Social Action Recognition in Egocentric RGB-D Videos

Ching-Hsiang Hsu, Kung-Hung Lu, Zhi-Qiang Zhong and Chieh-Chih Wang
Department of Computer Science and Information Engineering
Graduate Institute of Networking and Multimedia
National Taiwan University, Taipei, Taiwan
Email: [TomHsu1990,forsetideity,zqzhong1211]@gmail.com, bobwang@ntu.edu.tw

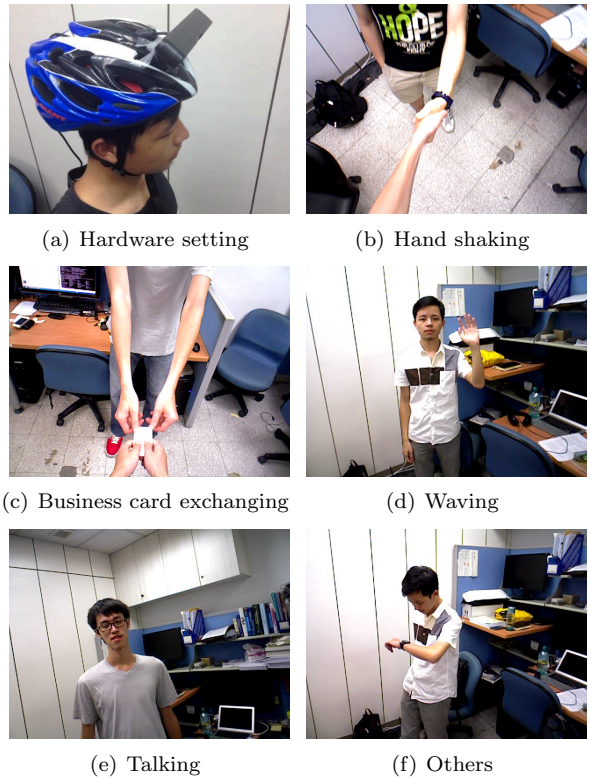
ABSTRACT

In this paper, social action recognition from the first-person point of view in unstructured indoor environments using a RGB-D camera is proposed. While the most of widely used features could be insufficient or incomplete due to the limited perspective view from egocentric vision, we propose dense 3D trajectory features with dimension reduction to enhance the recognition performance. Egocentric RGB-D videos on five common social actions including talking, hand shaking, waving, business card exchanging, and others are collected for training and testing of the proposed approach. Compared with the method using dense 2D trajectory features, our recognition system based on 3D trajectory features achieves 94% accuracy in 65 testing videos.

1. INTRODUCTION

While wearable computing has attracted considerable attention recently, it is believed that egocentric (first-person) vision will be one of the most critical machine perception techniques for enabling wearable computing applications. Although RGB cameras are still the dominant sensing devices for wearable computing, we believe that RGB-D cameras have the potential to fulfill the requests of real world applications in the near future. While there are a number of interesting research and development topics in egocentric vision, we focus on the problems of social action recognition in this paper. Egocentric RGB-D videos on five common social actions including talking, hand shaking, waving, business card exchanging, and others are collected and tested as illustrated in Fig. 1.

One of the most popular features for recognizing actions is to use the outputs from Microsoft Kinect skeleton tracking or similar approaches. Unfortunately, RGB-D skeleton tracking could often fail in egocentric vision due to the limited perspective view of the existing RGB-D cameras as illustrated in Figure 2. In this paper, with the use of the state-of-the-art multi-class recognition library, we pro-



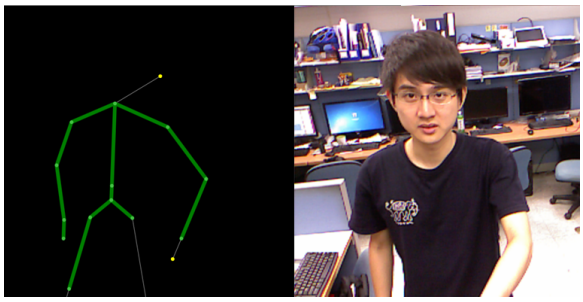
(a) Hardware setting (b) Hand shaking

(c) Business card exchanging (d) Waving

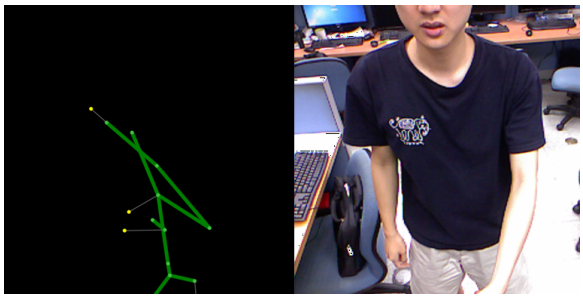
(e) Talking (f) Others

Figure 1: The settings of social action recognition in egocentric RGB-D videos.

pose dense 3D trajectory features with dimension reduction to enhance the social action recognition performance. Compared to the middle-level action information from skeleton tracking, we concentrate on low-level motion information from social interactions to solve the occlusion issue as well as to enhance the overall recognition performance. In the preprocessing stage, the background is removed for obtaining essential motions. The bag-of-words (BoW) model is used for reducing the dimensions of the extracted 3D trajectory features. Multi-classification tree based on the Matlab fit classification tree function is applied to accomplish the learning and classification part of the proposed system. The experimental results on our RGB-D social action datasets show a significant performance enhancement compared to the approach using 2D trajectory features.



(a)



(b)

Figure 2: During a hand shaking scenario from the first-person view of point, the extracted skeleton becomes highly unstable due to that only joints at the upper body are detectable.

The rest of the paper is organized as follows. Section 2 reviews related works and discusses the pros and cons. The proposed approach is described in detail in Section 3. Section 4 presents our datasets and shows the experimental results of social action recognition. Finally, Section 5 gives a conclusion and mention our future work.

2. RELATED WORKS

The problem of social action recognition has attracted much more attention recently. The most recent works use the datasets in the third-person view with only RGB information. The approaches using sparse spatio-temporal features [6] [4] [2] have been widely accepted due to its reliable performance. Accordingly, the hierarchical approach was proposed in which the input data are divided into pyramid subsets to ensure the access of key sub-events [7] [5]. With the use of RGB-D cameras, there are numerous approaches which referred to skeleton information [1] [11]. In the conditions with sufficient perspective view and no occlusion, the skeleton information would be stable and the action recognition performance should be accessible. However, the recognition performance degrades significantly under occlusion as the skeleton estimates could be unstable or incorrect.

One alternative to tackle social action recognition is to recognize postures and gestures of hand and body. During social interactions, hand postures could play a dominant role. In [10], an accurate hand posture recognition system was proposed using the Adaboost learning algorithm with the scale invariant feature transform (SIFT) features. However, it should be unable to fully support scenarios with multiple hands because of the similarity of features. Some works focus on gesture recognition as the interactions between people could be formed by gestures [8]. But these approaches would fail in the situations that the camera is moving or the actions are not depend on gestures such as the talk interaction. In [13], the ego-motion of the user/camera can be robustly estimated by a modified random sample consensus paradigm (RANSAC) which is also applied in the proposed system.

Dense trajectory features have been used for accomplishing action recognition [9]. The trajectories are established to obtain spatio-temporal information of actions, and the quality of the trajectories would influence directly the recognition performance. In the work of Ryoo and Matthies [9],

the spatio-temporal information is utilized by dividing the results of dense optical flow into several classes. The histogram of each class is then used as the features for recognition. Our approach is close to this work which uses dense optical flow results without dimension reduction. It would cause redundant information be used which will not only lead to inefficiency but also lower the performance. To the best of our knowledge, this paper is the first one to recognize first-person social action by dense 3D trajectories with dimension reduction based on RGB-D data.

3. FEATURE EXTRACTION AND DIMENSION REDUCTION

3.1. Dense 3D Trajectory

Derived from [12] which extracts dense 2D trajectories by multi-scale optical flow estimation, feature points are sampled on a dense grid map in multiple scales. First, a dense optical flow estimation is extracted by using the Färneback algorithm [3]. They are tracked by median filtering over the dense optical flow field. To avoid drifting, trajectory which reaches a maximum length of frames will be removed. Then, a new trajectory is initialized from the spatial grid if there is no trajectory within a $W \times W$ pixels grid around it. In addition, static trajectories and trajectories with sudden large displacements are excluded. Besides, points in homogeneous image areas without any structure will not be included in the tracking process. It works well with stationary camera. However, there will be camera motion in most of interactions among people and without ego-motion estimation it cannot handle this situation. So, with assumption that the background occupied most of the RGB-D image, we conduct background removal by using RANSAC based Iterative Closest Point algorithm (ICP).

$$\min_{\mathbf{T}}(\sum_{i}[(\mathbf{T}x_i - y_i) \cdot n_i]^2) \quad (1)$$

where x_i is the point cloud in current frame; y_i is the point cloud in previous frame. n_i is the corresponding normal to the points in x_i. By minimizing Equation 1, we can obtain reliable motion estimation and remove background. Figure 3 (d)-(f) shows the results of background removal. Then, by using these results as the input to get 2D dense trajectories in Figure 3 (g)-(i).

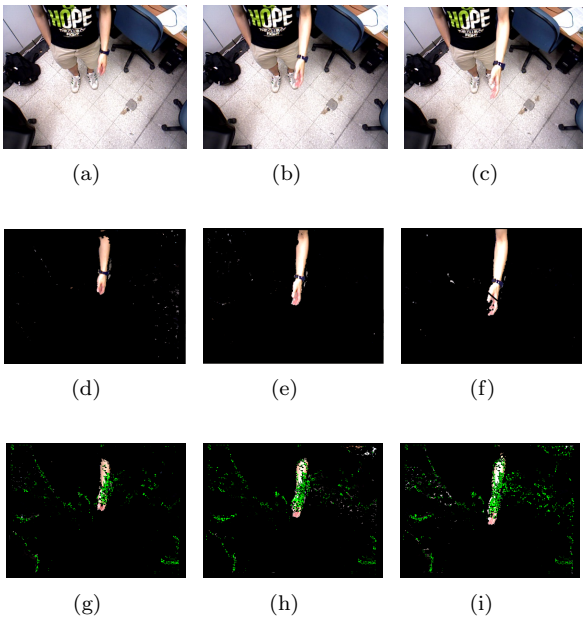


Figure 3: (a)-(c) are raw images. (d)-(f) are images with background removal. In subfigures (g)-(i), by connecting the green dots, they represent the 2D trajectories.

Additionally, we utilize the RGB-D image to get 3D trajectories in Figure 4 by interpolation method. In Figure 5, it illustrates the idea that each 3D trajectory is tracked in its corresponding scales. The cube represents the 3D space and different scale is represented by the different thickness of lines.

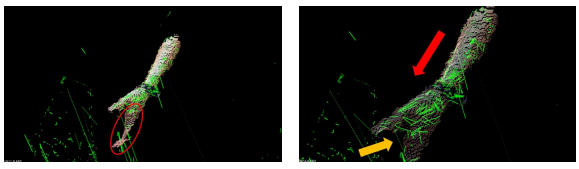


Figure 4: Green lines in (a) and (b) are the 3D trajectory results. In (a), one hand from another person circuits by a red circle tries to reach the hand of first-person. In (b), it shows there are two dominant trajectory directions. The yellow direction is from first person. The red direction is from the other.

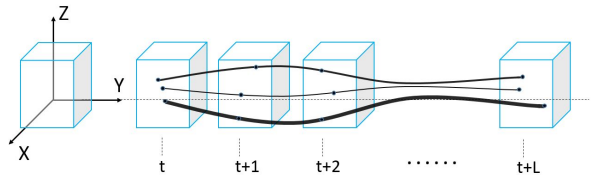


Figure 5: Illustration of 3D Dense Trajectory

3.2. Bag-of-Words

After feature extraction, we get lots of trajectories in each video. However, there are two problems about how to use these massive data. First, there are too many trajectories to deal with. Each video has approximately 90 thousand trajectories, so in order to process all videos, it has total 3 million trajectories needed to be trained at the same time. Second, each video has its own trajectories without knowing the relations among all videos. We overcome this problem by aligning each dimension of data to understand its physical meaning. To solve these two problems, Bag of Words method is adopted to find certain number of significant clusters of trajectories over all data. In Figure 6, it shows that videos in each class own similar distribution of dense 3D trajectories with dimension reduction.

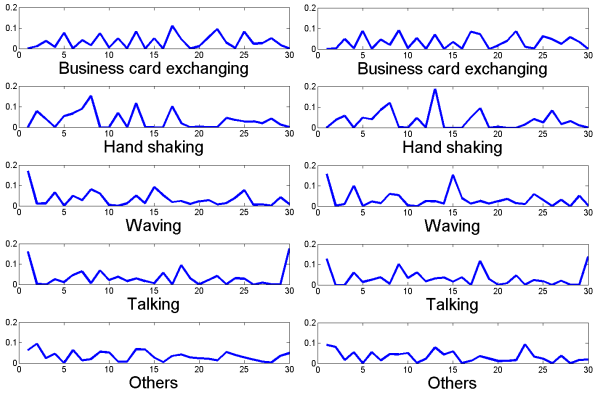


Figure 6: The corresponding class of subfigures in the same row yields similar distribution after data dimension reduction. On the other hand, the different class of subfigures in the column results in distinct distribution. In the subfigures, the X axis means the dimension of data; the Y axis stands for the proportion of each dimension.

All trajectories are clustered into certain number of classes by K-means algorithm. Then, the proportion of each cluster in every videos is treated as the descriptor of training data. Meanwhile, the center of each final trajectory cluster is reserved for future testing process. After finishing doing Bag of Words, all the descriptor of training data are fed to the training module. To achieve multi-classification, classification tree is used as training model by applying the fitctree library in Matlab.

To sum up, integrating all elements mentioned above, in Figure 7, it shows the framework of our approach step by step. After removing background, we calculate dense 2D and 3D trajectories. By applying Bags of Words, the clustered trajectory features are put into multi-classification for training and testing.

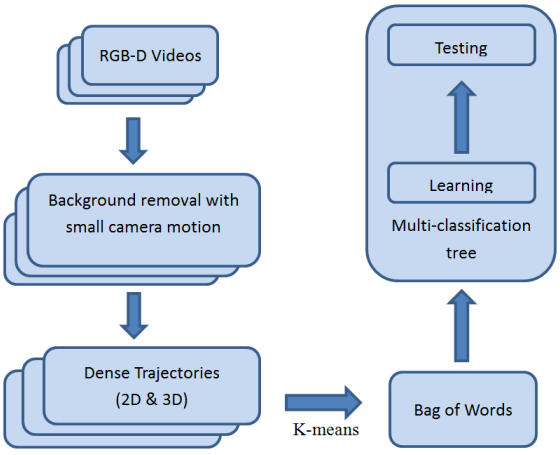


Figure 7: Framework of Proposed Approach

4. EXPERIMENTS

We choose five common social actions in the social interaction scenario, including talking, hand shaking, waving, business card exchanging, and others which includes drinking water, using cell phone, reading books, standing, hanging around, reading the watch ... etc. The RGB-D camera is mounted on a helmet to simulate first-person-view in Figure 1 (a). In Figure 1 (b)-(f), we show five actions which are hand shaking, business card exchanging, waving, talking, and others. Only one of actions in each category is shown. The interaction envi-

ronment is unstructured and there are four testers in our scenario. Each RGB-D data of one kind of class is continuously taken by an ASUS Xtion pro structured light range sensor. Totally, there are 23 RGB-D clipped videos for each class. Then, they are divided into two genres. 10 videos are for training; another 13 videos are for testing.

4.1. Result

The results are compared with four kind of settings of data usage: raw RGB data, raw RGB-D data, RGB data with background removal and RGB-D data with background removal. In each setting, the number of cluster in trajectories is set as 30. For each class, there are 10 videos as training data, so there are totally 50 videos in training dataset. The testing datasets have 13 videos in each class. Once, the trajectories of a testing video has been calculated, each trajectory is compared to the mean values of 30 trajectory clusters from training dataset and choose the one which has minimal distance from testing trajectory. Then, we count the proportions of each class in a video as Bag of Words features. The Table 1 below is the testing results from each setting. The accuracy indicates the proportion of data whose testing result is the same as its label regard to the labelled ground truth. We find that

	(a)	(b)	(c)	(d)
Accuracy	86%	82%	52%	94%

Table 1: (a) RGB data. (b) RGB-D data. (c) RGB data with background removal. (d) RGB-D data with background removal.

the raw RGB data and RGB-D data almost have the same accuracy in the experimental results. And both performances are great. However, considering the data with background removal, the accuracy between RGB-D data and RGB data has a great improvement. From our observation, the reason might be due to the loss of information which after background removal, there are lack of data for representing its property of motion. On the contrary, RGB-D data has more information than RGB data, and depth data can help to understand the motion more accurately, especially when the direction of motion is parallel to the captured axis of camera such as the shaking hands scenario. Therefore, the RGB-D data after removing background would not have insufficient data, and it can also deal with noise. So the performance of RGB-D data without background has promoted generally. For each class, the comparison results are showing below.

	(a)	(b)	(c)	(d)
Business card exchanging	92.3%	76.9%	76.9%	76.9%
Hand shaking	84.6%	76.9%	38.4%	100%
Waving	76.9%	84.6%	15.3%	84.6%
Talking	69.2%	92.3%	53.8%	53.8%
Others	61.5%	15.3%	61.5%	76.9%

Table 2: (a) RGB data. (b) RGB-D data. (c) RGB data with background removal. (d) RGB-D data with background removal.

In this part, there are two major interesting insights. First, the result in the talking class in RGB-D data without background becomes worse since the motion is too small to be recorded by the RGB-D camera because of the resolution. In addition, with background removal, the rest pixels are too few to generate sufficient trajectories. So, the amount of trajectories of the talking class maybe too few to be utilized. Second, the performance of the other class performs poor among all settings except RGB-D without background. The variety of this class may be the main cause for this result. Because the inconsistent data could diverse the distribution of trajectory, the testing data could be clustered to other class easily. And RGB-D data is influenced most since the uncertainty becomes larger as dimension of data grows higher. Videos showing trajectories for image without background shown in Figure 3 are available on the website. [1].

5. CONCLUSION

In conclusion, with apparent actions such as waving and hand shaking, proposed approach can achieve great classification result by using dense 3D trajectory with background removal. However, in the exchanging business cards scenario, RGB data with background gets an outstanding performance. We think it is likely that when exchanging business cards, people will focus on the card without ego-

[1] https://www.youtube.com/watch?v=HL83G9Zmifc

motion. Besides, sometimes, the noise in the scene may cause outlier trajectories. In future work, this problem needs to be improved. Furthermore, while when lack of conspicuous actions or data needed to be captured from a high resolution RGB-D camera, it is arduous to classify actions like talking. To figure out how to solve the problem with RGB-D images. Maybe we can adopt more features or descriptors, for example, Histogram of Oriented Gradients (HOG) or histograms of optical flow (HOF) or take background motion into consideration to recognize the status of the interaction. What's more, since the action is too slight to be exploited, maybe the sound information can be hybrid-used with RGB-D images to overcome the problem.

ACKNOWLEDGEMENT

This work was supported in part by Taiwan Ministry of Science and Technology (National Science Council) and National Taiwan University under Grants MOST(NSC) 102-2221-E-002-179 and 103-2623-E-002-010-D.

REFERENCES

[1] Chen, Hsuan-Sheng and Chen, Hua-Tsung and Chen, Yi-Wen and Lee, Suh-Yin. "Human action recognition using star skeleton" *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks. ACM*, pp. 171-178, 2006.

[2] Dollár, Piotr and Rabaud, Vincent and Cottrell, Garrison and Belongie, Serge. "Behavior recognition via sparse spatio-temporal features" *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2nd Joint IEEE International Workshop*, pp. 65-72, 2005.

[3] Farnebäck, Gunnar. "Two-frame motion estimation based on polynomial expansion" *Image Analysis*, pp. 363-370, 2003.

[4] Qiu, Qiang and Jiang, Zhuolin and Chellappa, Rama. "Sparse dictionary-based representation and recognition of action attributes" *Computer Vision (ICCV)*, pp. 707-714, 2011.

[5] Kovashka, Adriana and Grauman, Kristen. "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition" *Computer Vision and Pattern Recognition (CVPR)*, pp. 2046-2053, 2010.

[6] Schuldt, Christian and Laptev, Ivan and Caputo, Barbara. "Recognizing human actions: a local SVM approach" *Pattern Recognition*, pp. 32-36, 2004.

[7] Sun, Ju and Wu, Xiao and Yan, Shuicheng and Cheong, Loong-Fah and Chua, T-S and Li, Jintao. "Hierarchical spatio-temporal context modeling for action recognition" *Computer Vision and Pattern Recognition(CVPR)*, pp. 2004-2011, 2009.

[8] Ryoo, Michael S and Aggarwal, Jake K. "Recognition of composite human activities through context-free grammar based representation." *Computer Vision and Pattern Recognition(CVPR)*, pp. 1709-1718, 2006.

[9] Ryoo, Michael S and Matthies, Larry. "First-Person Activity Recognition: What Are They Doing to Me?" *Computer Vision and Pattern Recognition (CVPR)*, pp. 2730-2737, 2013.

[10] Wang, Chieh-Chih, and Wang, Ko-Chih. "Hand Posture recognition using Adaboost with SIFT for human robot interaction." *Proceedings of the 13th International Conference on Advanced Robotics (ICAR)*, pp. 317-329, August 2007.

[11] Wang, Jiang and Liu, Zicheng and Wu, Ying and Yuan, Junsong. "Mining actionlet ensemble for action recognition with depth cameras" *Computer Vision and Pattern Recognition (CVPR)*, pp. 1290-1297, 2012.

[12] Wang, Heng and Kläser, Alexander and Schmid, Cordelia and Liu, Cheng-Lin. "Dense trajectories and motion boundary descriptors for action recognition" *International Journal of Computer Vision*, 103(1), pp. 60-79, 2013.

[13] Yang, Shao-Wen and Wang, Chieh-Chih. "Multiple-Model RANSAC for Ego-motion Estimation in Highly Dynamic Environments." *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3531-3538, May 2009.