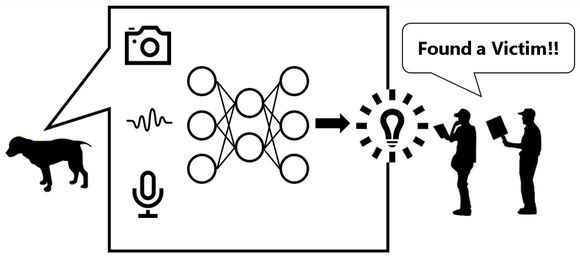
Dog-Centric Activity Recognition by Integrating Appearance, Motion and Sound

Tsuyohito Araki (Univ. Electro-Comm., Tokyo), Ryunosuke Hamada (Tohoku University), Kazunori Ohno (Tohoku University), Keiji Yanai (Univ. Electro-Comm., Tokyo)

Background & Objective

Rescue Dog

・A dog which assists rescue activity in the scene of natural disasters
・We aim to identify the behavior of the rescue-dog from the sensor attached to the dog.



Dataset

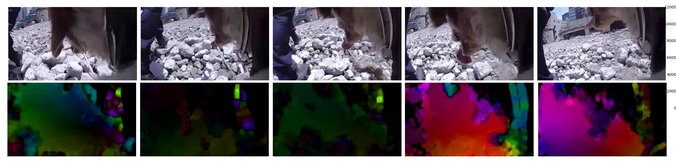
Rescue Dog suit

・Wearable measurement recording device of rescue-dog behavior, developed by Ohno et al. Of Tohoku University

RescueDog training dataset

・**Data of rescue training**
Seven ego-centric videos (2min to 20min),
Total 57min 40sec,
29.97FPS,

Down-sampled to 5 FPS, and used 103,696 frames.

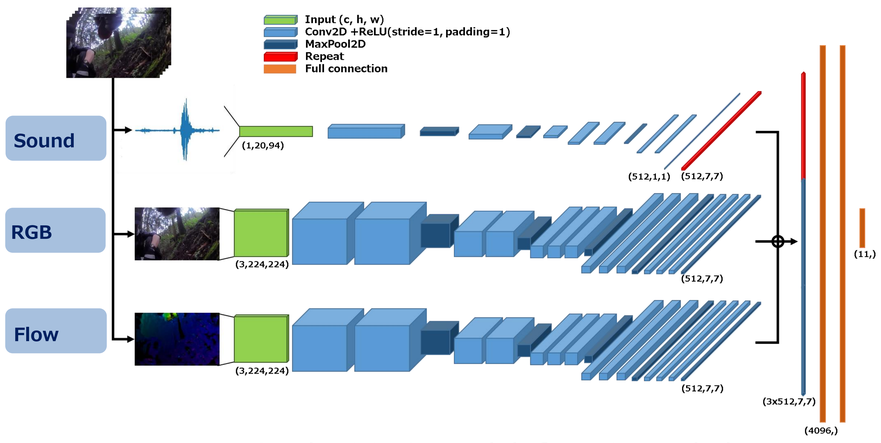


bark	Cling	Com-mand	eat-drink	handler	run	see victim	shake	sniff	stop	walk-trot
1,744	1,127	2,439	343	2,011	98	1,549	239	7,719	6,384	8,764 (frames)

Method

Sound/image-based Three Stream Network

・Multi-labeled action recognition by using RGB images, optical flow images, and sound.
・The three-stream network consists of the audio stream and two image streams.



・We use the ImageNet pretrained VGG16 model for RGB and Flow streams.
・The MFCC spectrogram extracted from the sound is an input as an image to the sound stream.
・Estimation is executed per frame.

Experiments

Comparative

・We made eight experiments with different combination of the inputs.
・We compared the performance of the proposed three-stream network with that of the one/two-stream networks as ablation studies.

	RGB	Flow	Sound	
(1)	✓	-	-	VGG16 pre-trained
(2)	-	✓	-	VGG16 pre-trained
(3)	-	-	1D	Sound stream with Conv1D
(4)	-	-	2D	Sound stream with Conv2D
(5)	✓	✓	-	2-stream CNN
(6)	✓	-	2D	Sound/image-based 2-stream CNN
(7)	-	✓	2D	Sound/image-based 2-stream CNN
(8)	✓	✓	2D	**Sound/image-based 3-stream CNN**

・For each video, 70% of the front was used for learning and 30% of the later was used for evaluation.

・In each case, **sound data is effective for multi-class estimation**, and accuracy decreases when combined with the other one.
・Feature extraction is not working from RGB and Flow.
・In case of **combining Sound with RGB and Flow, the accuracy is improved** compared to sound data only and image/sound-based two-stream.

・In the proposed method, the result was obtained high accuracy initially expected with from compensating each other's missing information.

	bark	cling	command	eat-drink	look at handler	run	see victim	shake	sniff	stop	walk-trot	Whole
(1)	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)	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)	***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)	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)	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)	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)	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)	0.577	***0.135***	0.186	***0.066***	***0.183***	***0.026***	***0.433***	0.409	***0.53***	***0.779***	*0.725*	***0.518***

Conclusions

・Result

・Estimated accuracy 51.8% with the proposed method.
・Each of the three types of data has the necessary information. 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

・Future Work

・Accuracy improvement
- Pre-processing of dog first person view.
- Data set enhancement.
・Real-time processing

Refference

[1]Y. Aytar, C. Vondrick, and A. A. Torralba. Soundnet: Learningsound representations from unlabeled video. InAdvances inNeural 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.InProc.of IEEE Computer Vision and Pattern Recognition,2009.
[3]C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolu-tional two-stream network fusion for video action recognition.InProc.of IEEE Computer Vision and Pattern Recognition,2016.
[4]B. Gedas, S. P. Stella, X. Y.and Hyun, and S. Jianbo. Am Ia baller? basketball skill assessment using first-person cam-eras. InProc.of IEEE International Conference on ComputerVision, 2016.
[5]Y. Iwashita, A. Takamine, R. Kurazume, and M. S. Ryoo.First-person animal activity recognition from egocentricvideos. InProc. of International Conference on PatternRecognition (ICPR), 2014.
[6]Y. Komori, T. Fujieda, K. Ohno, T. Suzuki, and S. Tadokoro.Detection of continuous barking actions from search and res-cue dogs' activities data. InProc.of IEEE/RSJ InternationalConference on Intelligent Robots and Systems (IROS), pages630–635, 2015.
[7]M. Minghuang, F. Haoqi, and M. K. Kris. Going deeper intofirst-person activity recognition. InProc.of IEEE ComputerVision and Pattern Recognition, 2016.
[8]K. Simonyan, A. Vedaldi, and A. Zisserman. Very deep con-volutional networks for large-scale image recognition. InProc. of International Conference on Learning Representa-tions, 2015.
[9]K. Simonyan and A. Zisserman. Two-stream convolutionalnetworks for action recognition in videos. InAdvances in Neu-ral Information Processing Systems, pages 568–576, 2014.