RECOGNIZING HAND-OBJECT INTERACTIONS IN WEARABLE CAMERA VIDEOS

Tatsuya Ishihara[⋆] *Kris M. Kitani*[†] *Wei-Chiu Ma*[†] *Hironobu Takagi*[⋆] *Chieko Asakawa*[⋆†]

[⋆]IBM Research - Tokyo
[†]The Robotics Institute, Carnegie Mellon University

ABSTRACT

Wearable computing technologies are advancing rapidly and enabling users to easily record daily activities for applications such as life-logging or health monitoring. Recognizing hand and object interactions in these videos will help broaden application domains, but recognizing such interactions automatically remains a difficult task. Activity recognition from the first-person point-of-view is difficult because the video includes constant motion, cluttered backgrounds, and sudden changes of scenery. Recognizing hand-related activities is particularly challenging due to the many temporal and spatial variations induced by hand interactions. We present a novel approach to recognize hand-object interactions by extracting both local motion features representing the subtle movements of the hands and global hand shape features to capture grasp types. We validate our approach on multiple egocentric action datasets and show that state-of-the-art performance can be achieved by considering both local motion and global appearance information.

Index Terms— Wearable cameras, first-person point-of-view, activity recognition

1. INTRODUCTION

Wearable cameras do not restrict the user's activity and can easily record daily activities from a first-person viewpoint. Analysis of these videos has been actively explored for different applications, such as recognizing events [1][2], interactions [3][4], ego-actions [5] and handled objects [6][7]. In this work, we focus on recognizing hand-object interactions in wearable videos. Recognizing hand and object interactions while eating or preparing foods can be useful for monitoring the wearer's diet.

Recognizing what the hands are doing is a challenging task for two primary reasons. The first reason is the difficulty of recognizing actions from the first-person viewpoint. The videos recorded by wearable cameras contain continuous motions caused by the platform's own motions, cluttered backgrounds, and sudden changes of scenery. These characteristics make automatic recognition harder. The second reason is the difficulty of recognizing the activities of the hands. Hand motions have significant temporal and spatial variations, and similar motions and configurations of a hand may be related to completely different activities. For example, grasping an object with two fingers and grasping the object with all fingers share similar movements at the level of some parts of the hand.

Activity recognition from videos recorded in third-person viewpoints has been extensively studied [8], with particular focus given to features based on optical flows such as HOF [9] and MBH [10]. These features can extract local motion features from keypoints, and previous research has shown that these features are effective for recognizing whole-body activities. Wang et al. [11] proposed dense trajectories to effectively sample the keypoints for activity recognition. The advantage of their approach is that it can extract statistically reliable features by densely sampling keypoints. They calculated the motion features at each keypoint by using dense optical flows [12] and tracked the keypoints for fixed time periods to avoid drifting. From each keypoint, the HOF and MBH are extracted to represent the local motion features and the HOG [13] is extracted to represent the local shape features. They also introduced trajectory features that represent the relative movement of each keypoint. In the recent activity recognition challenge [14], Peng et al. [15] showed that dense trajectories gave the best results. Although there have been significant improvements in whole-body activity recognition, hand activity recognition is still challenging because local features do not contain enough information and information about hand configuration is needed to disambiguate similar actions.

To improve the performance of hand activity recognition in the first-person views, past research [16][17] has shown that recognizing handled objects helps to infer hand activities. Complementary to previous work, our proposed work performs hand activity recognition by focusing on the hands. Baradi et al. [18] applied dense trajectories to gesture recognition in the first-person views. They introduced pixel-level hand detection [19] to reduce tracking keypoints and improved both accuracy and performance. Although they demonstrated that dense trajectories can be applied for gesture recognition in first-person view scenarios, they used only local features from dense trajectories.

In this paper, we introduce a novel approach for hand activity recognition in the wearable viewpoint to overcome the difficulties described above. We extract both local features representing the movements of each part of a hand and

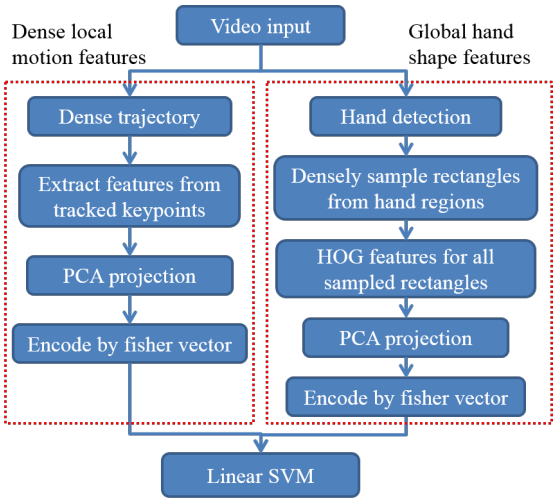

Fig. 1. Proposed method

the global shape features representing the hand's grasp type. These features can be robustly extracted even in a cluttered environment by using pixel-level hand detection. Our experiments showed that the proposed method outperforms the state-of-the-art action recognition methods currently used to recognize hand activities in first-person viewpoint scenarios.

2. COMBINED MOTION AND SHAPE APPROACH

To distinguish hand activities that share similar motions but different hand configurations, we introduce a novel complementary feature approach that consider two different types of features: dense local motion features and global hand shape features. Fig. 1 shows an overview of our approach.

Dense Local Motion Feature: The steps of extracting the local motion features are shown in the left part of Fig. 1. We first use dense trajectories [11] to sample and track the keypoints from the pyramidal images of the input video. The top left image in Fig. 2 shows all of the keypoints sampled with dense trajectories. We extracted HOF/MBH/HOG features from all keypoints because past research has shown that extracting different types of features from all keypoints improved the performance of activity recognition [11][18]. MBH is particularly helpful in our scenario because it is robust against camera motion [10].

Global Hand Shape Feature: The steps of extracting the hand shape features are shown in the right part of Fig. 1. We first detect hands at the pixel level with [19], computing a hand probability value for each pixel on the basis of the color and texture of a local surrounding image patch. The output probabilities are averaged over recent frames to remove noise. The top right figure in Fig. 2 visualizes the estimated hand probabilities. We create the binary image by using thresholding to the hand probability of each pixel and then find the contours of the binary image. We remove any contours that are too small or too large on the basis of the resolution of

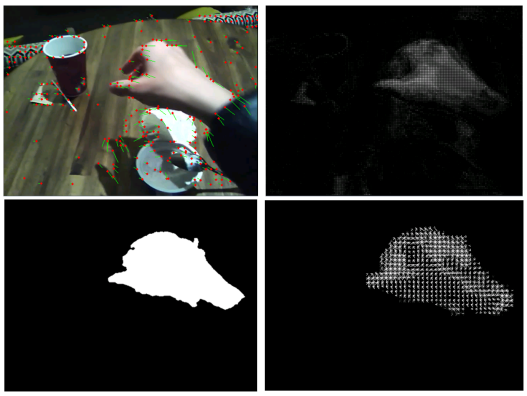

Fig. 2. Top left: Keypoints detected by dense trajectories, Top right: Estimated hand probabilities, Bottom left: Estimated hand regions, Bottom right: HOG features extracted by estimated hand regions

the input videos. The bottom left figure in Fig. 2 shows the contours of hand regions. We calculate shape features only from these estimated hand regions. The bottom right figure in Fig. 2 visualizes HOG features in the estimated hand regions. The detection of hand regions to extract hand shape features is itself a difficult task, and we cannot avoid every false positive hand detection. Therefore we reduce the effects of false positive detections in several steps. First, we densely sample the points only from the estimated hand regions in the pyramidal images, and extract the HOG features from fixed size rectangles whose centers are equal to the sampled points. These HOG features are calculated for all of the pyramidal images. We densely sampled 96×96 rectangles, and the block size, cell size, and number of histogram bins for computing the HOG features were 16×16, 8×8, and 9, respectively. We set these parameters empirically and obtained a 324 dimension HOG feature for each sampled rectangle.

Feature Selection and Encoding: To classify the actions in the input video, we slide the fixed-length time window as we retrieve each new frame from the video and encode all the features extracted within the time window. We use Fisher vectors [20] as the encoding method. For frame t, we extract the dense local motion features m_t from all keypoints and the global hand shape features s_t from all sampled rectangles in the estimated hand regions. These features are high dimensional and may contain redundant information. To avoid overfitting, we apply principal component analysis (PCA) to m_t and s_t. We empirically reduce both of these vectors to 32 dimensional vectors $\hat{m}_t$ and $\hat{s}_t$. Given the fixed length of time window L, we obtain two sets of features $(\hat{m}_t, ..., \hat{m}_{t-L})$ and $(\hat{s}_t, ..., \hat{s}_{t-L})$ for frame t. These two sets of features are encoded into Fisher vectors ϕ_t^m and ϕ_t^s, respectively. The number of gaussian mixtures to encode Fisher vectors is set to 256 and the dimension of each encoded vector is 16,384. By concatenating two encoded vectors as one, we obtain a 32,768 dimensional feature vector φ_t for frame t.

Classifier: Now that we have a frame's features φ_t for frame t, we use one-against-one multi-class SVM classifier [21] to predict its corresponding class label of hand activities. We use a linear kernel function as it works effectively for high dimensional Fisher vectors[15].

3. EXPERIMENT

In this section, we first present a new dataset, the CMU Dining Activity (CMU-DA) dataset, which we believe is the first of its kind. Then, we evaluate the effectiveness of our approach using egocentric videos from four datasets: (1) the CMU-DA dataset, (2) the CMU Multi-Modal Activity (CMU-MMAC) dataset [22], (3) the GTEA dataset [6], and (4) the GTEA Gaze+ dataset [23]. For all of these datasets, we compared the following three baselines which also used linear SVM as classifiers, with our proposed approach:

- Bag of Words (BW): Extract SIFT features [24] and encode with bag of visual words [25]. The number of bag of visual words was set to 1,000.
- Dense Trajectories (DT): Detect keypoints by dense trajectories, then extract features as in [11] and encode with Fisher vectors.
- Reduced Dense Trajectories (RDT): Reduce keypoints in the background by using hand detection as in [18]. The other steps are the same as DT.

3.1. CMU Dining Activity (CMU-DA) Dataset

Although humans can differentiate types of dining activity at a glance, recognizing them accurately still remains a difficult task for a computer, since different hand activities may have similar motions and configurations. To improve this situation, we introduce a new dataset consisting of 167 video clips. The videos are recorded at 4 different locations under varying conditions of a person with a Looxcie® 2. The resolution is 640×480 pixels and the frame rate is 25 fps. This dataset consists of 6 types of dining activity: fork, grab, pinch, spoon, stir and none. Examples are shown in Fig. 3.

3.2. Recognition of Dining Activities

We evaluated the CMU-DA dataset by setting the length of the sliding window to 1 second and the gap of the sliding time window to 0.5 seconds. The keypoints were sampled from 2 levels of pyramidal images by using the dense trajectories. The sampling step size of the keypoints was set to 10 pixels, with the other parameters used to extract the trajectory features and the HOG/HOF/MBH features the same as in [11]. The dimension of the feature vector for each tracked keypoint was 30 for the trajectory features, 96 for the HOG features, 108 for the HOF features, and 192 for the MBH features. We generated 943 video clips, whose length is 1 second, from the original dataset and evaluated the performance with 5-fold cross validation.

Fig. 3. Types of dining activity: Top left: *none* (no dining activities), Top center: *fork* (use a fork), Top right: *grab* (grab a cup), Bottom left: *pinch* (pinch foods), Bottom center: *spoon* (use a spoon), Bottom right: *stir* (stir a coffee)

	BW	DT	RDT	Proposed
none	0.71	0.81	0.79	**0.87**
fork	**0.86**	0.84	0.71	0.85
grab	0.48	0.66	0.57	**0.67**
pinch	0.78	0.63	0.65	**0.97**
spoon	0.59	0.74	0.67	**0.88**
stir	0.60	0.62	0.62	0.62
Accuracy	0.70	0.78	0.73	**0.86**

Table 1. F-measure of each class and average accuracy for CMU-DA dataset

Table 1 shows the F-measure of each class and the last row shows the average accuracy of all classes. As shown, the proposed approach improved the average accuracy compared with the three baselines, and using both hand shape features and dense motion features improved the performance. RDT did not improve the results of DT except for the "*pinch*" class. This result is different from the results of the hand gesture recognition shown in [18], indicating that the hand motion information alone is not sufficient; the motion information of the interacting objects is also important to recognize hand and object interactions.

The confusion matrix for DT and the proposed approach is shown in Fig. 4. The difference between these two approaches is especially clear for the "*pinch*" class. This activity does not contain any large movements at the level of the parts of the hand, and therefore the local motion features were not able to distinguish it from other hand activities. This demonstrates that our global shape features for the estimated hand region are effective for hand activities that are difficult to recognize from only the local motion features.

3.3. Recognition of Cooking Activities

To evaluate the performance under more challenging circumstances, we validate the effectiveness of our approach using three public datasets consisting of cooking activities recorded by wearable cameras.

CMU-MMAC Dataset [22]: This dataset consists of recordings of cooking activities for 5 different recipes per-

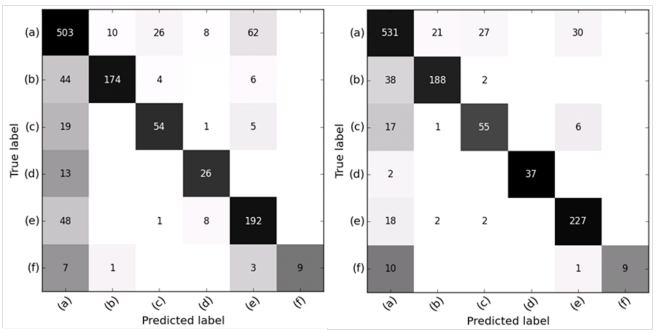

Fig. 4. Confusion matrix for CMU-DA dataset. Left: DT. Right: Proposed ((a) *none*, (b) *fork*, (c) *grab*, (d) *pinch*, (e) *spoon*, (f) *stir*).

formed by 25 subjects. For the videos of *Brownies*, this dataset provides annotations of the actions. We used the data for 12 subjects who made both *Brownies* and *Salad*. For training the hand detection model of each subject, we used the *Salad* videos. The *Brownies* videos were used only for training and testing the hand activity recognition. This dataset contains annotations for 43 verbs representing the wearers' actions, and we selected 8 actions for our evaluation: *no action*, *stir big bowl*, *crack egg*, *pour brownie bag into big bowl*, *open brownie bag*, *pour oil into big bowl*, *pour water into big bowl*, and *pour big bowl into big baking pan*. We selected these actions because they occur frequently and include hand and object interactions. In total, we used 189 video clips for all 8 of the actions. This dataset also includes data for the inertial measurement units (IMU) and extra videos recorded by fixed cameras. We evaluated only the videos recorded by the wearable cameras. We resized all of these videos to 480×360 pixels before processing. The length of the sliding time window was set to 2 seconds and the gap of the sliding time window was set to 0.5 seconds. Other parameters were the same as for the CMU-DA dataset. We evaluated the performance with leave-one-subject-out cross validation.

GTEA dataset [6]: This dataset was collected from 4 subjects and records the cooking activities for 7 recipes. We evaluated this dataset using the settings in [6]. The videos recorded by Subject 2 were used only for testing and all other videos for other subjects were used only for training. The hand detection model for Subject 2 was trained by using the videos of the other subjects, and the hand detection models for the other subjects were trained by the videos recorded by the same subjects. This dataset has annotations for 10 actions (*open*, *close*, *fold*, *pour*, *put*, *scoop*, *shake*, *spread*, *stir*, and *take*), all of which we used for our evaluation. We also added a "*none*" action type for this dataset. In total, we used 1,082 video clips for the 11 classes of action. We resized all of the videos to 720×404 pixels before processing. We evaluated this dataset by setting the length of the sliding time window to 1 second, and the sliding time window was moved for every frame of the 15 fps videos. Other parameters were the same as for the CMU-MMAC dataset.

	BW	DT	RDT	Proposed
CMU MMAC	0.59	0.81	0.68	**0.84**
GTEA	0.39	0.54	0.51	**0.58**
GTEA Gaze+	0.50	0.52	0.36	**0.56**

Table 2. The average accuracy for the cooking activities datasets

GTEA Gaze+ dataset [23]: This dataset was collected from 5 subjects, and records the cooking activities for 7 recipes. From this dataset, we used the videos for the three recipes (*American Breakfast*, *Pizza*, and *Afternoon Snack*) that were recorded for all 5 subjects. The *Afternoon Snack* videos were used only for training the hand detection model for each subject. The *American Breakfast* and *Pizza* videos were used for both training and testing the hand activity recognition. This dataset has annotations for 31 actions for these 3 recipes, and we selected 6 actions for our evaluation (*cut*, *distribute*, *mix*, *move around*, *pour*, and *spread*). We selected frequent actions that include hand and object interactions similar to the CMU-MMAC dataset. We also added a "*none*" action type. In total, we used 437 video clips for all 7 classes of actions. All videos contain eye tracking data. We evaluated this dataset using only the videos recorded by the wearable cameras. We resized all of the videos to 480×360 pixels before processing. Other parameters are the same as for the CMU-MMAC dataset. We evaluated the performance with leave-one-subject-out cross validation.

Table 2 shows the average accuracy for all classes evaluated for the three datasets. For all datasets, our proposed approach improved the average accuracy compared with the baselines. Note that these three datasets include various types of different hand and object interactions, and include different test subjects. The results show that our approach can be applied for a wide range of applications.

4. CONCLUSION

Recognizing hand and object interactions is essential for expanding the first-person view applications. We are working on a novel approach to recognize hand activities. We extract both the local motion features and the shape features of the hands. The local motion features represent the movements of each part of a hand and the shape features represent the hands configurations. Even against a cluttered background, the shape features can be robustly extracted from the hand regions that were estimated from the pixel-level hand detection. Our experiments showed our new approach was better than the state-of-the-art action recognition approaches for different datasets in which the hands interact with objects. We also showed that our approach is especially effective for hand activities that share similar motions at the level of the parts of the hands, which are difficult to recognize based only on the local motion features.

5. REFERENCES

[1] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 1346–1353.

[2] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 2714–2721.

[3] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 1226–1233.

[4] M. S. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 2730–2737.

[5] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 3241–3248.

[6] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 3281–3288.

[7] X. Ren and C. Gu, "Figure-ground segmentation improves handled object recognition in egocentric video," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3137–3144.

[8] J. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, pp. 16, 2011.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.

[10] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision (ECCV)*. Springer, 2006, pp. 428–441.

[11] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 3169–3176.

[12] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian Conference on Image Analysis*. Springer, 2003, pp. 363–370.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005, pp. 886–893.

[14] S. Escalera, M. A. Bautista, M. Madadi, M. Reyes, H. J. Escalante, J. Shotton, and I. Guyon, "Chalearn looking at people challenge 2014: Dataset and results," in *European Conference on Computer Vision Workshop (ECCVW)*, 2014.

[15] X. Peng, L. Wang, Z. Cai, and Y. Qiao, "Action and gesture temporal spotting with super vector representation," in *European Conference on Computer Vision Workshop (ECCVW)*, 2014.

[16] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 407–414.

[17] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2847–2854.

[18] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara, "Gesture recognition in ego-centric videos using dense trajectories and hand segmentation," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2014, pp. 702–707.

[19] C. Li and K. M. Kitani, "Pixel-level hand detection in egocentric videos," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 3570–3577.

[20] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision (ECCV)*. Springer, 2010, pp. 143–156.

[21] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[22] E. H. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2009, pp. 17–24.

[23] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 314–327.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.

[25] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision (ICCV)*. IEEE, 2003, pp. 1470–1477.