

# Rescue Dog Action Recognition by Integrating Ego-centric Video, Sound and Sensor Information

Yuta Ide<sup>1</sup>, Tsuyohito Araki<sup>1</sup>,  
Ryunosuke Hamada<sup>2</sup>, Kazunori Ohno<sup>2</sup>, and Keiji Yanai<sup>1</sup>

<sup>1</sup> Department of Informatics, The University of Electro-Communications, Tokyo

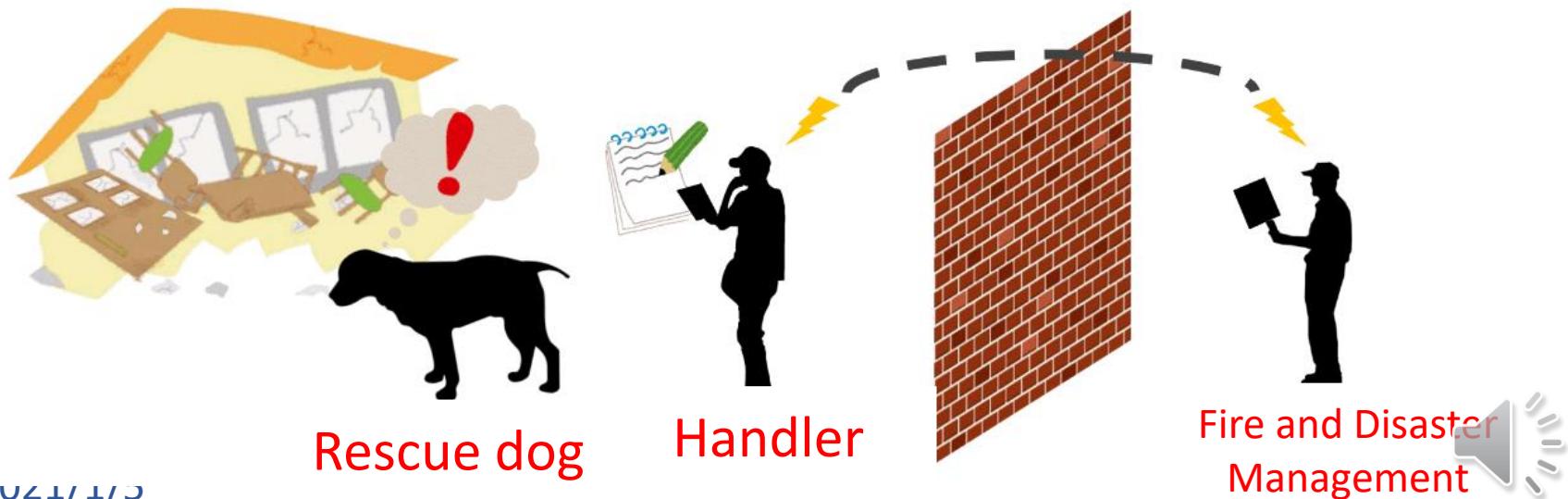
<sup>2</sup> NICHe, Tohoku University

{ide-y, araki-t, yanai}@mm.inf.uec.ac.jp



## What is a rescue dog?

- A dog that does not perform rescue work, but searches for victims in disaster areas.
- Handler manually records actions and verbally explains them to commander
- Make a disaster rescue plan based on the information from rescue dogs and handlers.



# Challenges in Utilizing Rescue Dogs

- Manual recording by handlers is insufficient in terms of objectivity and quantity of information.
- Verbal explanation is insufficient to ensure accuracy of information

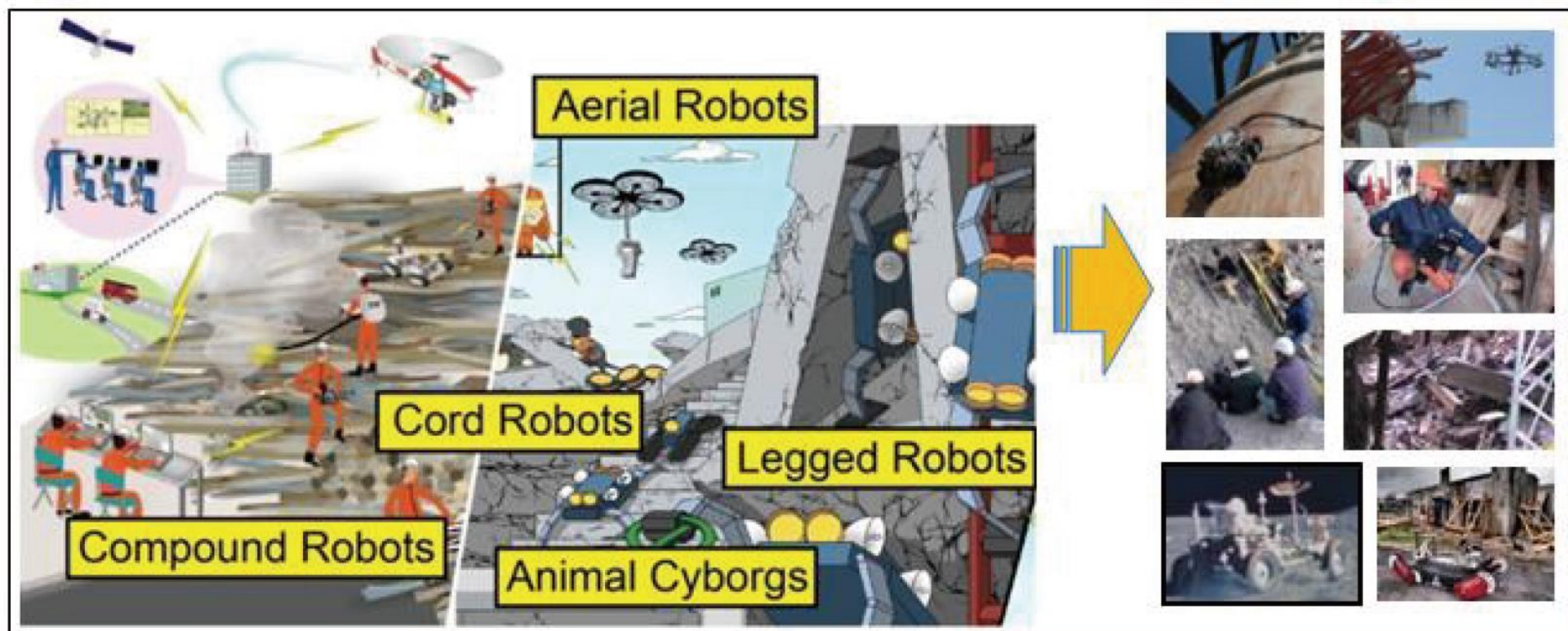


**• More and more accurate information on the rescue dog is required**



# Background

## ImPACT



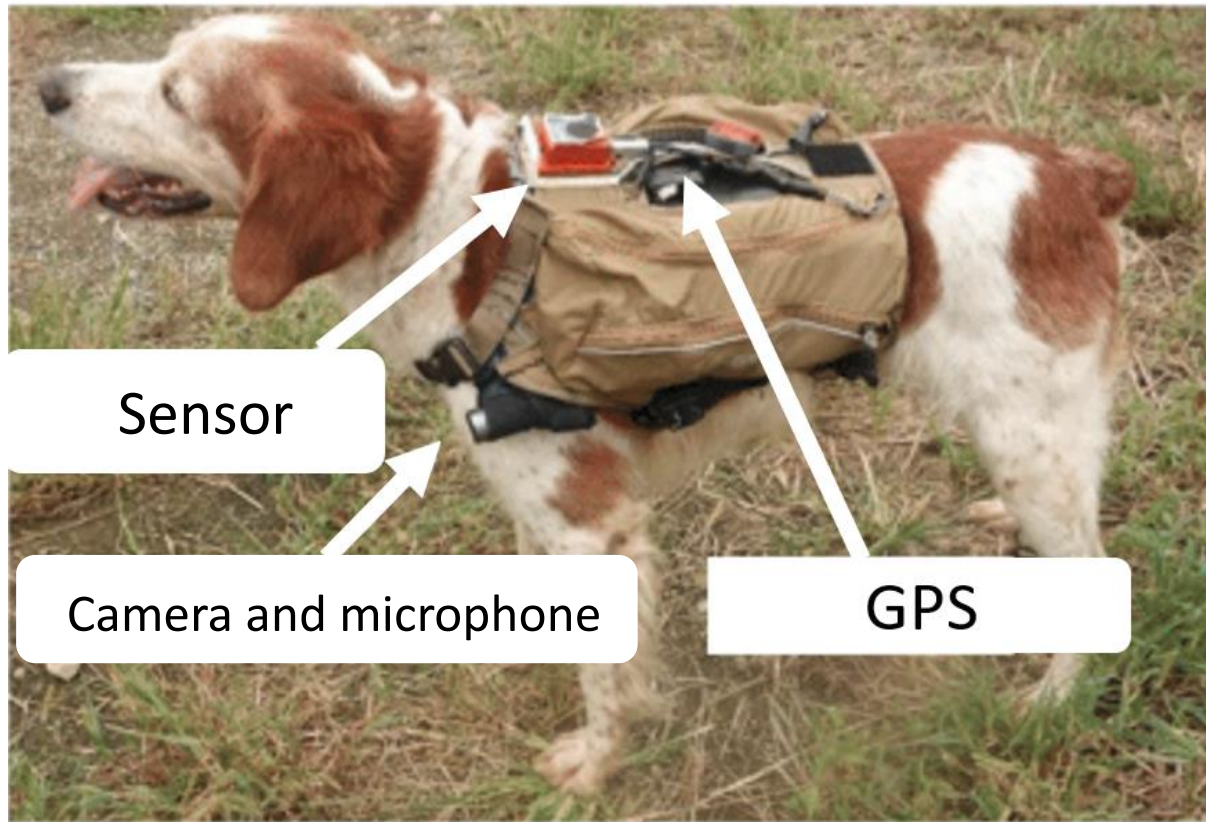
<https://www.jst.go.jp/impact/program/07.html>



# Cyber suit

Development of a wearable measurement and recording device

Equipped with camera, microphone, inertial sensor, etc.

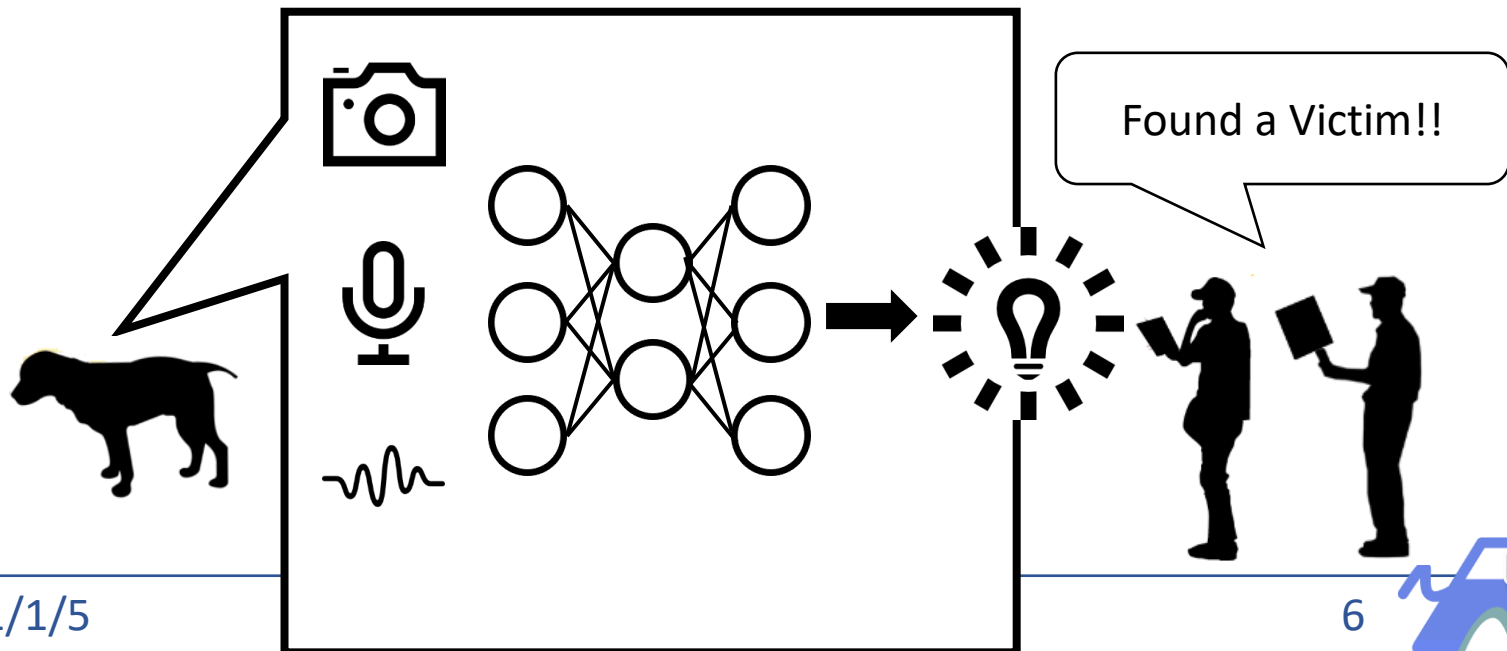


[K. Yuichi, at el. ICIR, 2015]



# Objective

- Using Deep Learning to Recognize Rescue Dog Behavior
- Using the multimodal rescue dog data (video, audio, and sensor information) provided by Professor Ohno of Tohoku University.



# Related Work

- Two-Stream CNN  
[K. Simonyan, NIPS 2014]

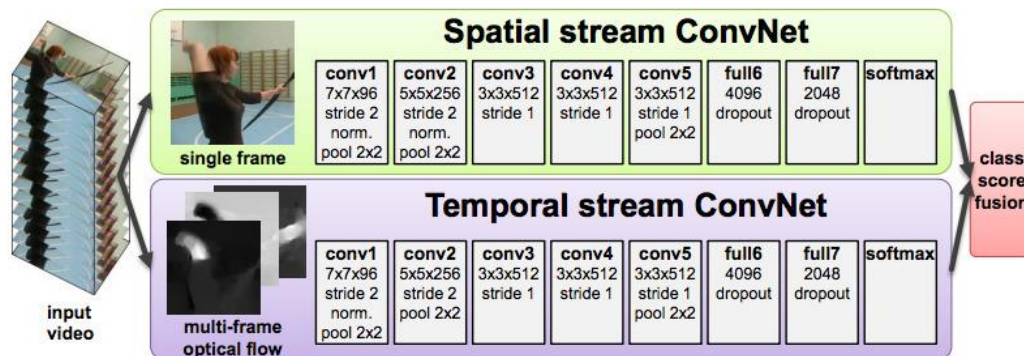
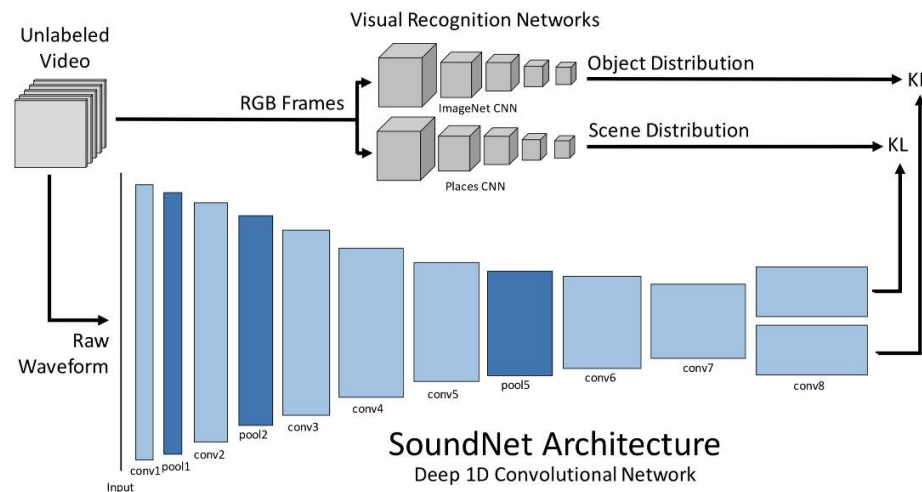


Figure 1: Two-stream architecture for video classification.

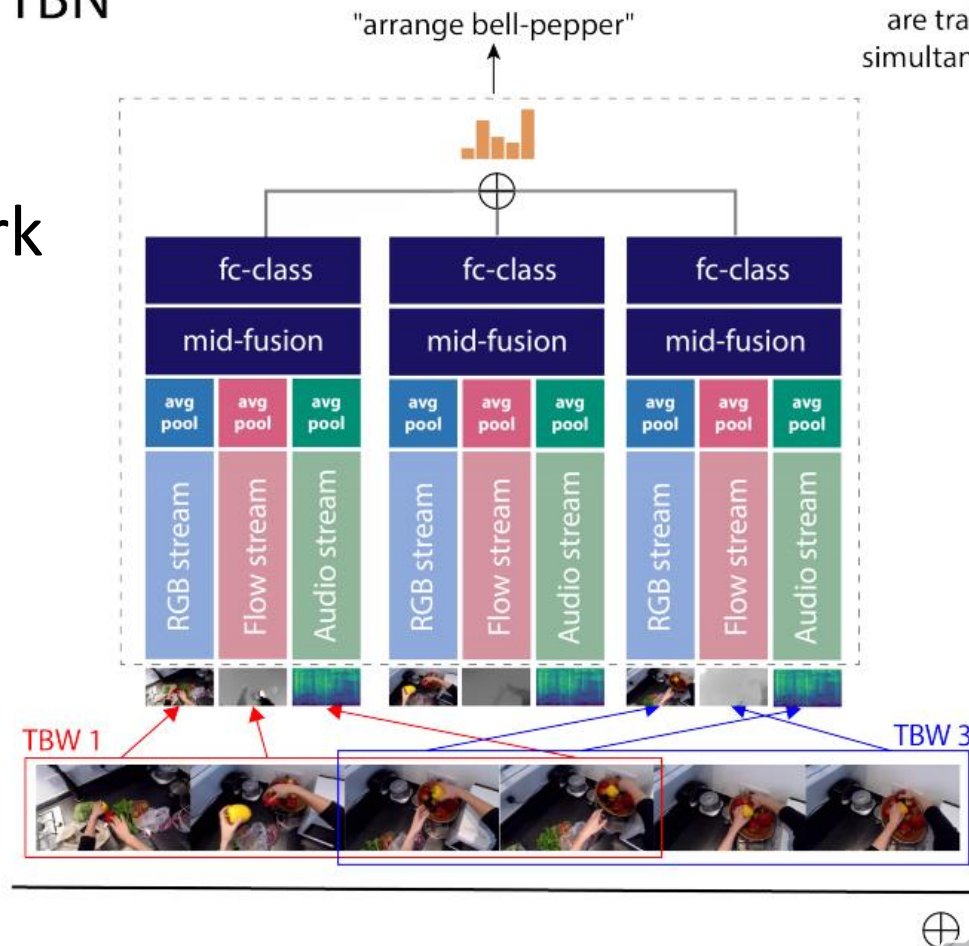
- Sound Net  
[Y. Aytar, NIPS 2016]



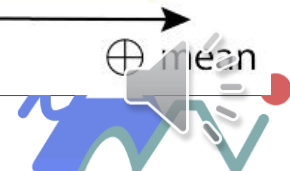
# Related Work

TBN

All modalities are trained simultaneously

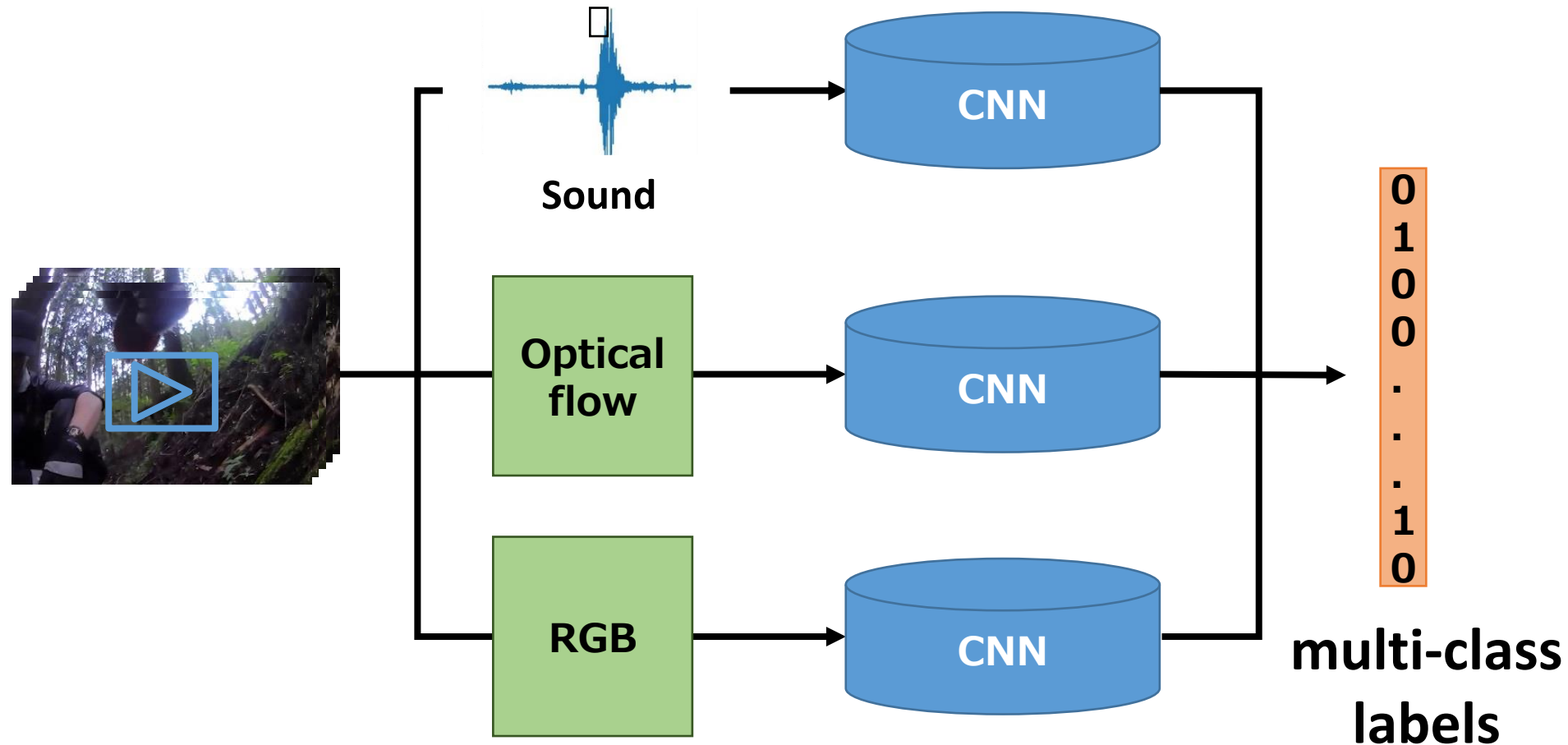


- Temporal Binding Network [K.Evangelos, ICCV 2019]





# Previous Work



•Dogcentric activity recognition by integrating appearance, motion and sound. [T. Araki, EPIC 2019]



# Dataset

## Rescue Dog Data Set

- A group of videos showing rescue dogs in training.



rescue dog's point



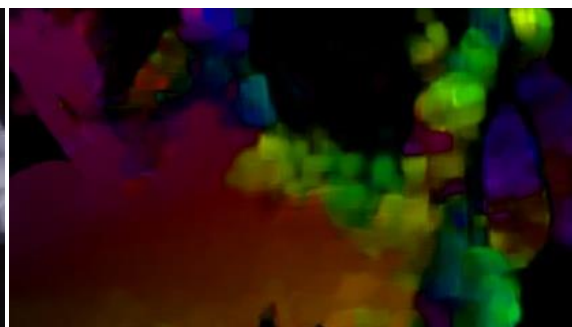
the handler's point



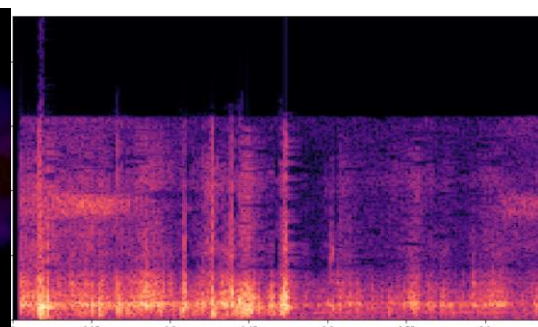
third party's point



RGB Image



Optical flow Image



STFT



## Rescue Dog Data Set

- Information obtained from sensors

- $x$  sens time:second(s)
- $G_x, G_y, G_z$ : angular velocity(deg/s)
- $A_x, A_y, A_z$ : acceleration(m/s<sup>2</sup>)
- Roll,Pitch,Yaw: posture(degree)

- $M_x, M_y, M_z$ : geomagnetism( $\mu$ T)
- Pressure(hPa)
- Temperature



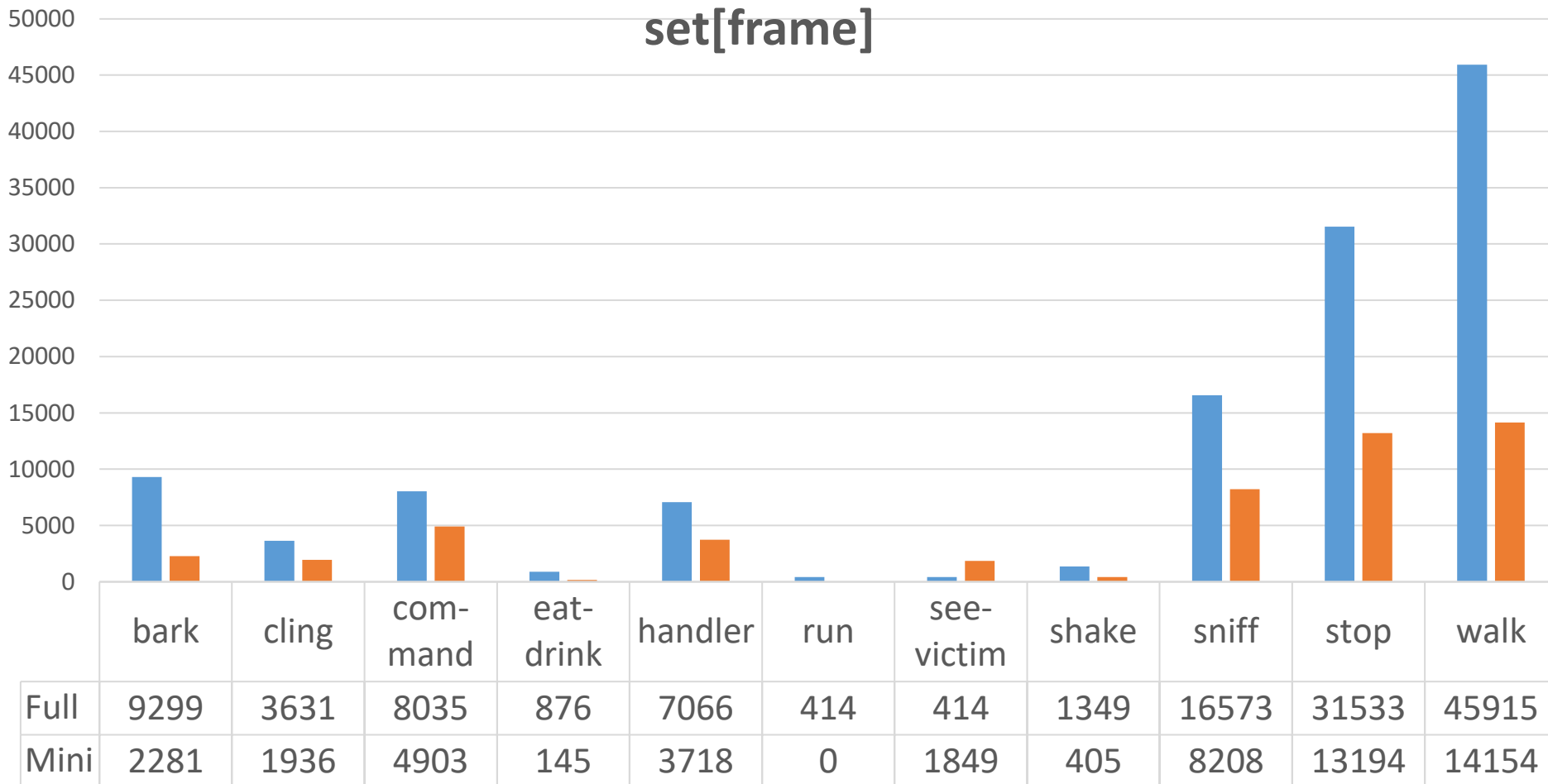
**Data used**

- <https://www.amtechs.co.jp/product/gps/post-130.html>



# Dataset

Number of frames appearing in each class per data set[frame]



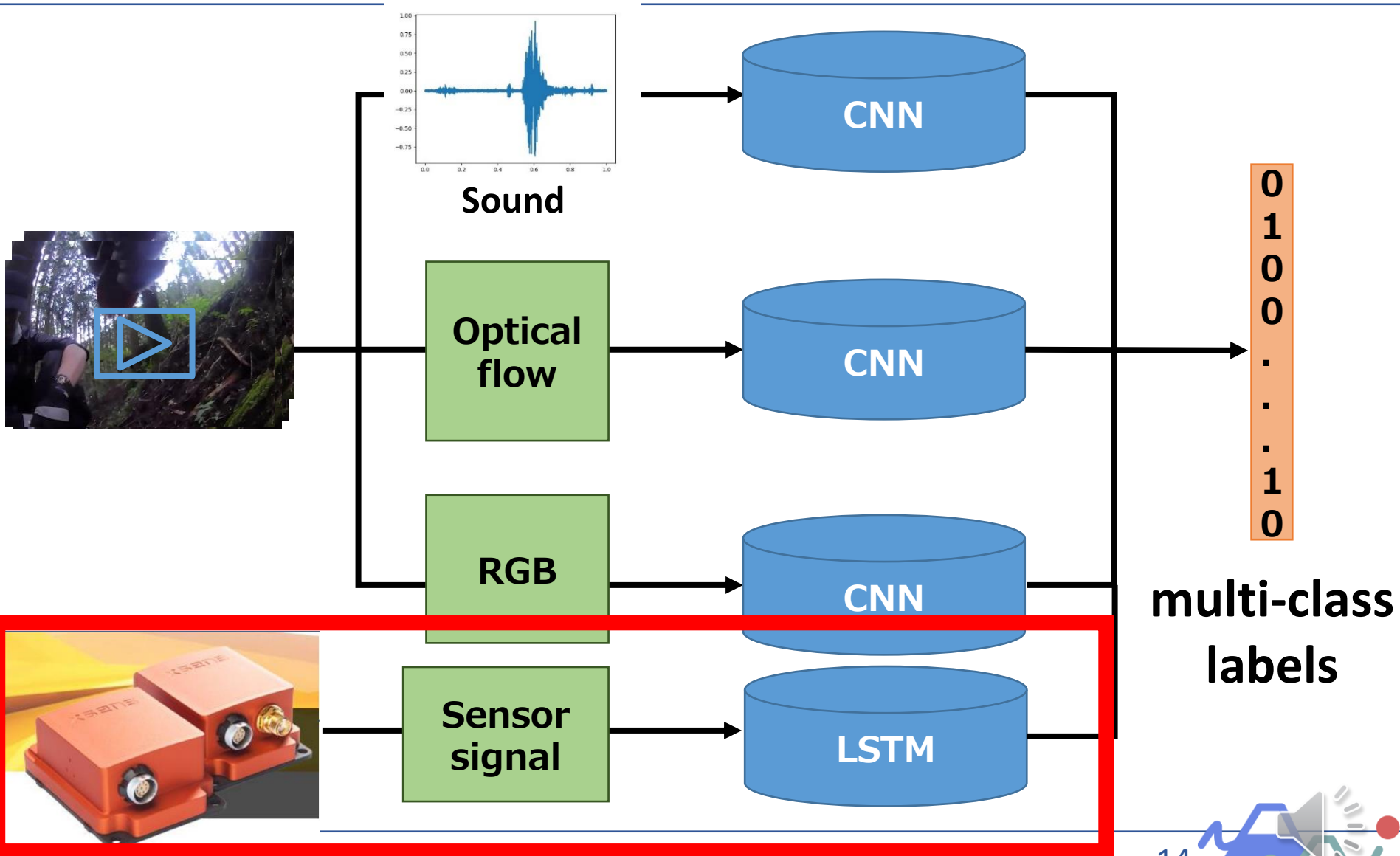
# Dog Activity Classes

Some of the behavior classes are described in detail

- cling : The situation in which a dog is sniffing with the nose close to the smell.
- command : The situation in which the dog is being instructed by the handler.
- look at handler : The situation in which the dog is looking at the handler. Hereinafter, this action is called just “handler”.



# Outline of the proposed method



## MultiLabel SoftMarginLoss

$$\begin{aligned} \text{loss}(x, y) = & -\frac{1}{C} * \sum_i(\{y_i\} * \log((1 + \exp(-x_i))^{-1}) \\ & + (1 - y_i) * \log(\frac{\exp(-x_i)}{1 + \exp(-x_i)})) \end{aligned}$$

x : output of the network

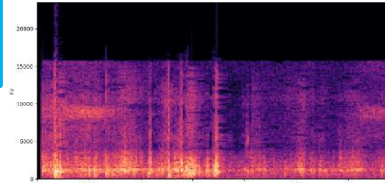
y : target,

C : numer of the classes



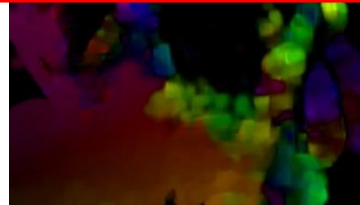
# Method

**Sound stream**  
(1,224,224)



Resnet-101

**RGB stream**  
(3,224,224)



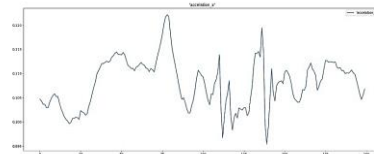
VGG-16

**Optical flow stream**  
(3,224,224)



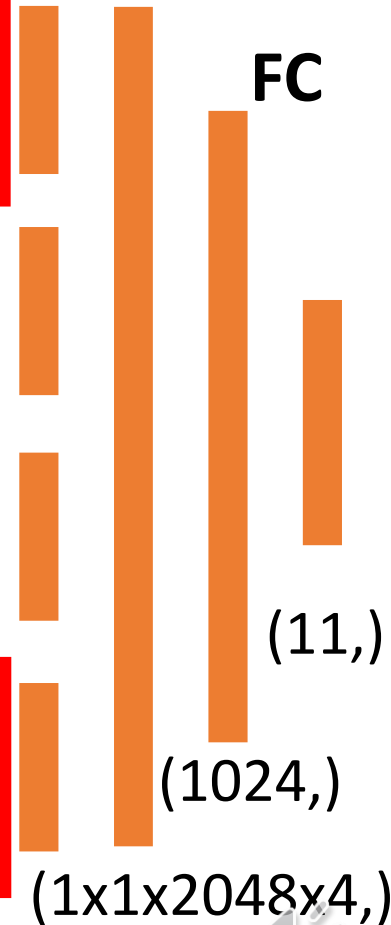
VGG-16

**Sensor stream**  
(310,9)

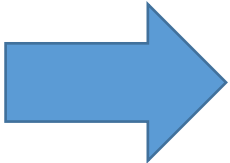


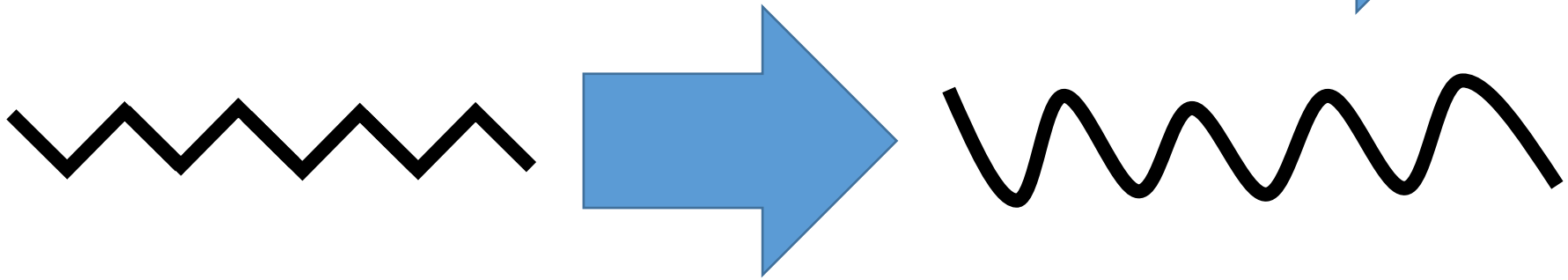
Bi-LSTM

concat





- Sensor information(310,9)
  - Sensor acquires information every 0.005 seconds  
 $0.005 \times 320 = 1.6(s)$   
Smoothing with 10 pieces of data 320  310



9 :  $G_x, G_y, G_z$ : angular velocity(deg/s)

$A_x, A_y, A_z$ : acceleration (m/s<sup>2</sup>)

Roll,Pitch,Yaw: attitude (degree)



- Data set used

Rescue dog training data recorded on July 10, 2016.  
(about 12 minutes)

Rescue dog training data recorded on November 11, 2016.  
(about 4 minutes and 50 seconds)

- Training data set

First 80% of the data set used (22979 frames)

- Data set for evaluation

Second half 20% of the dataset used (6643 frames)

- Indicators for accuracy comparison

Jaccard coefficient

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

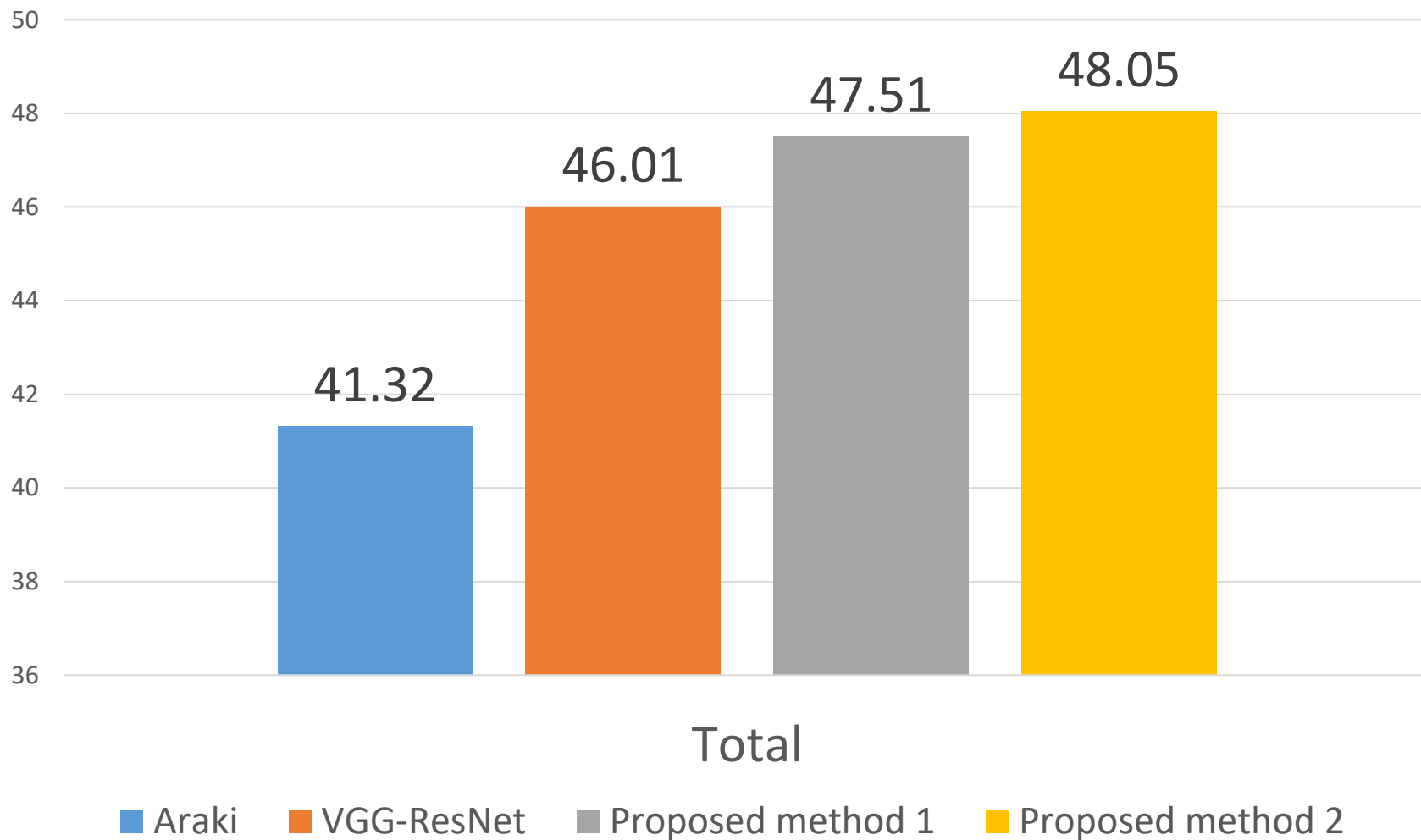


# Differences between the previous study and the proposed method

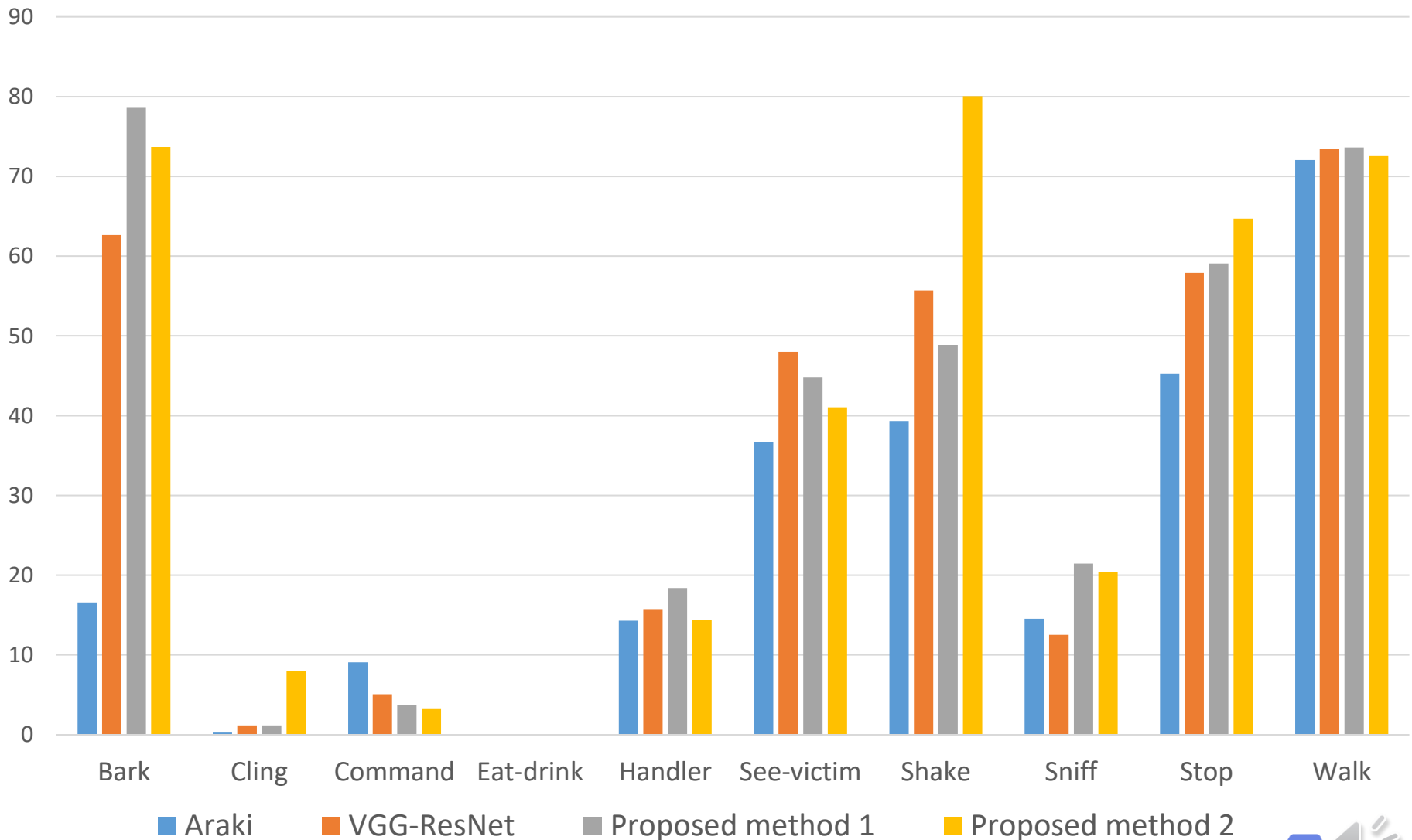
	Araki et al.	VGG-ResNet	Proposed method 1	Proposed method 2
RGB	VGG-16	VGG-16	VGG-16	VGG-16
Optical Flow	VGG-16	VGG-16	VGG-16	VGG-16
Sound	2D Conv (MFCC)	ResNet-101 (STFT)	ResNet-101 (STFT)	ResNet-101 (STFT)
Sensor	None	None	Bi-GRU	Bi-LSTM



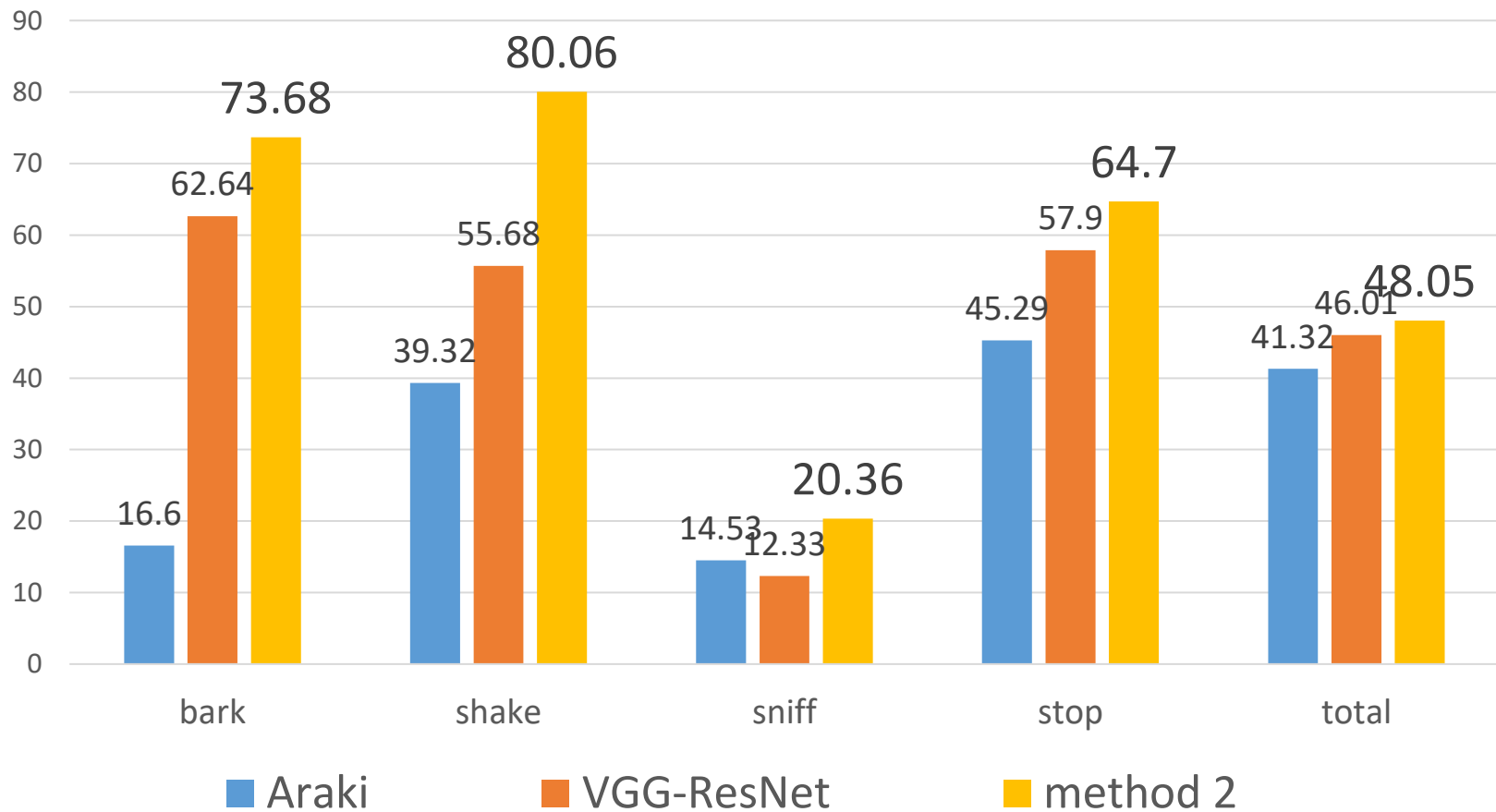
# Comparison with each method [%]



# Comparison with each method [%]



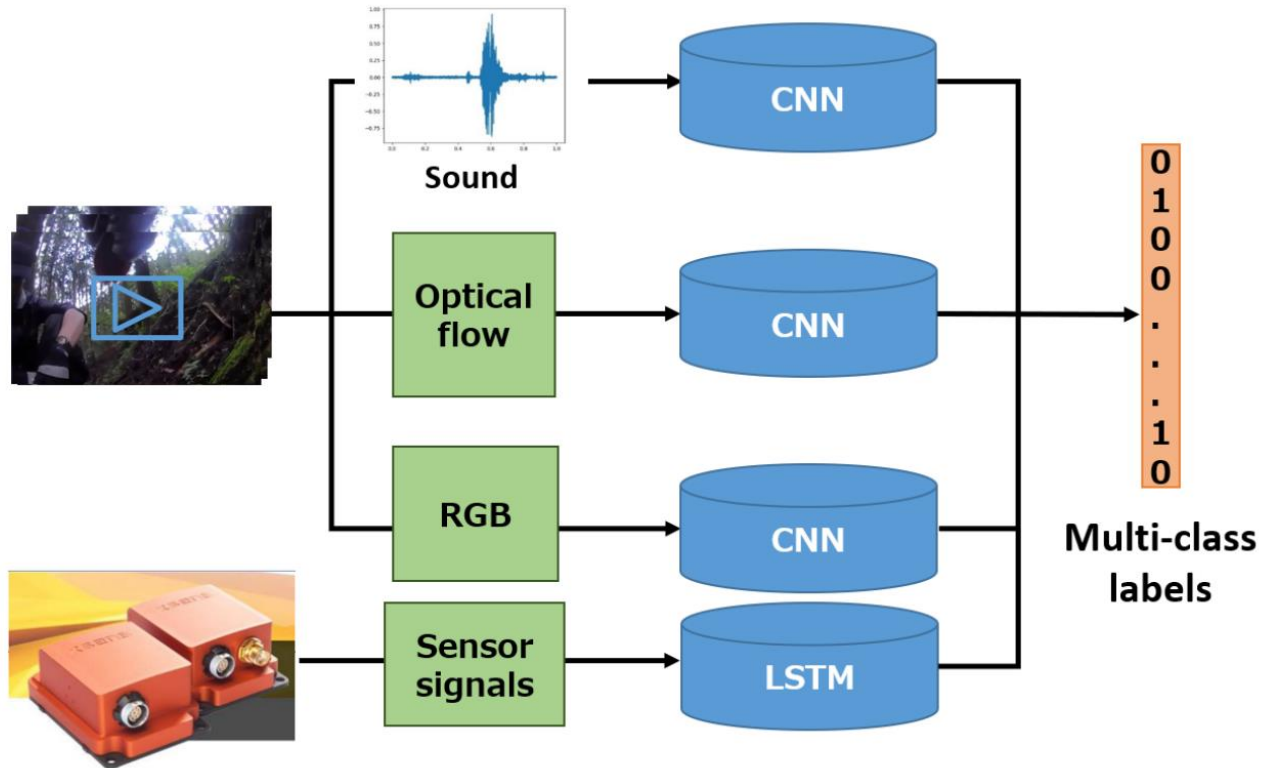
Comparison between Araki's method and the proposed method[%]



- Use of sensor information is essential.



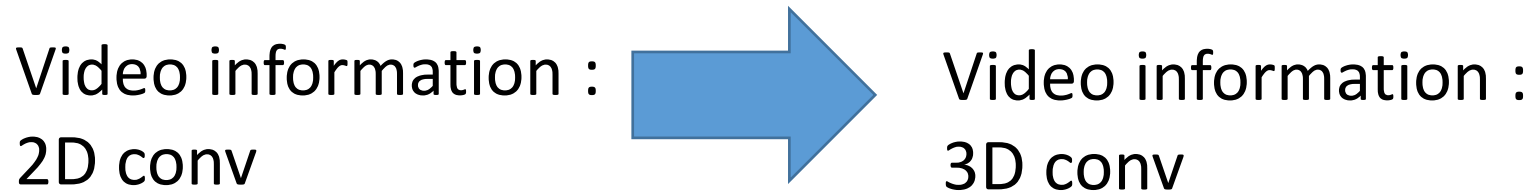
# Conclusions



- Proposed an image/sound/sensor-based four-stream CNN
- The effectiveness of using sensor information is effective.



- Use of time series information



- Expansion of the data set

Increase in the number of datasets that include video and sensor information





# UEC

TOKYO

A large, stylized version of the UEC TOKYO logo. The letters 'U', 'E', and 'C' are in a bold, blue, serif font. The 'U' and 'E' are light blue, while the 'C' is a darker blue. The word 'TOKYO' is in a smaller, dark blue, serif font below the 'C'. The logo is decorated with three glowing blue spheres and curved lines that suggest motion or a path.