2015 IEEE/SICE International Symposium on System Integration (SII)
December 11-13, 2015. Meijo University, Nagoya, Japan

# First-Person Activity Recognition with C3D Features from Optical Flow Images

Asamichi Takamine, Yumi Iwashita, and Ryo Kurazume

*Abstract*— **This paper proposes new features extracted from images derived from optical flow, for first-person activity recognition. Features from convolutional neural network (CNN), which is designed for 2D images, attract attention from computer vision researchers due to its powerful discrimination capability, and recently a convolutional neural network for videos, called C3D (Convolutional 3D), was proposed. Generally CNN / C3D features are extracted directly from original images / videos with pre-trained convolutional neural network, since the network was trained with images / videos. In this paper, on the other hand, we propose the use of images derived from optical flow (we call this image as "optical flow image") as input images into the pre-trained neural network, based on the following reasons; (i) optical flow images give dynamic information which is useful for activity recognition, compared with original images, which give only static information, and (ii) the pre-trained network has chance to extract features with reasonable discrimination capability, since the network was trained with huge amount of images from big categories. We carry out experiments with a dataset named "DogCentric Activity Dataset", and we show the effectiveness of the extracted features.**

## I. INTRODUCTION

First-person videos capture various information such as activities of a person wearing a camera (e.g. walking and sitting down), activities of people and objects around him, and activities between the person and people/objects around him. Due to the nature that videos are taken from a viewpoint of the person wearing the camera, first-person videos are suitable for recognizing his activities for the purpose of assisting / supporting people, life-log, and so on.

The study of daily activities from the first-person videos are popular topic [1] [2]. Fathi et al. [3] analyzed the cooking activity based on the consistent appearance of objects, hands, and actions. Different from their work, Kitani et al. [4] analyzed sports activities from the first-person video using motion-based histograms. Most of existing work have focused on first-person vision from human, but different from the existing work, we proposed the concept of "first-person animal vision" [5]. In [5] a dataset named "DogCentric Activity Dataset" was constructed by mounting a camera on each of four pet dogs (Fig. 1), and it consists of 10 different activities including a heavy amount of ego-motion. Compared with public datasets taken from a viewpoint of a person, it is shown that the DogCentric Activity Dataset includes a huge amount of ego-motion. This heavy ego-motion makes the first-person activity recognition more challenging, and the average correct classification ratio with standard techniques for activity recognition was low (60.5 %) in [5].

A. Takamine, Y. Iwashita, and R. Kurazume are with School of Information Science and Electrical Engineering, Kyushu University, Japan takamine@irvs.ait.kyushu-u.ac.jp

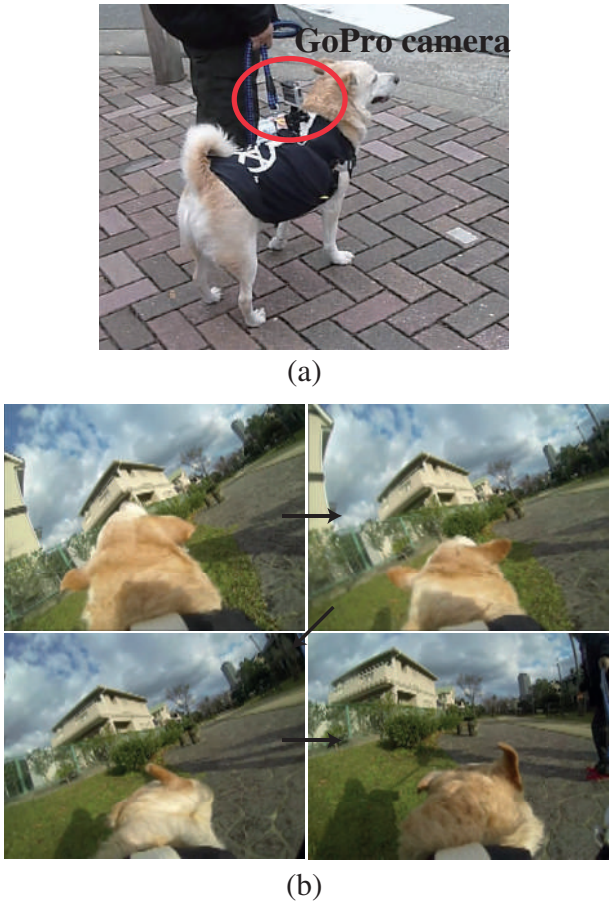



Fig. 1. (a) The setup of the dog, (b) example snapshot images captured while the dog was shaking his body.

Ryoo et al. proposed a method for efficient feature encoding, called "Pooled Time Series (PoT)" [6], and they showed that the combination of newly proposed feature encoding method (PoT) and features from convolutional neural network [8], which is pre-trained with huge amount of images, outperformed the performance in [5]. Features from convolutional neural network (CNN) attract attention from computer vision researchers due to its powerful discrimination capability, but it is designed for 2D images, not videos. Recently a convolutional neural network for videos, called C3D (Convolutional 3D), was proposed [10]. Generally CNN / C3D features are extracted directly from original images /

978-1-4673-7242-8/15/$31.00 ©2015 IEEE

videos with pre-trained convolutional neural network, since the network was trained with images / videos.

In this paper we propose the use of images derived from optical flow (we call this image as "optical flow image") as input images into the pre-trained neural network, based on the following reasons; (i) optical flow images give dynamic information which is useful for activity recognition, compared with original images, which give only static information, and (ii) the pre-trained network has chance to extract features with reasonable discrimination capability, since the network was trained with huge amount of images from big categories. Thus we propose new features, which are extracted from the pre-trained network of CNN and C3D, from optical flow images. We carry out experiments with the DogCentric Activity Dataset, and we show activity recognition results of the extracted features with various feature encoding methods.

### A. Related works

In general there are 3 steps for activity recognition from videos; (i) feature extraction from videos, (ii) feature encoding for more efficient representation of motion information, and (iii) classification with machine learning techniques. Regarding the feature extraction, it can be divided into two categories, global features and local features. Global features are necessary to better understand activities of a person / dog wearing a camera, and local features are suitable to recognize activities of people / objects around the person. In [5], two global features are obtained from dense optical flows [12] and local binary patterns [13], and three sparse spatio-temporal features are extracted as local features, based on a cuboid feature detector [14] and a STIP detector [15]. Wang et al. proposed a better method to extract global feature based on optical flow, called "improved trajectories", and this feature showed good results with various datasets including Hollywood2 [7].

In terms of the feature encoding, other than the bag of visual words, there are more advanced techniques are proposed such as Vector of Locally Aggregated Descriptors (VLAD)[16] and Improved Fisher Vector (IFB)[18]. These techniques encoded feature with histograms which abstract all feature descriptors in one frame or several frames. Ryoo et al. proposed an encoding method called PoT, which basically keeps tracking how each of feature descriptor element changes in a video and summarize them based on pooling technique [6]. PoT showed better performance than VLAD and IFV, and especially the combination of CNN features and PoT showed the best performance (74.5 %).

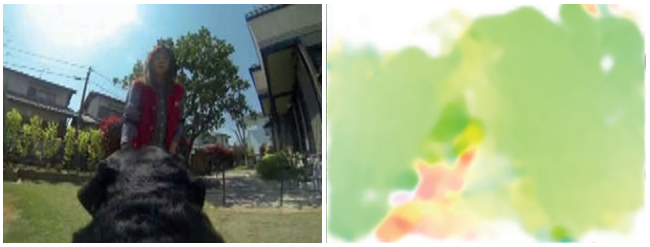
## II. ACTIVITY RECOGNITION

In this section first we briefly describe the way how features are extracted from CNN and C3D, and explain the advantage of features extraction from optical flow images. Next, we abstract the idea of IFV and PoT, followed by the classification.

### A. Convlutional neural network (CNN) and convolutional 3D (C3D)

In this paper we used 'Caffe' for CNN [8] [9], which provides a neural network pre-trained with ImageNet. The neural network consists of 5 convolutional layers, some of which are with pooling layers, and 3 fully-connected layers with a final softmax. In our experiments we used outputs from the 6th layer as features. As we mentioned above, CNN was designed for 2D images, but C3D [10] was designed for videos. C3D was trained on Sports-1M dataset [11], and its network consists of 5 convolutional layers with pooling layer for each, and 2 fully-connected layers. Basically the differences between the network of CNN and that of C3D are that in C3D 3D convolution and 3D pooling are applied. In our experiments we used outputs from the 6th layer as features. For more details about the CNN and C3D, please refer to [8] and [10].

We extract features with the pre-trained neural network from both original and optical images / videos. We calculate dense optical flow with OpenCV library and Fig. 2 shows an example of actual images and its corresponding optical flow image, where hue and saturation values show the direction of each optical flow and its velocity, respectively. Optical flow images give dynamic information, which is useful for activity recognition, compared with original images, which give static information. Thus the combination of features from original and optical flow images (videos) may give more stronger discrimination capability.



(a) (b)

Fig. 2. (a) An example of actual images, and (b) the optical flow image of (a).

### B. Improved Fisher vector and pooled time series

The most common approach to encode features is to use the bag of visual words (BOV) histograms. In BOV each feature descriptor is assigned to its closest visual vocabulary, and each video has histograms which abstract all feature descriptors in one frame or several frames by counting the total number of visual vocabularies. To improve the performance of the BOV, there are several approaches, such as increasing the number of visual vocabularies and using spatial pyramids, which result in high computational cost. The Fisher vector, on the other hand, was proposed to deal with this problem, by applying the Fisher Kernel to image classification [18]. The advantage of the Fisher vector is that it encodes additional information about the distribution of the feature descriptors.

PoT [6] was proposed to focus on capturing changes in feature descriptor elements,, on the other hand, the the BOV-based methods including the IFV and VLAD do not capture changes because of the use of visual vocabularies. The main idea of PoT is to keep tracking how each of descriptor element changes in a video and summarize them based on pooling techniques, which are defined as follows.

$$P_i^{\Sigma} = \Sigma_t f_i(t) \tag{1}$$

$$P_i^{max} = \max_t f_i(t) \tag{2}$$

$$\begin{aligned} P_i^{\Delta^+} &= \|t\| f_i(t) - f_i(t-1) > 0\| \\ P_i^{\Delta^-} &= \|t\| f_i(t) - f_i(t-1) < 0\| \end{aligned} \tag{3}$$

$f_i(t)$ is $i$-th element of feature on $t$-th frame / clip. In experiments with the DogCentric Activity Dataset, PoT outperformed the BOV-based methods.

### C. 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. We utilize a linear kernel for features encoded with IFV method and a histogram intersection kernel[12] for features with PoT.

## III. EXPERIMENTS

In this section we explain the DogCentric Activity Dataset, and we show experimental results with the dataset.

### A. DogCentric Activity Dataset

The video contains various activities, which include 10 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 3 shows example snapshots of the activities in our dataset.

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.

### B. Implementation

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.

### C. Evaluation

In the first experiment, we extracted features with pre-trained neural networks of CNN and C3D from original and optical flow image / videos, and we used the Improved Fisher vector (IFV) for feature encoding. Table I (second row) shows results of each original and optical flow images with CNN features, and Table I (b) shows results of each original and optical flow images with C3D features. From these results, it is shown features extracted from optical flow images show better performance than features from original images, and we can see the advantage of the new features which are based on optical flow images experimentally. Moreover, the combination of original and optical flow images by concatenation show the best performance.

TABLE I

PERFORMANCE COMPARISON BETWEEN FEATURES EXTRACTED FROM ORIGINAL IMAGES, OPTICAL FLOW IMAGES, AND COMBINATION OF ORIGINAL AND OPTICAL FLOW IMAGES [%] (IFV).

| | Original image | Optical flow image | Combination of original and optical flow images |
|---|---|---|---|
| CNN | 57.2 | 63.2 | 65.8 |
| C3D | 57.1 | 60.0 | 68.4 |

In the next experiment, we used the PoT for feature encoding. Table II shows results of CNN features with each PoT pooling technique. From these results, CNN features of optical flow images show better performance than those of original images. The combination of original and optical flow images show the best performance, 66.4 %. The state-of-the-art result of CNN features with PoT was 63.9 % ([6]), and our new features outperformed it.

Table III shows results of C3D features with each PoT pooling technique. Compared with results of CNN features, those of C3D features (Table III) gave much better results, 71.4 %. From this result, we could see that C3D features have stronger discrimination capability compared with CNN features. However, results with C3D features from optical flow images were worse than those from original images, and we believe this is because the C3D network was trained with only sport videos. However, the DogCentric Activity Dataset includes daily life scenes. Thus the C3D could not extract features with enough discrimination capability.

As we mentioned in Section 1, the state-of-the-art result was 74.5 %. This was obtained with features with combination of CNN features and improved trajectory feature (ITF) [7]. Compared with the result of CNN feature in [6], our new C3D features, which are extracted from original and optical flow images, show better result. Thus it has big chance to produce better result (i.e. more than 74.5 %), if we combine new C3D features and ITF. This is left as a future work.

## IV. SUMMARY

In this paper,we proposed new features from optical flow images with CNN and C3D pre-trained networks. We carried out experiments with the DogCentric Activity Dataset, and experimental results showed that the combination of the


Fig. 3. 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.

TABLE II

PERFORMANCE COMPARISON BETWEEN FEATURES EXTRACTED FROM ORIGINAL IMAGES, OPTICAL FLOW IMAGES, AND COMBINATION OF ORIGINAL AND OPTICAL FLOW IMAGES [%] (POT).

| | Original image | Optical flow image | Combination of original and optical flow images |
|---|---|---|---|
| CNN ($P^{\Sigma}$) | 60.6 | 62.3 | 65.3 |
| CNN ($P^{\max}$) | 62.4 | 65.9 | 66.4 |
| CNN ($P^{\Delta}$) | 20.4 | 19.2 | 21.0 |

TABLE III

PERFORMANCE COMPARISON BETWEEN FEATURES EXTRACTED FROM ORIGINAL IMAGES, OPTICAL FLOW IMAGES, AND COMBINATION OF ORIGINAL AND OPTICAL FLOW IMAGES [%] (POT).

| | Original image | Optical flow image | Combination of original and optical flow images |
|---|---|---|---|
| C3D ($P^{\Sigma}$) | 68.2 | 61.3 | 70.6 |
| C3D ($P^{\max}$) | 67.1 | 59.6 | 70.5 |
| C3D ($P^{\Delta}$) | 69.7 | 55.7 | 71.4 |

proposed features from optical flow images and features from original images outperformed those of original images. Our future work includes experiments with the combination of the proposed features and the ITF.

## V. ACKNOWLEDGMENTS

This work is supported by a Grant-in-Aid for Exploratory Research (26630099).

## REFERENCES

[1] Z. Lu and K. Grauman, *Story-Driven Summarization for Egocentric Video*, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2013.
[2] H. Pirsiavash and D. Ramanan, *Detecting activities of daily living in first-person camera views*, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2012.
[3] A. Fathi, A. Farhadi, and J. M. Rehg, *Understanding egocentric activities*, In ICCV, 2011.
[4] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, *Fast unsupervised ego-action learning for first-person sports videos*, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2011.
[5] Y. Iwashita, A. Takamine, R. Kurazume, and M, S. Ryoo, First-person activity recognition from animal videos, Int. Conf. on Pattern Recognition, pp. 4310-4315, 2014.
[6] M. S. Ryoo, B. Rothrock, L, Matthies, *Pooled Motion Features for First-Person Videos*, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015.
[7] H. Wang and C. Schmid, *Action Recognition with Improved Trajectories*, Int. Conf. on Computer Vision (ICCV) 2013.
[8] A. Krizhevsky, I. Sutskever, and G. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, Advances in Neural Information Processing Systems 25, 2012.
[9] *http://caffe.berkeleyvision.org/*
[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, *Learning Spatiotemporal Features with 3D Convolutional Networks*, Int. Conf. on Computer Vision (ICCV) 2015.
[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, *Large-scale video classification with convolutional neural networks*, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014.
[12] M. S. Ryoo and L. Matthies, *First-Person Activity Recognition: What Are They Doing to Me?*, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2013.
[13] 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.
[14] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, *Behavior recognition via sparse spatio-temporal features*, In IEEE Workshop on VS-PETS, 2005.
[15] I. Laptev, *On Space-Time Interest Points*, Int. J. of Computer Vision, Vol.64, No.2-3, pp.107-123, 2005.
[16] H. Jegou, M. Douze, C. Schmid, P. Perez, *Aggregating local descriptors into a compact image representation*, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp.3304-3311, 2010.
[17] F. Perronnin, C. Dance, *Fisher kernels on visual vocabularies for image categorization*, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp.1-8, 2007.
[18] F. Perronnin, J. Sanchez, T. Mensink, *Improving the fisher kernels for large-scale image classification*, European Computer Vision Conference (ECCV), pp.143-156, 2010.