Dog-Centric Activity Recognition by Integrating Appearance, Motion and Sound

Tsuyohito Araki† Ryunosuke Hamada‡ Kazunori Ohno‡ Keiji Yanai†
†The University of Electro-Communications, Tokyo, Japan
‡NICHe, Tohoku University, Japan

Abstract

In this work, we analyze the ego-centric videos of "rescue dogs" and recognize dog activities by using not only by videos but also sounds. To do that, we propose a three-stream-based action recognition network. As a result, the proposed method which integrates appearance, motion and sound achieved the highest accuracy, 51.8%. This result is relatively high as a recognition result of ego-centric video recognition.

1. Introduction

A dog which assists rescue activity in the scene of disasters such as earthquakes and landslides is called a "disaster rescue dog" or just a "rescue-dog". In rescue activities in the disaster areas, trained rescue-dogs may conduct exploration as human assistants. The rescue-dog makes a pair with a human and investigates disaster areas by making use of special characteristics as a dog. The big problem on a joint rescue activity by a handler and a rescue-dog is lack of information on the surrounding environment and victims for triage [1].

Figure 1. A rescue dog wearing a "cyber-rescue suit" [7].

For this situation, in Japan where earthquakes happen frequently, a research project on "Cyber-Rescue" is being organized for more efficient rescue activities. In the project, to analysis the activities of rescue dogs in the scene of a disaster, "Cyber-Dog Suits" equipped with sensors, a camera and a GPS were developed [7] (Figure 1).

In this study, we aim to estimate the behavior of rescue-dogs using sensor data obtained from the cyber-dog suits. We analyze ego-centric videos taken by a camera on the cyber-dog suits and recognize dog activities by using not only by videos but also sounds. This is expected to make it possible to determine automatically what the rescue-dog is doing now. Information necessary for triage is organized and disaster rescue activities will be more efficient by the rescue-dog investigation.

To analyze ego-centric videos with audio, we propose an image/sound-based three-stream CNN for dog activity recognition which integrates sound as well as motion and appearance. We conducted some experiments for multi-label activitity categorization using the proposed method. As a result, the proposed method achieved the highest accuracy, 51.8%. This result is relatively high as a recognition result of ego-centric video recognition.

2. Related Work

Third-person activity recognition: Two-stream CNN is a method for video classification [10]. It classifies video categories by integrating motion information represented as optical flows and appearance information. There are many kinds of the researches on derived networks based on the Two-stream CNN. Convolutional Two-Stream Network Fusion [4] is one of the variant methods.

First-person activity recognition: First person vision has been actively studied so far. There are so many researches such as [8][5]. Minghuang*et al*. [8] proposed a twin stream network that integrates hand segmentation, target object localization and motion.

Dog-Centric activity modeling: There are a few studies on first person video analysis from dog's view. Ehsan *et al*. estimated dog activity from dog-centric video [3]. They modeled dog activity and estimate how dogs will move.

[1] "Triage" means the process of deciding who receives medical treatment first, according to how seriously someone is injured.

They used only a video in the experiments, and did not use multi-modal information such as sound. Iwashita *et al*. also published a Dog-Centric Activity Dataset (DCAD) [6]. This dataset is used for dog behavior classification from dog-centric movies.

Sound Recognition: As a study of movie classification, "SoundNet" has been proposed [1]. Semantic information of audio has been shown to be important for movie recognition, though we do not aim classification from the sound only. In our work, we build the architecture in which an appearance/motion two-stream network is connected with a deep audio network.

3. Dataset

The rescue-dog training dataset consists of a group of data collected by the sensors embedded in the cyber-rescue suits. The dataset is still growing, and we are collecting data at the time of rescue training of rescue-dogs in the simulated disaster sites. Due to privacy and ethical issue, it does not contain the data recorded in the actual disaster sites. It consists of about 2 minutes to 20 minutes of seven videos with audio. The total time is 57 minutes and 40 seconds, the number of frames per second is 29.97fps, and the total number of frames is 103,696. The videos are annotated by specifying a time range for each of 11 activity classes. Multiple classes are sometimes overlapped with each other at the same time. Therefore, we treat this task as a multi-label classification task.

3.1. 11 Dog Activity Classes

We explain each of the 11 classes of rescue dog activities. Their frequencies in the dataset are not uniform as shown in Table 1. **bark:** Typically a rescue dog barks when finding out victims. **cling:** The situation in which the dog has an obsession with something smells. It is more detailed than "*sniff*," and this is labeled it always overlaps with "*sniff*." **command:** The situation in which the dog is being instructed by the handler. **eat-drink:** The situation in which the dog is eating or drinking something. **look at handler:** The situation in which the dog is looking at the handler. **run :** There is a feeling of floating on the screen compared to the *walk-trot* class, and it usually contains intensive shaking and sounds. **see victim:** The situation in which the victim is appearing in the camera. **shake:** The situation in which the dog is vigorously waving. During this activity, sound clatters on the camera on the dog's back. **sniff:** This can be the indicator that measures the dog's motivation for exploration. **stop:** The situation in which the dog is not stepping and stays in the same spot, which includes stepping on the same spot. **walk-trot:** Walking, not running. The order of footing is different from "*run*" class. As examples, the scenes of "*see victim*" and "*stop*" are shown in

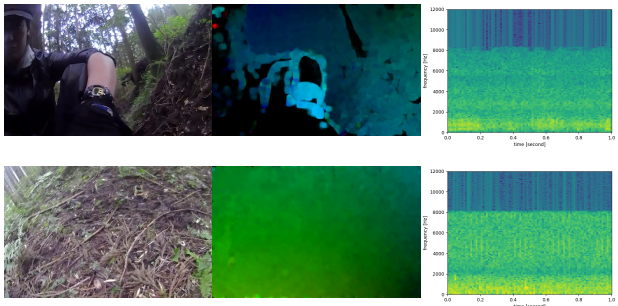

Figure 2. Examples of the ego-centric rescue-dog video dataset, which shows"*see victim*" class (upper) and "*stop*" class (lower). From the left, RGB images, optical flow images, and the sound images visualized by the MFCC spectrogram.

Figure 2.

4. Method

In this study, we perform multi-label estimation of rescue-dog activities from ego-centric dog videos and audio. Our method is based on Convolutional Two-Stream Network Fusion [4] and SoundNet [1].

The proposed network which takes two images (appearance RGB images and optical flow images) and sounds as inputs is called the image/sound-based three-stream network. The detail is shown in Figure 3.

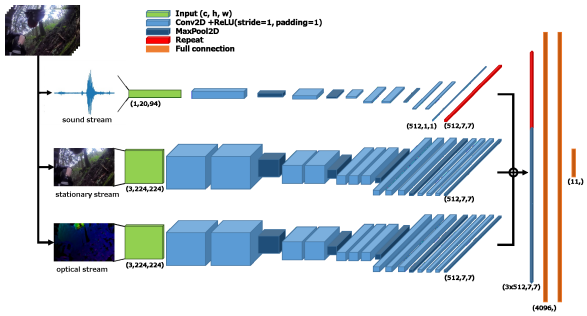

Figure 3. Proposed method architecture of image/sound-based three-stream CNN.

4.1. The detail of the image/sound-based three-stream CNN

The proposed image/sound-based three-stream CNN is a network for activity estimation by integrating appearance, motion and sound.

The window size which is the unit size for activity classification is 31 frames, and a RGB image of the center frame and the corresponding optical flow image immediately after a RGB image are taken out. We fine-tune the ImageNet [2] pretrained VGG16 [9] model for both the RGB and optical flow streams. Regarding audio information, we extract

Table 1. The frequencies of the 11 dog activity classes in the rescue-dog dataset.

classes	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk
num of frames	1744	1127	2439	343	2011	98	1549	239	7719	6384	8764

MFCC spectrogram from the 31 frames which corresponds to 0.5 seconds before and after the center frames of the window, and provide it to the sound stream. The sound stream consists of 2-d convolution, since MFCC features has the shape of feature maps. This is different from SoundNet [1] consisting of 1-d convolutions which takes raw sound waves as an input. The three outputs obtained from each stream are simply concatenated in the direction of channel, and it is provided to three FC layers.

5. Experiments

We made the experiments to show the effectiveness of the proposed three-stream network. We explain eight types of experiments in this section. We compared the performance of the proposed three-stream network with that of the one/two-stream networks as ablation studies. Table 2 shows eight kinds of the networks and the streams each of them used.

Table 2. Modality table including the proposed method. The sound stream experiments have two types convolutional layer of the conv2D and conv1D.

	stationary	optical flow	sound
(1)	✓	×	×
(2)	×	✓	×
(3)	×	×	1D
(4)	×	×	2D
(5)	✓	✓	×
(6)	✓	×	2D
(7)	×	✓	2D
(8)	✓	✓	2D

We used the first half 70% of each video for training, and the latter half 30% for evaluation.

In all the tables showing experimental results, the accuracy for each class and the total accuracy are represented by Precision, Recall, and Jaccard coefficient. Note that the Jaccard coefficient is represented by

$$\frac{TP}{FP+FN+TP}$$

and a more rigorous value can be obtained compared to the F scale. We used this coefficient to emphasize both Precision and Recall in the rescue-dog's activity estimation.

Multi-label estimation from only RGB images:

The estimated accuracy is shown in Table 3 (1). *Eat* class and the *run* class show particularly low precision because of the small numbers of training samples. Table 3 show strong relationship between the number of samples and the accuracy. Exceptionally, *sniff* class and *command* class show zero scores, which indicates that there exist some classes which are difficult to recognize from the RGB images.

From optical flow images:

The estimated accuracy is shown in Table 3 (2). It can be seen that the estimation is difficult because the appearance features are lost.

From only the sound data:

The network was built with reference to the audio classification network of [1]. We use MFCC to extract features from the sound data, and the output is used as input to our sound network.

1D-Convolutional Network:

Overall, the accuracy is better than the RGB and optical flow images. In particular, the accuracy increased on the *bark* class, the *command* class, the *shake* class, and the *sniff* remarkably, and it was shown that the sound feature was important for estimation of some classes.

2D-Convolutional network:

Compared with the 1D-Convolutional network, the output of feature quantities increased, but the overall accuracy did not differ much.

From the RGB image and the optical flow image:

In multi-label estimation from the RGB image with the optical flow images, we trained non-pre-trained VGG-16 models and estimated from the result of combining the two outputs. The estimated accuracy is shown in Table 3 (5). The *sniff* class, in particular, has dramatically improved regarding accuracy. It is considered that training can be performed using motion features obtained from optical flow in addition to image features of the RGB image. In contrast, some classes such as the *shake* can not be classified at all as RGB images, and optical flow images are adversely affected by each other as well.

From the RGB image with audio data:

Only when comparing accuracy by class is the accuracy of the *victim* class is better. In the *see victim* class, in addition to the sound of dog's barking, the victims are frequently shown on the camera. It can be inferred that the accuracy has increased because sound data and the RGB image compensate each other for the lack of information. Overall, the result of using sound data only is better.

The classes with better accuracy in using the RGB image only are *cling* and *eat* classes. These class features tend to appear in images, and sound data is not important for estimation for humans. The classes with better accuracy in

Table 3. Comparison of each experiment result by the Jaccard coefficient.

	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk	all
num. of sample frames	1744	1127	2439	343	2011	98	1549	239	7719	6384	8764	103696
(1) RGB images	0.244	0.066	0.0	0.024	0.057	0.0	0.204	0.0	0.0	0.588	0.51	0.436
(2) optical flow	0.141	0.0	0.0	0.0	0.017	0.0	0.017	0.0	0.0	0.586	0.476	0.406
(3) sound (Conv1D)	**0.669**	0.078	0.22	0.023	0.138	0.0	0.274	**0.44**	0.502	0.745	0.704	0.512
(4) sound (Conv2D)	0.563	0.04	0.188	0.001	0.059	0.0	0.201	0.304	0.524	0.744	**0.74**	0.512
(5) RGB+optical	0.11	0.018	0.043	0.0	0.155	0.0	0.259	0.0	0.426	0.705	0.668	0.435
(6) RGB+sound	0.662	0.031	0.195	0.018	0.115	0.002	0.308	0.402	0.498	0.726	0.694	0.5
(7) optical flow+sound	0.667	0.054	**0.234**	0.014	0.123	0.01	0.223	0.356	0.487	0.759	0.692	0.493
(8) RGB+optical+sound	0.577	**0.135**	0.186	**0.066**	**0.183**	**0.026**	**0.433**	0.409	**0.53**	**0.779**	0.725	**0.518**

using sound data only are *stop* and *walk* classes. Although these three classes are easy for human to estimate, they are somewhat difficult to understand from only RGB images.

From the flow images with audio data:

The class with significantly improved accuracy is the *command* and *stop* classes. With that, the data set basically has a walk or stop class and compare Precision and Recall of the *walk* and *stop* class, we can say that the increase in the number of *stop* classes caused a decrease in the *walk* class in the number of detection. While the *stop* class is characteristic, the *walk* class has various patterns. Therefore, it is considered that learning the *stop* class could not lead to learning in the *walk* class. The *see victim* class is less accurate for the same reasons as the *stop* class, compared to the sound data with the RGB image.

Image/sound-based three-stream CNN:

The accuracy is improved compared to sound data only and image/sound-based two-stream. The result was obtained high accuracy initially expected with from compensating each other's missing information. Unlike in the case of two inputs, credible results were respected, and there were fewer cases where the legs were pulled, leading to an overall increase accuracy.

6. Conclusions

We proposed an image/sound-based three-stream CNN, and estimated the rescue-dog's behavior using the proposed network. As a result, we obtained the accuracy, 51.8%. Ablation studies were performed on three inputs and compared to the proposed method. From the results, it was shown that although audio data is powerful for class estimation, necessary information is included in each of three data of sound, RGB image, and optical flow image.

The purpose of this study is the life-saving task thorough rescue-dog's behavior estimation. The result was still insufficient for making decisions on the real field.

Future work: There is much room for ingenuity in feature extraction from dog-centric videos. We should consider how to extract more semantic features from the three or more information. For example, as with the arm segmentation network used in the human first-person viewpoint classification method [8]. Enrichment of training data is also one of the most important issues. For example, when adopting the segmentation method, it is necessary to prepare a new dataset. The amounts of the rescue-dog training dataset is still not enough. In particular, the *eat*, *shake*, and *run* classes have very little data. Making the expansion dataset for rescue-dogs is one of the most important issues.

References

[1] Y. Aytar, C. Vondrick, and A. A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2009.

[3] K. Ehsani, H. Bagherinezhad, J. Redmon, R. Mottaghi, and A. Farhadi. Who let the dogs out? modeling dog behavior from visual data. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2018.

[4] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2016.

[5] B. Gedas, S. P. Stella, X. Y.and Hyun, and S. Jianbo. Am I a baller? basketball skill assessment using first-person cameras. In *Proc.of IEEE International Conference on Computer Vision*, 2016.

[6] Y. Iwashita, A. Takamine, R. Kurazume, and M. S. Ryoo. First-person animal activity recognition from egocentric videos. In *Proc. of International Conference on Pattern Recognition (ICPR)*, 2014.

[7] Y. Komori, T. Fujieda, K. Ohno, T. Suzuki, and S. Tadokoro. Detection of continuous barking actions from search and rescue dogs' activities data. In *Proc.of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 630–635, 2015.

[8] M. Minghuang, F. Haoqi, and M. K. Kris. Going deeper into first-person activity recognition. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2016.

[9] K. Simonyan, A. Vedaldi, and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of International Conference on Learning Representations*, 2015.

[10] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.