First-Person Animal Activity Recognition from Egocentric Videos

Yumi Iwashita Asamichi Takamine
Ryo Kurazume
School of Information Science and Electrical Engineering
Kyushu University, Fukuoka, Japan
yumi@ieee.org

M. S. Ryoo
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA
mryoo@jpl.nasa.gov

***Abstract*—This paper introduces the concept of *first-person animal activity recognition*, the problem of recognizing activities from a view-point of an animal (e.g., a dog). Similar to first-person activity recognition scenarios where humans wear cameras, our approach estimates activities performed by an animal wearing a camera. This enables monitoring and understanding of natural animal behaviors even when there are no people around them. Its applications include automated logging of animal behaviors for medical/biology experiments, monitoring of pets, and investigation of wildlife patterns. In this paper, we construct a new dataset composed of first-person animal videos obtained by mounting a camera on each of the four pet dogs. Our new dataset consists of 10 activities containing a heavy/fair amount of ego-motion. We implemented multiple baseline approaches to recognize activities from such videos while utilizing multiple types of global/local motion features. Animal ego-actions as well as human-animal interactions are recognized with the baseline approaches, and we discuss experimental results.**

I. Introduction

First-person activity recognition (i.e., recognition of activities from egocentric videos) is receiving an increasing amount of attention from computer vision researchers. In the first-person computer vision research, the images/videos obtained are in a viewpoint of a person wearing a camera, and its objective is to analyze objects around the person, understand activities performed by the person, and predict his/her intention. Similar to first-person computer vision works focusing on humans, we propose the new concept of first-person animal computer vision focusing on animals. Particularly, in first-person animal activity recognition, the approach is required to classify activities performed by an animal wearing a camera. This enables automated understanding of what animals are doing, and this can be done regardless of the presence of the humans around them. The concept can be applied to a wide range of animal studies, including wildlife. To the best of our knowledge, this is the first paper such a concept is being introduced.

Different characteristics between human and animal make first-person animal videos unique: (i) moving behavior - biped versus quadruped walk, and (ii) daily activity motion - the more prevalent use of hands by humans (e.g., eating). In addition, since animals, such as dogs and cats have a nature to move dynamically compared with humans, first-person animal videos may contain a huge amount of ego-motion. To better understand the difficulties of first-person animal activity recognition, we provide a dataset composed of first-person animal videos obtained by mounting a camera on each of the four dogs and discuss experiments with baseline approaches.

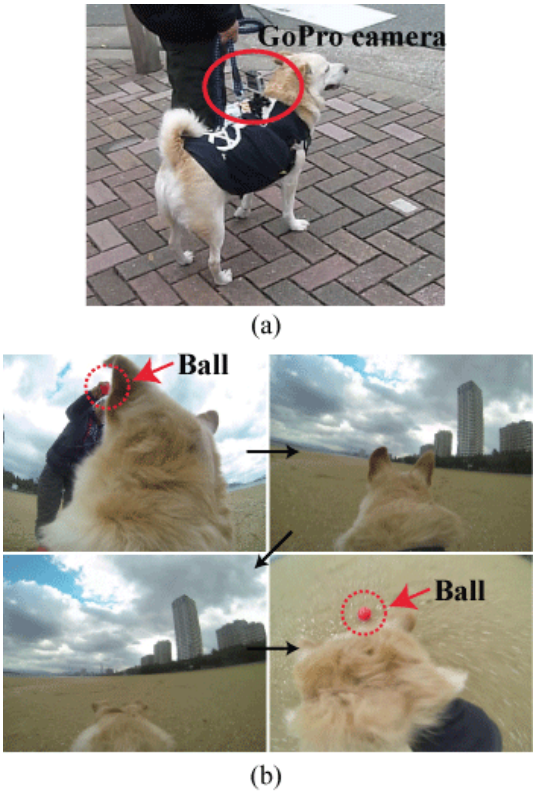


Fig. 1. (a) The setup of the dog, (b) example snapshot images captured while the dog was chasing a ball (in the first frame, the dog covers half of the image, since he stood up).

The dataset contains 10 different types of activities, including activities performed by the dog himself/herself (e.g., running, body shaking, etc), interactions between people and the dog (e.g., petting, feeding, etc), and activities performed by people or cars (e.g., approaching the dog). Figure 1 shows the setup of a dog and example images captured while the dog was chasing a ball. Videos of these three categories of activities tend to display different visual characteristics, implying that multiple types of features are necessary to correctly capture their motion information: (i) global features are necessary for the first type of activities mainly composed of the dog's motion, (ii) both global and local features are useful for the

Fig. 2. Ten classes of activities in our dataset. (a) playing with a ball, (b) waiting for a car to passed by, (c) drinking water, (d) feeding, (e) turning dog's head to the left, (f) turning dog's head to the right, (g) petting, (h) shaking dog's body by himself, (i) sniffing, and (j) walking.

second type composed of the combination of actions by people and the dog, and (iii) local features are suitable for the third type composed of people or car motion.

Multiple baseline approaches are implemented and tested with our new dataset, and we present their results in this paper. In the baseline approaches, five types of global and local features, which are common in first-person activity recognition, are extracted from the first-person animal videos. More specifically, two global features are obtained from dense optical flows [1] and local binary patterns [2], and three sparse spatio-temporal features are extracted as local features, based on a cuboid feature detector [3] and a STIP detector [4]. For a more efficient representation of motion information in videos, we employ the concept of visual words. Finally, for activity recognition, we use SVM classifiers with non-linear kernels.

We emphasize that some of the first-person videos in our dataset display an extreme amount of ego-motion, which is unobservable in previous video datasets. Our dataset is composed of various types of videos with very heavy ego-motion and (almost) without any ego-motion. This makes us believe that our dataset will help general understanding/study of 'egocentric videos' by covering extreme cases. We also believe that our videos may assist development of approaches for first-person recognition of 'human' activities with heavy ego-motion (e.g., sports) by serving as their testbed.

A. *Previous works*

Low-cost high-quality wearable cameras have been available in the market for more than 6 years. Thanks to this, the first-person video analysis have received a lot of attention in the computer vision community. In the first-person vision the study of daily activities are popular topic [5] [6]. Fathi et al. [7] analyzed the cooking activity based on the consistent appearance of objects, hands, and actions. Different from their work, Kitani et al. [8] analyzed sports activities from the first-person video using motion-based histograms. Ryoo et al. [1] recognized interaction-level human activities using local and global motion features. Motivated by the above works focusing on the first-person vision, this paper proposes the concept of the first-person animal vision and the baseline algorithm to recognize activities from first-person animal videos.

Different from a 3rd-person vision, which most of previous works focused on the past decade [9] [10] [11], the first-person vision and the first-person animal vision involve a huge amount of ego-motion such as running and jumping. This results in not only local motion but also global motion in the captured videos. As we mentioned above, activities contains either (i) global motion or local motion, or (ii) both global motion and local motion. In other words, different features are optimal for different types of activity. Thus features from both local motion and global motion should be integrated optimally. For the purpose of combining multiple features which are extracted from the first-person video, Ryoo et al. [1] proposed a method based on a multi-channel version of histogram intersection kernel. Laptev also utilized a multi-channel χ^2 kernel [10] to combine features, which are obtained from a 3rd-person video.

In the baseline approaches for the first-person animal activity recognition, we utilized multi-channel kernels [1] [10]. To our knowledge, our paper is the first paper to recognize activities from an animal's viewpoint.

II. First-person animal video dataset

We construct a new first-person animal video dataset, named 'DogCentric Activity Dataset'. We attached a GoPro camera to the back of each of the four dogs, and Fig. 1 (a) shows an example snapshot of a dog. The four dogs have different owners, and their owners took them on a walk to their familiar walking routes. The walking routes are in various environments, such as residential areas, a park along a river, a sand beach, a field, streets with traffic, etc. Thus even though different dogs do the same activity, their background varies.

The video contains various activities, and we chose 10 activities of interest as our target activities. 'Playing with a ball', 'waiting for a car to passed by', 'drinking water', 'feeding', 'turning dog's head to the left', 'turning dog's head to the right', 'petting', 'shaking dog's body by himself', 'sniffing', and 'walking' are the activities of importance we chose to recognize. Figure 2 shows example snapshots of the activities in our dataset. Figure 3 shows example sequences of frames of 'playing with a ball', 'shaking dog's body by himself', and 'waiting for a car to passed by'. Each activity involves both local motion and global motion. For example in

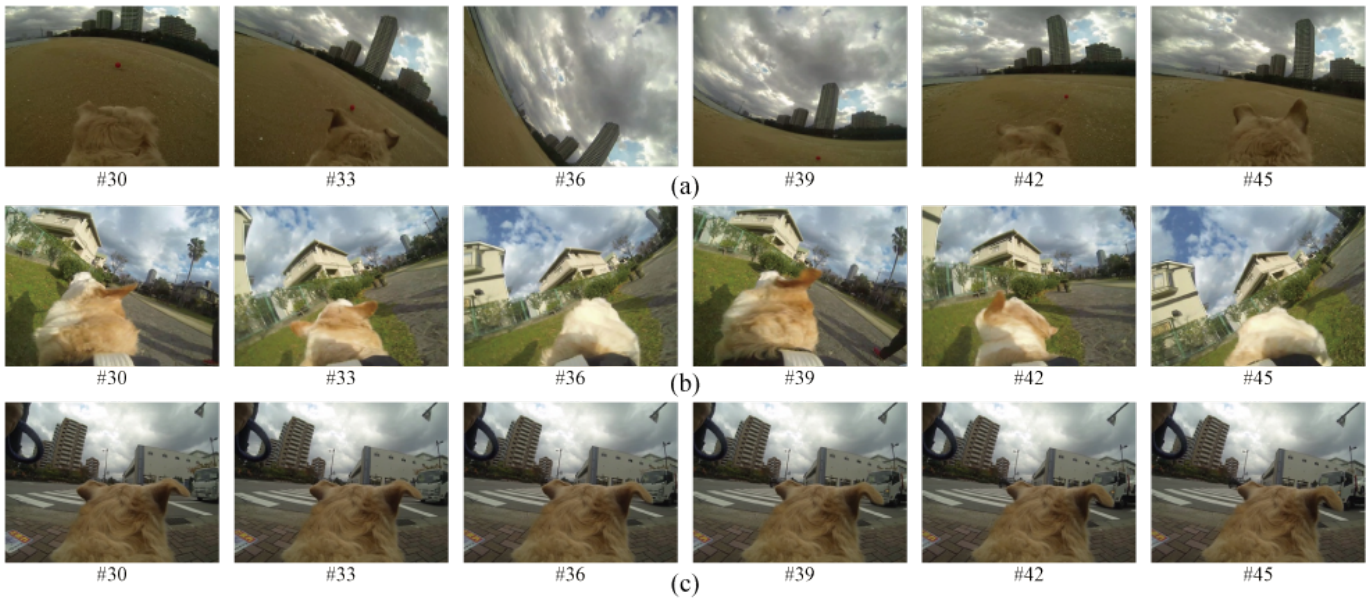


Fig. 3. Example sequences of frames of (a) 'playing with a ball', (b) 'shaking the dog's body by himself' and (c) 'waiting for a car to passed by'.

the category of 'waiting for a car to passed by', the car moving in video is considered as local motion. At the same time the dog also moves his body, which produces global motion. In the category of 'feed', the owner produces moves his hands to give foods to the dog, and at the same time the dog jumped to get his foods. These two motions produces both local motion and global motion.

The videos are in 320 $\times$ 240 image resolution, 48 frames per second. Each continuous video is temporally segmented into multiple videos, so that each video contains a single activity. The number of activities for each dog is shown in Table I, and the total number of video segments in the dataset is 209.

TABLE I. THE NUMBER OF VIDEOS OF ALL ACTIVITIES IN 'DOGCENTRIC ACTIVITY DATASET'

	Dog A	Dog B	Dog C	Dog D	Total (category)
Ball play	6	5	3	0	14
Car	7	1	14	4	26
Drink	5	2	2	1	10
Feed	7	3	8	7	25
Turn head (left)	8	4	3	6	21
Turn head (right)	7	2	4	5	18
Pet	8	4	8	5	25
Body shake	8	2	3	5	18
Sniff	8	7	7	5	27
Walk	7	4	7	7	25
Total (dog)	71	34	59	45	**209**

As shown in Figs. 1 (b), 3 (a), and 3 (b), the dog's motion induces a huge amount of ego-motion. On the other hand, Fig. 3 (c) shows that the amount of ego-motion is relatively small. We quantitatively evaluated the amount of ego-motion displayed in each activity, by estimating rotation angle between frames. Rotation angles are obtained by estimating fundamental matrix between frames, followed by decomposition into rotation matrix, translation vector, and intrinsic parameters. We randomly choose 3 video segments of each activity, and calculated average angle and standard deviation of estimated angles at each activity as shown in Fig. 4 (a). In Fig. 4 (a), the category of 'All' shows average angle and standard deviation of all activities. We also evaluated other two state-of-the-art first-person video datasets and compared them with ours: one is sports activities [8] and the other is JPL-Interaction dataset [1] containing interaction-level activities. The frame rate of the two datasets is 30 Hz, so we interpolate linearly estimated rotation angles into 48 Hz.

The results in Fig. 4 confirm that (1) our new dataset is a good mixture of heavy ego-motion videos and low ego-motion videos, (2) some of our videos display an extreme amount of ego-motion much greater than previous datasets, and (3) motion variance in our videos in general is quite high (i.e., motion is very dynamic) compared to previous datasets. For instance, the results of the sports activity dataset [8] shows that some of its activities have a fair amount of ego-motion (although not as heavy as our 'shaking' and 'ball' activity). However, its variance is rather small. The results of JPL-Interaction dataset shows that 'punching' has heavy ego-motion and high variance, but variance of the other activities are very small (i.e., they are less dynamic).

III. FEATURE EXTRACTION

In this section, we explain motion features we extracted from our first-person animal videos. We utilize a total of five types of features, two global motion descriptors and three local motion descriptors, which are explained in the subsections below.

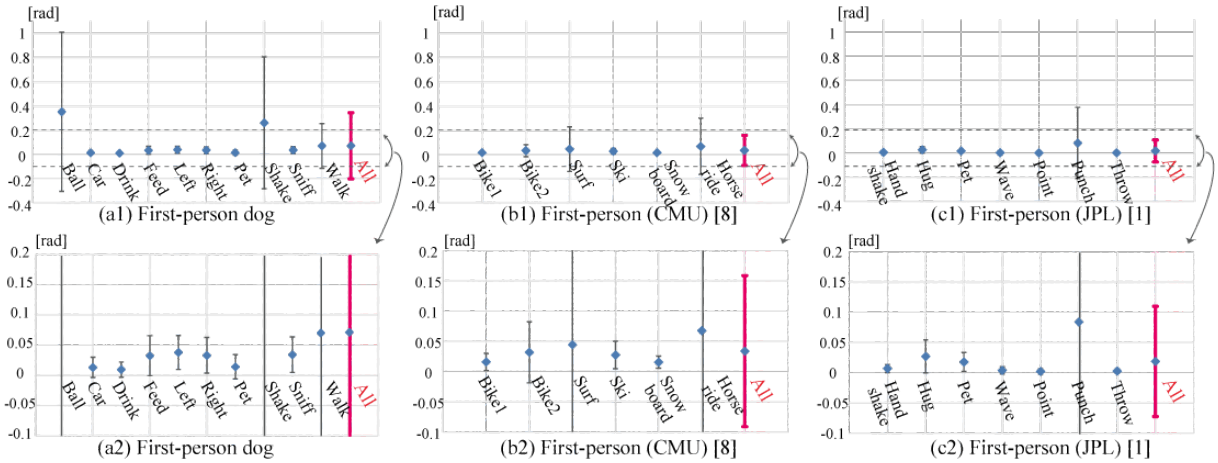


Fig. 4. Ego-motion amount comparison: average value and standard deviations of estimated rotation angles. A higher mean value indicates that the activity contains a greater amount of ego-motion (i.e., heavier). A higher standard deviation indicates that videos of the activity tends to contain dynamic ego-motion, instead of static/monotonic/periodic motion. We are able to observe that ego-motion in our dog dataset is heavier and more dynamic than the previous datasets in general. Particularly, our 'ball play' and 'shaking' show an extreme amount of ego-motion.

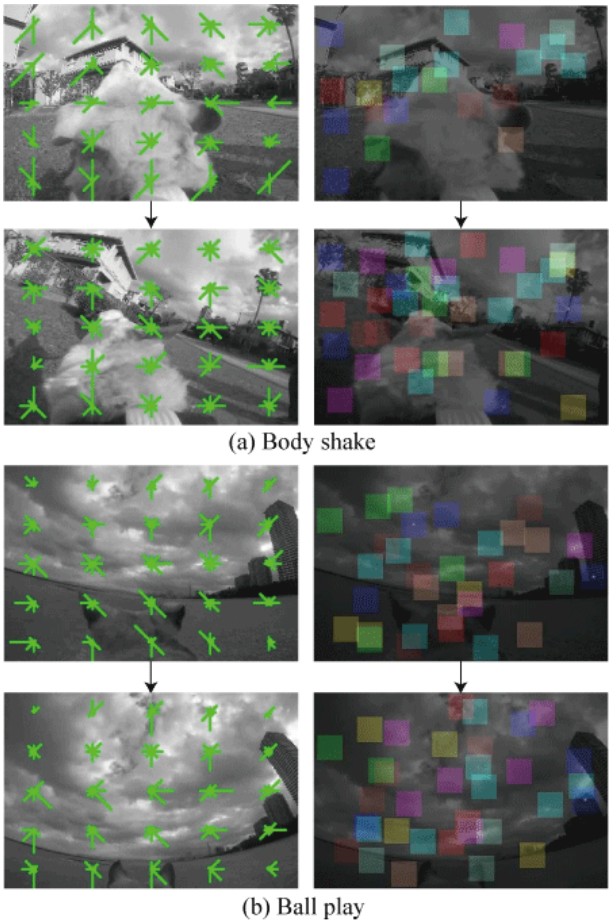
A. *Feature extraction*

In this subsection, we discuss motion features to represent global motion and local motion in first-person animal videos. In the next subsection, we will cluster features to form visual words and obtain histogram representations.

1) Global motion descriptors: Our approach describes global motion in a first-person animal videos using (1) dense optical flows and (2) local binary patterns (LBP).

The global motion descriptor of the optical flows is defined as a histogram of extracted optical flows as described in [1]. Depending on location and direction, the observed optical flows are separated into categories; and the number of flows in each category is counted. As for location, each scene is divided into a grid of s by s (e.g., 3 by 3), and for direction, 8 representative motion directions are considered. Thus, it results in a histogram of optical flow with s-by-s-by-8 bins. The descriptor in each grid is constructed by the sum of optical flows in a given time interval (e.g., 0.2 seconds). The left column on Fig. 5 shows example images of global motion descriptors of the optical flows for two activities 'body shake' and 'ball play'.

The local binary pattern (LBP) [2] is appearance-based features, which showed good performance in analysis and classification of gray scale images. We use the LBP as a global motion descriptor in our method. The LBP is a local transformation that contains the relations between pixel values in a neighborhood of a reference pixel, and in our videos we extracted rotation-invariant LBP [12]. The LBP feature is calculated as a local histogram of quantized local binary patterns, in our case 256 bins at each pixel. The system places each of the computed LBP features into one of the predefined s-by-s-by-256-by-t bins, where we spatially divide an image into s by s grids and t is the number of temporal windows which is explained below. At each grid the LBP features are collected in a fixed time duration (e.g., 0.2 seconds), and this results in s-by-s-by-256 bins. To generate motion feature from LBP features, we concatenate s-by-s-by-256 bins for t times (e.g., t=2). We apply a dimensionality reduction method (the principal component analysis) to compute the LBP-based global motion descriptor having 100 dimensions.



(a) Body shake

(b) Ball play

Fig. 5. (a) Example motion descriptors for 'body shake'. Left column shows global motion descriptor of the optical flows, and right column shows local motion descriptor of the cuboids. Different color shows different cuboids. Note that features are extracted from spatio-temporal patches. (b) example motion descriptors for 'ball play'. Left column shows global motion descriptor of the optical flows, and right column shows local motion descriptor of the cuboids.

2) Local motion descriptors: We extract multiple types of features capturing local motion information in first-person ani-

mal videos and use them as our descriptors. More specifically, sparse 3D XYT spatio-temporal features are extracted. The video is interpreted as a 3D volume of 2D XY frames in sequence along the time dimension T (thus forming a 3D XYT volume). A spatio-temporal feature extractor searches for a set of small XYT patches which contains interest points with salient motion changes inside. We have chosen a cuboid feature detector [3] and a STIP detector [4] as our spatio-temporal feature extractors. As a feature descriptor for cuboids, we use normalized pixel values. As a feature descriptor for STIP, we use histograms of oriented gradients (HOG) and histograms of optical flow (HOF). We apply a dimensionality reduction method to compute our local motion descriptors having 100 dimensions. Figures on right column on Fig. 5 show example images of local motion descriptors of cuboids, for 'body shake' and 'ball play'. In this figure different color shows different types of cuboids.

B. Visual words

For a more efficient representation of motion information in videos, we employ the concept of visual words. We use k-means clustering; each motion descriptor is interpreted as an occurrence of a visual word (one of the w possible) (w=500).

After clustering motion descriptors and obtaining the visual words, each video v_i gets associated a computed histogram, representing its motion. The histogram H_i is a w dimensional vector H_i=$[h_{i1}h_{i2}\dots h_{iw}]$, in which h_{im} is the number of mth visual words identified in the video v_i.

The construction of visual words takes place separately for all local and global motion descriptors. Thus, five histograms are obtained: two histograms are obtained from global motion descriptors (optical flow and LBP) and three are obtained from local motion descriptors (Cuboids, STIP(HOG), and STIP(HOF)). The feature vector x_i is defined as $x_i = [H_i^1; H_i^2; H_i^3; H_i^4; H_i^5]$, in which $H_i^1 \sim H_i^5$ stands for the histograms of the optical flow, LBP, cuboids, STIP(HOG), and STIP(HOF), respectively.

IV. Classification

We use SVM classifiers to recognize first-person animal activities. A kernel $k(x_i, x_j)$ is a function defined to model distance between two vectors x_i and x_j. Learning a classifier (SVMs) with kernel function enables the classifier to estimate better decision boundaries. As we explain in our experiments section, different features are optimal for different types of activity. Thus, utilizing these multiple types of global/local motion features in an efficient way in terms of a non-linear kernel function is extremely crucial for the reliable recognition of activities, and we utilize multi-channel kernels proposed in [1] [10] for combining multiple feature vectors.

V. Experiments

In this section, we implement the baseline approaches and evaluate their performances on our new dataset.

A. Implementation

To obtain visual words, we randomly selected one video segment from each activity and used all selected video segments for k-means clustering. For activity recognition, the selected video segments were removed from the dataset and the rest of video segments were used. We use a repeated random sub-sampling validation to measure the classification accuracy of the baseline approaches. At each round, we randomly selected half video sequences of each activity from our dataset as training dataset and use the rest of sequences for the testing. The mean classification accuracy was obtained by repeating this random training-test splits for 100 times. In addition to two state-of-the art multi-channel kernels [1] [10], we implemented two baseline kernels (linear kernel and RBF kernel). The two multi-channel kernels are a multi-channel χ^2 kernel [10] and a multi-channel histogram intersection kernel [1], which are defined as follows.

$$K(x_i, x_j) = \exp(-\sum_{n=1}^{N} D_n(H_i^n, H_j^n)) \quad (1)$$

where $D_n(H_i^n, H_j^n)$ is the χ^2 kernel [10] is defined as

$$D_n(H_i^n, H_j^n) = \frac{1}{2M_n} \sum_{m=1}^{w} \frac{(h_{im} - h_{jm})^2}{h_{im} + h_{jm}}. \quad (2)$$

Here, M_n is the mean distance between training samples. In [1], $D_n(H_i^n, H_j^n)$ is the histogram distance defined as

$$D_n(H_i^n, H_j^n) = 1 - \frac{\sum_{m=1}^{w} min(h_{im}, h_{jm})}{\sum_{m=1}^{w} max(h_{im}, h_{jm})}. \quad (3)$$

A variance in each kernel was chosen as a value which showed the best performance with training datasets.

B. Evaluation

We first apply the χ^2 kernel to each feature type separately. The motivation is to evaluate the performance of each individual feature type on recognition of activities, and investigate their characteristics. Figures 6 (a) $\sim$ (e) show the confusion matrix of optical flow, LBP, cuboids, STIP(HOG), and STIP(HOF), respectively. The figures show that different features are suitable for different types of activity. For example STIP(HOF) performs better on activities of 'walk' than other features. On the other hand cuboids works good on an activity of 'pet', which the STIP(HOF) performs worse than cuboids. The average classification accuracies of the features were 41.7 % (optical flow), 34.5 % (LBP), 55.3 % (cuboids), 48.6 % (STIP(HOG)), and 51.2 % (STIP(HOF)).

Next, we integrate all five features using a linear kernel and three non-linear kernels (RBF kernel, multi-channel χ^2 kernel [10], and multi-channel histogram intersection kernel [1]). Since all five features have different scale, we normalized these features. Table II shows the average classification accuracies of all kernels. These results suggest that the multi-channel χ^2 kernel successfully integrate global and local motion features, compared with the other three kernels. Figure 6 (f) shows the confusion matrix of all features, and its average classification accuracy was 60.5 %. This result shows that the kernel successfully integrate optimal features for each activity.

VI. Conclusion

In this paper, we provided the dataset composed of first-person animal videos and the baseline algorithms. Experimental results of the baseline algorithms showed different descrip-

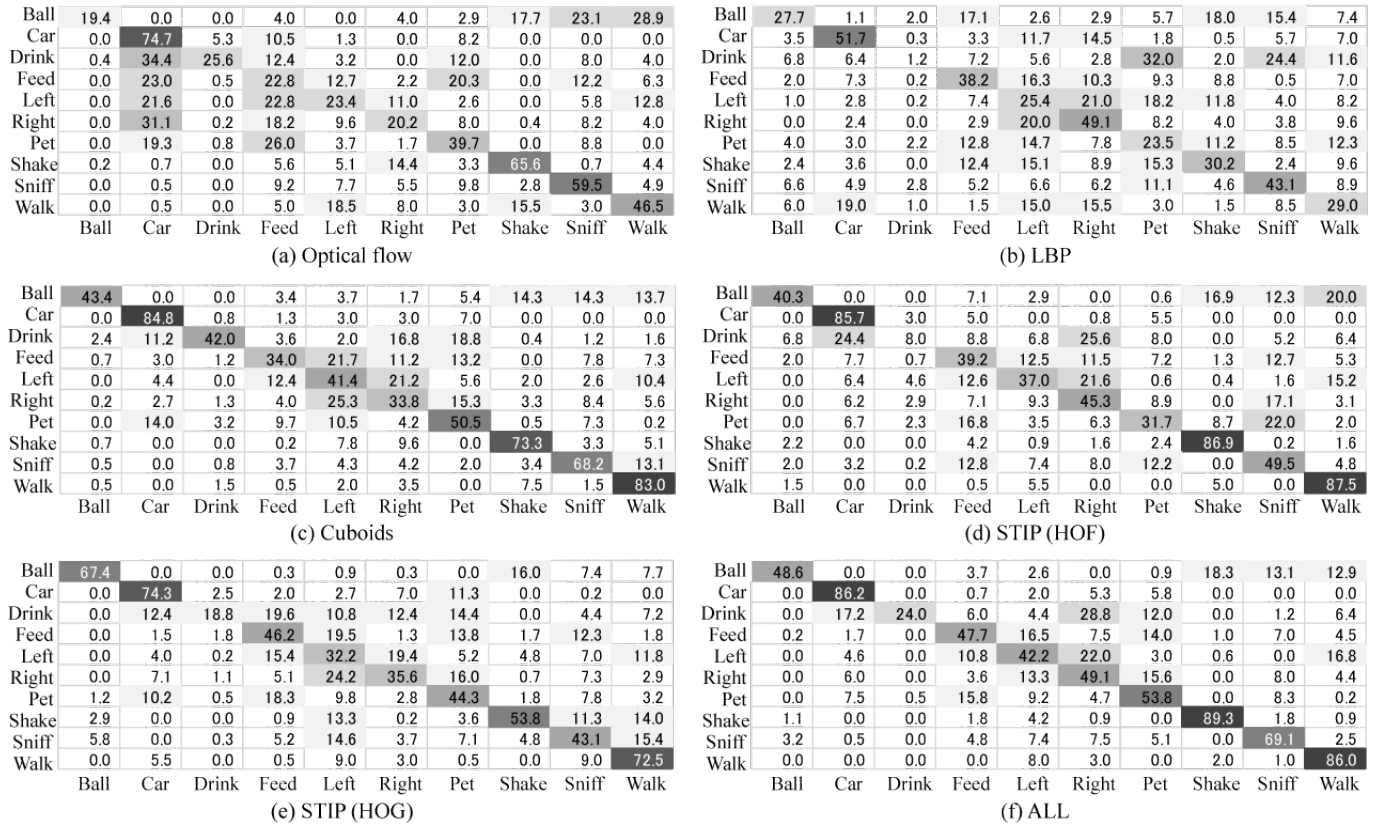


Fig. 6. Confusion matrices of (a) global motion descriptor (optical flow), (b) global motion descriptor (LBP), (c) local motion descriptor (cuboids), (d) local motion descriptor (STIP(HOF)), (e) local motion descriptor (STIP(HOG)), (f) combination of all descriptors.

TABLE II. COMPARISON OF THE CLASSIFICATION ACCURACIES OF LINEAR KERNEL, RBF KERNEL, MULTI-CHANNEL χ^2 KERNEL, AND MULTI-CHANNEL HISTOGRAM INTERSECTION KERNEL.

	Classification accuracy [%]
Linear kernel	52.6
RBF kernel	54.2
Multi-channel χ^2 kernel	**60.5**
Histogram intersection	57.3

tors characterized different activities and the combination of all descriptors achieves a good performance.

The future work includes using multiple cameras on a dog. The dataset was collected with a camera on the dog, and we found out that the position of the camera clearly influences the view. Mounting it on the back has the advantages of seeing for example interactions with people, for example patting the dog. On the other hand, it prevents from seeing exactly in front of the dog, for example what food is the dog eating. In that case a camera mounted on the dog collar, it offers a better view. However that does not allow to see the interaction from above, as is the case with the humans who approach the dog from above. Thus, more than one camera may be needed for a better immersion in the dog environment. Certainly this needs to be as little intrusive as possible. Having more than one GoPro size camera does not appear a good idea, but rather having miniaturized cameras, such as button-size camera.

Acknowledgments The present study was supported by a Grant-in-Aid for Exploratory Research (26630099).

REFERENCES

[1] M. S. Ryoo and L. Matthies, *First-Person Activity Recognition: What Are They Doing to Me?*, In CVPR, 2013.

[2] T. Ojala, M. Pietikainen, and T. Maenpaa, *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns*, IEEE Trans. Pattern Anal. Mach. Intell., 2002.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, *Behavior recognition via sparse spatio-temporal features*, In IEEE Workshop on VS-PETS, 2005.

[4] I. Laptev, *On Space-Time Interest Points*, Int. J. of Computer Vision, Vol.64, No.2-3, pp.107-123, 2005.

[5] Z. Lu and K. Grauman, *Story-Driven Summarization for Egocentric Video*, In CVPR, 2013.

[6] H. Pirsiavash and D. Ramanan, *Detecting activities of daily living in first-person camera views*, In CVPR, 2012.

[7] A. Fathi, A. Farhadi, and J. M. Rehg, *Understanding egocentric activities*, In ICCV, 2011.

[8] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, *Fast unsupervised ego-action learning for first-person sports videos*, In CVPR, 2011.

[9] J. Aggarwal and M. S. Ryoo, *Human activity analysis: A review*, ACM Computing Surveys, vol.43, no.3, 2011.

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, *Learning realistic human actions from movies*, In CVPR, 2008.

[11] J. Niebles, C. Chen, and L. Fei-Fei, *Modeling temporal structure of decomposable motion segments for activity classification*, In ECCV, 2010.

[12] G. Zhao, T. Ahonen, J. Matas, and M. Pietikainen, *Rotation-Invariant Image and Video Description With Local Binary Pattern Features*, IEEE Trans. on Image Processing, 2012.