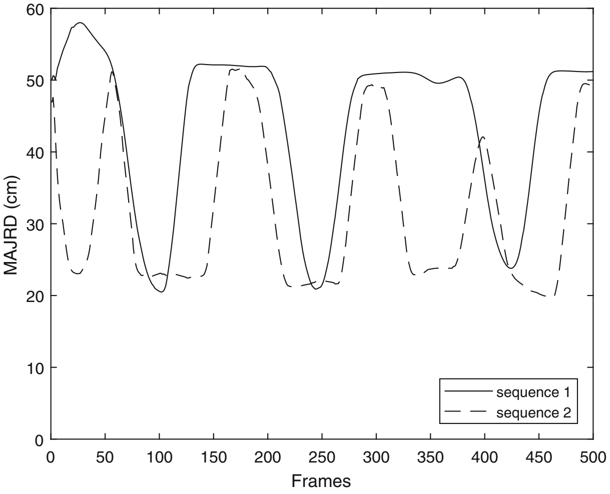


Figure 5 Two MAJRDs of eating fries in different feeding frequencies

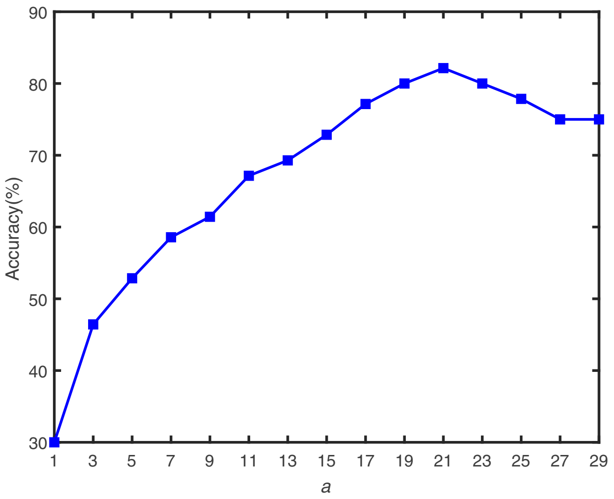


Figure 6 The detection accuracy of “MAJRD+K-means” with different *a* values

6.1 Experiment on self-feeding frequency estimation

As described in Section 5.3, we detect the feeding frames using K-means clustering on MAJRD between shoulder center and the feeding hand and then count the number of blocks with feeding frames to estimate the self-feeding frequency. In order to verify the effectiveness of our model, we compare our method (denoted as “MAJRD+K-means”) with the method applying K-means clustering on JRD without utilizing moving average (denoted as “JRD+K-means”), and two other joint smoothing methods called jitter removal filter [21] and Savitzky-Golay filter [20] also applying K-means clustering on JRD (denoted as “JRJRD+K-means” and “SGJRD+K-means” respectively). The ground truth of self-feeding frequencies of all the seven class sequences are given by manual annotation, and it is used to determine the accuracy of each method to evaluate the performance. The variation of “MAJRD+K-means” detection accuracy is performed with respect to different window size *a* as shown in Figure 6, and we can conclude that “MAJRD+K-means” gets the optimal performance when parameter *a* is set to be 21.

Table 1 illustrates the accuracies of self-feeding frequency detection for our dataset (described in Section 3) using the above four approaches. As indicated in Table 1, the estimated self-feeding frequency achieves the highest accuracy for our proposed approach “MAJRD+K-means”. Jitter removal filter is effective in dampening the spikes such that a

Table 1 Performance comparison for self-feeding frequency detection

Method	Accuracy (%)
JRD+K-means (baseline)	30.0
JRJRD+K-means [21]	75.7
SGJRD+K-means [20]	77.9
MAJRD+K-means (proposed)	82.1

spike will be replaced by the average of previous frames if a large disparity exists between the JRD of shoulder center and feeding hand in current frame and the average JRD of these two joints in previous frames. The JRDs of these two joints may vary significantly during eating, and applying jitter removal filter will make it less discriminative against feeding and non-feeding frames. Savitzky-Golay filter smooths skeleton data via minimizing the least-squares error, which is sensitive to outliers. This may reduce accuracy when estimating self-feeding frequencies. We further provide an example of MAJRD illustrating the clustering obtained from "JRD+K-means" and "MAJRD+K-means" under the same sequence of eating fries in Figure 7. The shaded areas represent the detected feeding frames. There are four feeding periods in this sequence from the ground truth annotation. We can see that "JRD+K-means" wrongly classifies the third feeding period into two feeding actions. This implies that "JRD+K-means" is unable to identify outliers, and "MAJRD+K-means" outperforms "JRD+K-means" to rapid fluctuations in captured data.

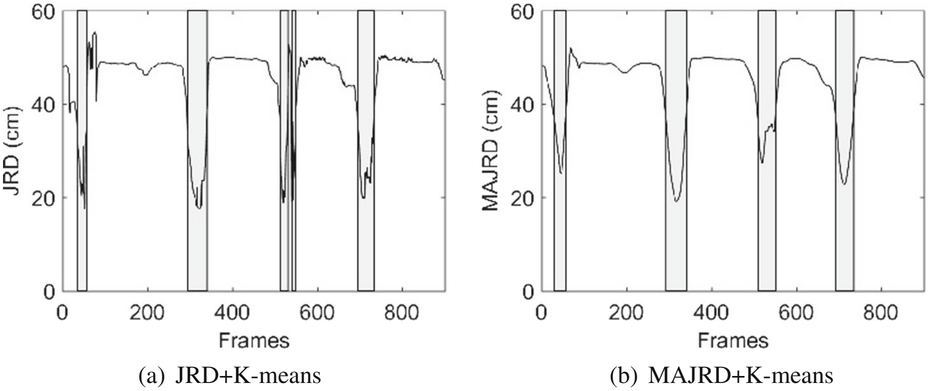
6.2 Experiment on eating/drinking action recognition

In this experiment, we execute classification of seven different eating or drinking actions in our dataset based on the features obtained in Section 5.4. The SVM classifiers are trained for action classification in a one-vs-all setting. In order to perform a subject-invariant experiment, each time the actions of three subjects are used for training, and the remaining two subjects are used for testing. The total number of combinations of training-testing settings is $\binom{5}{2} = 10$, and we take average of the recognition accuracies as the performance indicator.

To show the robustness of our proposed features, i.e., the MAJRD of the 45 joint pairs, we compare it with two other state-of-the-art skeletal representations, called the relative variance of JRD (RVJRD) [14], and the relative range of JRD (RRJRD) [15], which are defined for joint pair p in action M as follows:

$$RVJRD_M(p) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{JRD_M(t,p) - mean_t(JRD_M(t,p))}{mean_t(JRD_M(t,p))}\right)^2, \tag{4}$$

$$RRJRD_M(p) = \frac{\max_t(JRD_M(t,p)) - \min_t(JRD_M(t,p))}{mean_t(JRD_M(t,p))}. \tag{5}$$



(a) JRD+K-means (b) MAJRD+K-means

Figure 7 The detected feeding frames of "JRD+K-means" and "MAJRD+K-means" between shoulder center and the feeding hand under the same sequence of eating fries

As JRD is not robust enough to segment an eating or drinking action into feeding and non-feeding periods, the RVJRD and RRJRD are computed over the whole sequence. On the other hand, we make use of MAJRD to segment the sequence into feeding and non-feeding periods, and focus on the MAJRD from the feeding periods since it retains more representative characteristics of the eating or drinking action. The feeding frames we used are obtained from "MAJRD+K-means" with the best segmentation performance (i.e., $a = 21$). Thus we compare the RVJRD and RRJRD computed over the whole sequence with the relative variance of MAJRD (RVMAJRD) and the relative range of MAJRD (RRMAJRD) computed over the feeding frames. The RVMAJRD and RRMAJRD for joint pair p in action M are formally defined as follows:

$$RVMAJRD_M(p) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{MAJRD_M(t,p) - mean_t(MAJRD_M(t,p))}{mean_t(MAJRD_M(t,p))}\right)^2, \tag{6}$$

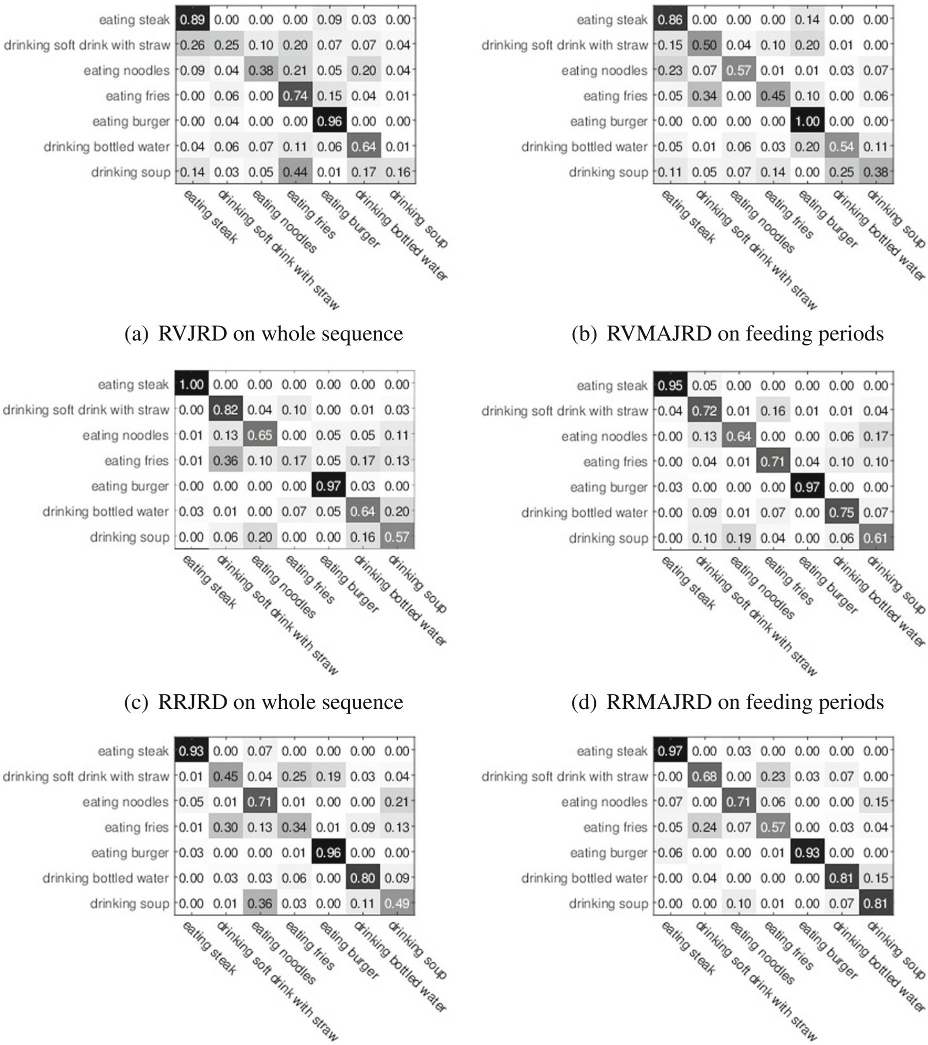
$$RRMAJRD_M(p) = \frac{\max_t(MAJRD_M(t,p)) - \min_t(MAJRD_M(t,p))}{mean_t(MAJRD_M(t,p))}. \tag{7}$$

We also compare the performance using the relative variance, the relative range as well as the mean of JRD (RVJRD, RRJRD vs $\overline{\text{JRD}}$) computed over whole sequence. In addition, we evaluate the recognition accuracies using RVMAJRD, RRMAJRD and $\overline{\text{MAJRD}}$ over feeding frames. The recognition accuracy results are shown in Table 2. It can be observed that all these three approaches computed from the feeding periods show higher recognition accuracies compared with the ones computed from the whole sequence. Since RRJRD is proposed under clear version data and RVJRD works better on noisy data, this may explain why RRJRD outperforms RVJRD after moving average step in eating and drinking behavior recognition. The performance of $\overline{\text{JRD}}$ is slightly lower than RRJRD, but after taking the moving average of its counterpart, $\overline{\text{MAJRD}}$ computed from feeding periods works better than RRMAJRD. The lower performance when the feature is extracted from the whole sequence (RVJRD, RRJRD or $\overline{\text{JRD}}$) is due to the fact that inter-class variations are small among various eating or drinking actions especially for $\overline{\text{JRD}}$. This is because resting and chewing occupy most of the time in an eating or drinking sequence, and the $\overline{\text{JRD}}$s over these inactive periods are similar among all the seven classes. On the other hand, when the feature is focused on the feeding periods (RVMAJRD, RRMAJRD or $\overline{\text{MAJRD}}$), it is more discriminative in representing self-feeding actions.

Figure 8 shows the confusion matrices for RVJRD, RRJRD, $\overline{\text{JRD}}$ on whole sequence as well as RVMAJRD, RRMAJRD, $\overline{\text{MAJRD}}$ on feeding periods. While our proposed method

Table 2 Performance comparison for eating/drinking action recognition

Method	Accuracy (%)
RVJRD on whole sequence	57.3
RVMAJRD on feeding periods	61.4
RRJRD on whole sequence	69.1
RRMAJRD on feeding periods	76.6
$\overline{\text{JRD}}$ on whole sequence	66.8
$\overline{\text{MAJRD}}$ on feeding periods	78.4



(a) RVJRD on whole sequence

	eating steak	drinking soft drink with straw	eating noodles	eating fries	eating burger	drinking bottled water	drinking soup
eating steak	0.89	0.00	0.00	0.00	0.09	0.03	0.00
drinking soft drink with straw	0.26	0.25	0.10	0.20	0.07	0.07	0.04
eating noodles	0.09	0.04	0.38	0.21	0.05	0.20	0.04
eating fries	0.00	0.06	0.00	0.74	0.15	0.04	0.01
eating burger	0.00	0.04	0.00	0.00	0.96	0.00	0.00
drinking bottled water	0.04	0.06	0.07	0.11	0.06	0.64	0.01
drinking soup	0.14	0.03	0.05	0.44	0.01	0.17	0.16

(b) RVMAJRD on feeding periods

	eating steak	drinking soft drink with straw	eating noodles	eating fries	eating burger	drinking bottled water	drinking soup
eating steak	0.86	0.00	0.00	0.00	0.14	0.00	0.00
drinking soft drink with straw	0.15	0.50	0.04	0.10	0.20	0.01	0.00
eating noodles	0.23	0.07	0.57	0.01	0.01	0.03	0.07
eating fries	0.05	0.34	0.00	0.45	0.10	0.00	0.06
eating burger	0.00	0.00	0.00	0.00	1.00	0.00	0.00
drinking bottled water	0.05	0.01	0.06	0.03	0.20	0.54	0.11
drinking soup	0.11	0.05	0.07	0.14	0.00	0.25	0.38

(c) RRJRD on whole sequence

	eating steak	drinking soft drink with straw	eating noodles	eating fries	eating burger	drinking bottled water	drinking soup
eating steak	1.00	0.00	0.00	0.00	0.00	0.00	0.00
drinking soft drink with straw	0.00	0.82	0.04	0.10	0.00	0.01	0.03
eating noodles	0.01	0.13	0.65	0.00	0.05	0.05	0.11
eating fries	0.01	0.36	0.10	0.17	0.05	0.17	0.13
eating burger	0.00	0.00	0.00	0.00	0.97	0.03	0.00
drinking bottled water	0.03	0.01	0.00	0.07	0.05	0.64	0.20
drinking soup	0.00	0.06	0.20	0.00	0.00	0.16	0.57

(d) RRMAJRD on feeding periods

	eating steak	drinking soft drink with straw	eating noodles	eating fries	eating burger	drinking bottled water	drinking soup
eating steak	0.95	0.05	0.00	0.00	0.00	0.00	0.00
drinking soft drink with straw	0.04	0.72	0.01	0.16	0.01	0.01	0.04
eating noodles	0.00	0.13	0.64	0.00	0.00	0.06	0.17
eating fries	0.00	0.04	0.01	0.71	0.04	0.10	0.10
eating burger	0.03	0.00	0.00	0.00	0.97	0.00	0.00
drinking bottled water	0.00	0.09	0.01	0.07	0.00	0.75	0.07
drinking soup	0.00	0.10	0.19	0.04	0.00	0.06	0.61

(e) JRD on whole sequence

	eating steak	drinking soft drink with straw	eating noodles	eating fries	eating burger	drinking bottled water	drinking soup
eating steak	0.93	0.00	0.07	0.00	0.00	0.00	0.00
drinking soft drink with straw	0.01	0.45	0.04	0.25	0.19	0.03	0.04
eating noodles	0.05	0.01	0.71	0.01	0.00	0.00	0.21
eating fries	0.01	0.30	0.13	0.34	0.01	0.09	0.13
eating burger	0.03	0.00	0.00	0.01	0.96	0.00	0.00
drinking bottled water	0.00	0.03	0.03	0.06	0.00	0.80	0.09
drinking soup	0.00	0.01	0.36	0.03	0.00	0.11	0.49

(f) MAJRD on feeding periods

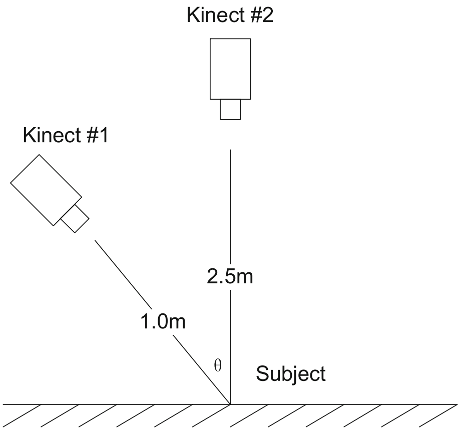
	eating steak	drinking soft drink with straw	eating noodles	eating fries	eating burger	drinking bottled water	drinking soup
eating steak	0.97	0.00	0.03	0.00	0.00	0.00	0.00
drinking soft drink with straw	0.00	0.68	0.00	0.23	0.03	0.07	0.00
eating noodles	0.07	0.00	0.71	0.06	0.00	0.00	0.15
eating fries	0.05	0.24	0.07	0.57	0.00	0.03	0.04
eating burger	0.06	0.00	0.00	0.01	0.93	0.00	0.00
drinking bottled water	0.00	0.04	0.00	0.00	0.00	0.81	0.15
drinking soup	0.00	0.00	0.10	0.01	0.00	0.07	0.81

Figure 8 Confusion matrices

can better recognize different classes of eating and drinking actions compared with other methods. The performance is lower when the eating or drinking actions have similar distances between shoulder center and the feeding hand during self-feeding actions, such as “eating fries” and “drinking soft drink with straw”, “eating noodles” and “drinking soup”.

Although our proposed $\overline{\text{MAJRD}}$ outperforms other features, it is still challenging to distinguish among some actions when they have similar $\overline{\text{MAJRD}}$ over feeding frames causing inconspicuous inter-class variations. Besides, different eating habits of people may cause large intra-class variations which provide another reason for causing poor performance.

Figure 9 Experiment configuration



6.3 Experiment on different distances between Kinect and subject

In order to test the influence of different distances between Kinect and subject as well as the robustness of our proposed $\overline{\text{MAJRD}}$ against RVMAJRD and RRMAJRD, we analyze skeletal representation of eating and drinking behaviors captured by two Kinects simultaneously. The experiment setting is demonstrated in Figure 9. The distances between subjects and two sensors are within the depth range of Kinect under seated mode (0.4m $\sim$ 3m). Two optical axes form a small angle θ to reduce interference between two sensors. In this experiment we have three subjects performing four actions: *drinking soup, drinking soft drink with straw, eating fries and eating noodles*. These actions are the most likely to cause confusions (described in Section 6.2) among all the seven eating and drinking behaviors, which makes the experiment more challenging. Analogous to our captured eating and drinking dataset (described in Section 3), each action is performed four times by each subject.

Table 3 gives out the recognition accuracies of the three mentioned methods RVMAJRD, RRMAJRD, and $\overline{\text{MAJRD}}$ under two Kinect-subject distances. For both near and remote modes, $\overline{\text{MAJRD}}$ shows the best performance among these three approaches, which shows the robustness of our method under different distances. In general, the performance of recognition under remote mode is lower than the one under near mode. This is because when the distance increases, the positions of joints are likely to be inferred which may cause inaccurate joint distances.

Table 3 Recognition accuracies (%) of eating/drinking action recognition under different Kinect-subject distances

Method	Near (1.0m)	Remote (2.5m)
RVMAJRD	58.3	56.3
RRMAJRD	77.1	66.7
$\overline{\text{MAJRD}}$	89.6	77.1

7 Conclusions

In this paper, we analyze eating and drinking motions through self-feeding frequency estimation and eating behavior recognition based on the skeleton data from our collected dataset. We introduce MAJRD by computing the moving average on the JRD between joint pairs as features. K-means clustering is applied to the MAJRD between shoulder center and the feeding hand to segment the motion into self-feeding and resting periods. The MAJRDs of all joint pairs from the upper body during the feeding periods are used as features for recognizing different eating and drinking actions. The experiment results show that our proposed method in self-feeding frequency estimation and eating and drinking action recognition performs better than other methods. In the future, we will extend this pioneer work by considering the trajectories of MAJRD in addition to the mean as features to improve the recognition accuracy. We also tend to combine skeleton motions with depth images to examine bits and swallows so that it will conduct a more precise self-feeding action detection and recognition. In the long run, a Web system could be developed to allow people to share and upload their eating and drinking behaviors such that big data analysis can be carried out to facilitate searching and getting feedback.

Acknowledgements The work described in this paper is supported by grants from City University of Hong Kong (Project No. 7004548 and 7004681), National Natural Science Foundation of China (Grant No. 61402205), China Postdoctoral Science Foundation (Grant No. 2015M571688), and Jiangsu University (Grant No. 13JDG085).

References

1. Abdur Rahman, M., Qamar, A.M., Ahmed, M.A., Ataur Rahman, M., Basalamah, S.: Multimedia interactive therapy environment for children having physical disabilities. In: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, pp. 313–314 (2013)
2. Amft, O., Bannach, D., Pirkl, G., Kreil, M., Lukowicz, P.: Towards wearable sensing-based assessment of fluid intake. In: The 8Th IEEE International Conference on Pervasive Computing and Communications Workshops, pp. 298–303 (2010)
3. Bian, Z.P., Chau, L.P., Magnenat-Thalmann, N.: Fall detection based on skeleton extraction. In: Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry, pp. 91–94 (2012)
4. Chang, C.Y., Lange, B., Zhang, M., Koenig, S., Requejo, P., Somboon, N., Sawchuk, A.A., Rizzo, A.A.: Towards pervasive physical rehabilitation using microsoft Kinect. In: The 6Th IEEE International Conference on Pervasive Computing Technologies for Healthcare, pp. 159–162 (2012)
5. Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., Yang, J.: Pfid: Pittsburgh fast-food image dataset. In: The 16Th IEEE International Conference on Image Processing, pp. 289–292 (2009)
6. Dong, Y., Hoover, A., Scisco, J., Muth, E.: A new method for measuring meal intake in humans via automated wrist motion tracking. Appl. Psychophysiol. Biofeedback **37**(3), 205–215 (2012)
7. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1110–1118 (2015)
8. Gao, C., Yang, L., Du, Y., Feng, Z., Liu, J.: From constrained to unconstrained datasets: an evaluation of local action descriptors and fusion strategies for interaction recognition. World Wide Web **19**(2), 265–276 (2016)
9. Hartigan, J.A., Wong, M.A.: Algorithm as 136: a k-means clustering algorithm. J. R. Stat. Soc.: Ser. A: Appl. Stat. **28**(1), 100–108 (1979)
10. Hondori, H.M., Khademi, M., Lopes, C.V.: Monitoring intake gestures using sensor fusion (microsoft Kinect and inertial sensors) for smart home tele-rehab setting. In: The 1St Annual IEEE Healthcare Innovation Conference (2012)

11. Kalantarian, H., Alshurafa, N., Sarrafzadeh, M.: A Wearable Nutrition Monitoring System The 11Th IEEE International Conference on Wearable and Implantable Body Sensor Networks, pp. 75–80 (2014)
12. Kalantarian, H., Alshurafa, N., Le, T., Sarrafzadeh, M.: Monitoring eating habits using a piezoelectric sensor-based necklace. Comput. Biol. Med. **58**, 46–55 (2015)
13. Kawano, Y., Yanai, K.: Real-time mobile food recognition system. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–7 (2013)
14. Li, M., Leung, H.: Graph-based approach for 3d human skeletal action recognition. Pattern Recogn. Lett. **87**, 195–202 (2017)
15. Li, M., Leung, H., Liu, Z., Zhou, L.: 3d human motion retrieval using graph kernels based on adaptive graph construction. Comput. Graph. **54**, 104–112 (2016)
16. Malnick, S.D., Knobler, H.: The medical complications of obesity. J. Assoc. Physicians **99**(9), 565–579 (2006)
17. Mastorakis, G., Makris, D.: Fall detection system using kinect's infrared sensor. J. Real-Time Image Proc. **9**(4), 635–646 (2014)
18. Mettel, M.R., Alekseew, M., Stocklöw, C., Braun, A.: Safety services in smart environments using depth cameras. In: European Conference on Ambient Intelligence, pp. 80–93. Springer, Switzerland (2017)
19. Nguyen, D.T., Cohen, E., Pourhomayoun, M., Alshurafa, N.: Swallownet: Recurrent neural network detects and characterizes eating patterns. In: IEEE International Conference on Pervasive Computing and Communications Workshops, pp. 401–406 (2017)
20. Press, W.H.: Numerical recipes 3rd edition: The art of scientific computing. Cambridge University Press, Cambridge (2007)
21. Qian, R.J., Sezan, M.I., Matthews, K.E.: A robust real-time face tracking algorithm. In: Proceedings of the IEEE International Conference on Image Processing, vol. 1, pp. 131–135. IEEE (1998)
22. Saini, S., Rambli, D.R.A., Sulaiman, S., Zakaria, M.N., Shukri, S.R.M.: A low-cost game framework for a home-based stroke rehabilitation system. In: IEEE International Conference on Computer & Information Science, vol. 1, pp. 55–60 (2012)
23. Shen, Y., Salley, J., Muth, E., Hoover, A.: Assessing the accuracy of a wrist motion tracking method for counting bites across demographic and food variables. IEEE J. Biomed. Health Inform. **21**(3), 599–606 (2017)
24. Stunkard, A., Wolff, H.: A mechanism of satiety-function and disorder in Human obesity. In: Psychosomatic Medicine, vol. 18, pp. 515–515 (1956)
25. Tang, J.K., Leung, H.: Retrieval of logically relevant 3d human motions by adaptive feature selection with graded relevance feedback. Pattern Recogn. Lett. **33**(4), 420–430 (2012)
26. Thomaz, E., Essa, I., Abowd, G.D.: A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1029–1040 (2015)
27. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 588–595 (2014)
28. Wu, W., Yang, J.: Fast food recognition from videos of eating for calorie estimation. In: IEEE International Conference on Multimedia and Expo, pp. 1210–1213 (2009)
29. Yang, S., Chen, M., Pomerleau, D., Sukthankar, R.: Food recognition using statistics of pairwise local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2249–2256 (2010)
30. Zhang, R., Amft, O.: Monitoring chewing and eating in free-living using smart eyeglasses. IEEE J. Biomed. Health Inform. **22**(1), 23–32 (2018)
31. Zhang, S., Ang, M.H., Xiao, W., Tham, C.K.: Detection of activities by wireless sensors for daily life surveillance: eating and drinking. Sensors **9**(3), 1499–1517 (2009)
32. Zhang, Z., Liu, W., Metsis, V., Athitsos, V.: A viewpoint-independent statistical method for fall detection. In: The 21St IEEE International Conference on Pattern Recognition, pp. 3626–3630 (2012)
33. Zong, Z., Nguyen, D.T., Ogunbona, P., Li, W.: On the combination of local texture and global structure for food classification. In: IEEE International Symposium on Multimedia, pp. 204–211 (2010)