

# Dog-centric Activity Estimation

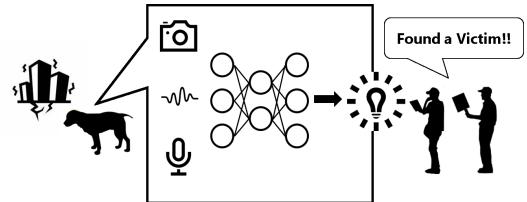
TSUYOHITO ARAKI<sup>1,a)</sup> YUTA IDE<sup>1</sup> KAZUNORI OHNO<sup>2</sup> KAZUNORI OHNO<sup>2,3</sup>  
KEIJI YANAI<sup>1</sup>

## Abstract

Dogs that assist disaster relief in the affected areas are called rescue dogs, and dogs equipped measuring device such as cameras are called cyber rescue dogs. In this study, we aim to identify the behavior of cyber rescue dog from the sensor attached to the dog. We proposed a Sound/image-based three-stream CNN that classifies dog activity from dog-centric movies and sounds. And experiment using the proposed method for multiclass estimation. Sound/image-based three-stream CNN is a movie identification network that receives still images, optical flow images, and audio obtained from movie. In comparison experiments to confirm the effectiveness of the proposed method, single pattern and multiple patterns were used for each of the three inputs. As a result, the highest accuracy was obtained with the proposed method at 51.8%. For estimation, we used a rescue dog training data set. This data set is still being created, and the amount of data needed for learning is not enough. The result is insufficient even if Given the complex dataset of the rescue dog training and the task of multi-class estimation from dog-centric movies. We need to further refine this method and extend the data set in the future.

## 1. Introduction

In rescue activities in the affected areas, trained rescue dogs may conduct surveys as human assistants. The rescue dog makes a pair with a human and investigate for a affected areas making use of the characteristics as a dog. Resque dogs can investigate even in environments where it is difficult for a person to traverse on such as a narrow space and a crevice or a collapsed building. And a rescue operation that relies on the developed sense of smell is possible. However, they have no language for humans. People in pairs with rescue dogs are called handler, and handlers must interpret from rescue dog's behavior the information that they collected. Under the present circumstances, a handler who directs a rescue dog manually marks the action of the rescue dog and orally transmits information to a commander such as a fire department. As a problem of joint investigation with this



**Fig. 1** Outline of the proposed method. Extract multiple data from the input image and enter them into different streams. Based on the output of each stream, estimate the rescue dog behavior.

rescue dog, there is a lack of information on surrounding environment and victims for triage. And, the records by the handler are subjective, so they lack objectivity. Verbal transmission also makes them less accurate. In this study, we aim to estimate the behavior of rescue dogs using data obtained by attaching sensors to rescue dogs (Figure1). The specific task is class estimation using multimodal information that uses not only video but also audio data. As a result, this makes it possible to determine mechanically what the rescue dog is doing now. And, information necessary for triage is organized, disaster rescue activities will be more efficiently.

## 2. Related Work

There are two-stream CNN as a study of video classification [4]. It is classify videos by learn motion information from optical flow image that calculated between frame and frame. Ohno, Shibata et al. Developed a wearable measurement recording device for monitoring of rescue dog behavior [2]. We show that attachable device with rescue dog Figure 2. Rescue dogs wearing that called cyber rescue dogs. This device can record measured data and broadcast information as a movie and so on. It makes handler can understand the beyond rescue dog activity that over the rubble.

Also, there are study that estimated dog activity from dog-centric video by Ehsan. at al. It is modeling dog activity and estimate how dogs will move. However, that study used only video, not used multimodal information such as audio. Rescue dogs are asked more information that can be judged not only from the dog's surrounding environment information and moving images. For example, rescue dogs are trained to stay on the spot and continue to bark when they find a victim. In this way, handler must understand the surrounding environment of the rescue dog using not only video

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

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

<sup>3</sup> Organic Physical Chemistry Laboratory

a) araki-t@mm.inf.uec.ac.jp



**Fig. 2** Wearable measurement and recording device [2].

data but also information such as voice data, inertia data and GPS data in combination. But, we have been provided only dog-centric motion pictures and voice information, in this study use these inputs for estimate the activity of rescue dogs.

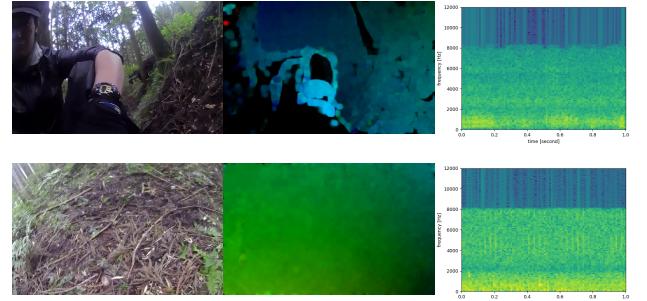
SoundNet as a study of movie classification focus it's sounds [1]. In the SoundNet, deep convolutional sound network is trained with visual supervision. This method achieved the state-of-the-art results on standard benchmarks for acoustic scene/object classification. Semantic information of audio has been shown to be important for movie recognition, we don't aimed classification from only audio. In this study build the architecture connected two-stream network with deep audio network while referring to SoundNet and two-stream CNN. SoundNet takes an audio waveform as an input. On the other hand, this study is different in that it extracts MFCC features from audio and then uses them as input to a stream.

There is research of multimodal learning using not only audio and moving pictures but also acceleration and speed information by Tanno et al [2]. That research reported that improvement and decrease in recognition accuracy by adding information. In particular, classification accuracy is improved by audio. It is the same theme as this study in terms of multimodal deep learning, but it differs from this study in terms of multiclass estimation.

### 3. Dateset

We use the Training Cyber Rescue Dog Dataset. This is data collection about training rescue dog wearing recording device. We are provided some piece of that data by Ohno et al. Provided data consists of seven movies, each movie has three view point: dog-centric, handler-centric and third-person view. Each view are labeled rescue dog activities based on time lange. There are 11 classes of rescue dog activities: bark, cling\_to\_something, command\_by\_handler, eat\_drink, look\_at\_handler, run, see\_victim, shake\_body, sniff, stop\_legs and walk-trot. Figure 3 show part of that. Each class annotated every movies to specifying time lange. Because multiple classes overlap at the same time, it is treated as multi-label data.

Total time is 57 minutes and 40 seconds, 29.97 FPS, the number of frames is 103,696. We thinned the frames to 6fps for use to train and evaluation. Table 1 shows the frequency



**Fig. 3** Class “see victim” (upper) and class “stop” (lower) from Training Cyber Rescue Dataset. From the left, still image, optical flow image and audio data visualized by MFCC spectrum.

of occurrence of each class. Class names are abbreviated for convenience of notation. Each class has 100 to 9,000 samples, so it looks numerically good enough. However, there are continuous scenes. Therefore, the content of the data is actually lacking in diversity. For example, the class “run” has 98 samples, but only 3 scenes as independent scenes. Also, the sum of scene “bark”, “eat-drink”, “see\_victim” and “shake\_body” is less than 100 scenes.

Three data were extracted for each frame: still image, optical flow image calculated from immediately frame and 15 frames of audio before and after the frame.

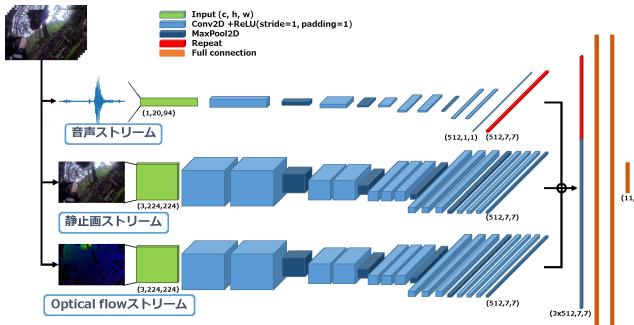
### 4. Proposed Method

本研究ではレスキュー犬の行動推定のために、動画像・音声のマルチラベル推定を行う。そのため、レスキュー犬訓練データセットを認識するための sound/image-based three-stream CNN を提案する。これは、two-stream CNN に音声ストリームを加えたモデルである。静止画像、optical flow 画像、音声を別々のストリームへの入力とし、それぞれの出力を元にレスキュー犬の行動を推定する。Sound/image-based three-stream CNN のアーキテクチャを Figure 4 に示す。

まず入力となる動画から静止画像のフレーム ( $F_t$ ) を取り出し、直後の  $F_{t+1}$  間との optical flow 画像 ( $O_t$ ) を生成する。次に両画像から同じ構造の 2 つのネットワークを用いて特徴量の抽出を行う。そして対応する音声 ( $A_t$ ) から MFCC ( $M_t$ ) を求め、前述とは異なる構造のネットワークを用いて ( $M_t$ ) から特徴量の抽出を行う。最後にこれら 2 つあるいは 3 つの特徴量の組み合わせ毎に結合し、分類ネットワークでクラス分類を行う。この際に、音声をフレームと同じサイズで切り出すと特徴が著しく失われるため入力音声には  $F_t$  の前後 0.5 秒ずつを用いた。動画あるいは音声から実際に犬の行動を推定する場合を想定し、現実的で取り扱いやすい時間としてこれを設定した。動画長は 31 フレームとし、中央から切り出した静止画像とそれに対応する直後の optical flow 画像をそれぞれ ImageNet で学習済みの VGG16 モデルに通し、畳み込み層の出力を結合した。動画から切り出した音声は畳み込みを繰り返すと特徴の縦横次元が小さくなるため、静止画像の畳み込みの出力と同じサイズになるように調整した。調整は畳み込みで奥行き次元を揃えた後、同じ特徴をリピートし目的の大きさになるまでコピーして結合した。分類はフレーム毎に行った。推定にはクラス毎に SoftMarginLoss を用いた。入力を  $x$ 、出力を  $y$ 、クラス数を  $C$  とすると、マルチ

**Table 1** Class frequency of occurrence of Training Cyber Rescue Dataset

Class name	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk
Occurrence	1744	1127	2439	343	2011	98	1549	239	7719	6384	8764



**Fig. 4** Sound/image-based three-stream アーキテクチャ. 音声, 静止画像, optical flow 画像それぞれを入力とする 3 ストリームからなり, それぞれの出力を結合してクラスを推定している.

クラス推定の損失関数 SoftMarginLoss は式 1 に定義される. 推定クラスが正解なら  $\Sigma$  内の第 2 項, 不正解なら第 1 項が計算に利用するように設計されており, 推定ラベルによって関数が変わる. 本研究では 11 クラスを取り扱うため, 出力  $y$  は 11 次元のバイナリとする. 閾値=0.5 とし, 閾値以上のクラスを推定クラスとした.

$$\begin{aligned} loss(x, y) = & -\frac{1}{C} * \sum_i y[i] * \log((1 + \exp(-x[i]))^{-1}) \\ & + (1 - y[i]) * \log(\frac{\exp(-x[i])}{1 + \exp(-x[i])}) \end{aligned} \quad (1)$$

学習は全てにおいて 100 エポック行った. 学習率は  $1e-03 \sim 1e-06$ , バッチサイズは 32 ~ 128 の範囲で学習毎に変え制限の中で最も良い精度の結果を評価した.

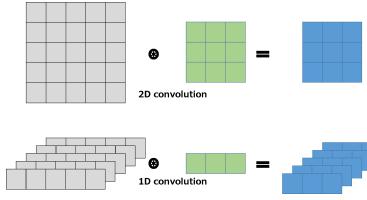
## 5. Experiments

3 つの入力について, それぞれ単体, 2 つずつの組み合わせおよび全てを統合した場合の 8 通りを行った. 画像単体を入力とした実験では ImageNet を学習した VGG16 の学習済みモデルを用い再学習した. 音声での学習は畳み込み層について 1D と 2D の 2 種類で行った. 畳み込み層の次元の差について Figure 5 に詳細を示す. 出力の向きを合わせるために, 提案手法の複数入力には 2D の畳み込みを用いた. 一人称視点動画像からのマルチ行動推定を行なった先行研究は見つけられなかっただため, 既存手法との比較は行なっていない. 表では, クラス毎の精度と全体を合計しての精度を Jaccard 係数で示している. なお, Jaccard 係数とは

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

で表され, Precision と Recall の両者について F 尺度と比較してより厳格な値が求まる. レスキュー犬の行動分類にあたり, Precision と Recall を共に重視するためにこの係数を採用した. よって, 本研究では Jaccard 係数がより大きいモデルは精度がより良いと表現する.

結果のまとめを表 2 に示す. 静止画像, optical flow 画像单



**Fig. 5** 1D convolution と 2D convolution の詳細. データは同様でも入力の形と処理が異なり, 出力の結果も異なっている.

体では精度が低く, これらを組合せたもので精度が上昇した. 比較して, 音声単体では 1D, 2D ともに精度が高かった. 音声と静止画像, 音声と optical flow 画像を組合せた実験では, 音声単体と比較して全体的な精度は低下した. しかし, 静止画像, optical flow 画像, 音声の 3 つを組合せた提案手法では全体的な精度の上昇が見られ, クラス別で見ても半数以上のクラスの数値が上昇している. 人間が耳で聞いた際にもその特徴を識別しやすい bark, shake クラスにおいては音声単体を 1D 畳み込み層で学習したものの方が数値が高い. これから, データセットに対する提案手法の有効性が示された.

## 6. Conclusion and Future Work

Sound/image-based three-stream CNN の提案と, 提案手法を用いたレスキュー犬の行動推定を行なった. 音声データはクラス推定に強力であるものの, 音声・静止画像・optical flow 画像の 3 つのデータにそれぞれ必要な情報が含まれていることがわかった. 提案手法が相対的に最も精度が高かったが, 51.8%と数値では決して高いとは言えない. 本研究の目的はレスキュー犬の行動推定という人命のかかったタスクである. ハンドラーの補助的な役割を任せた運用をこなせるとしても, 実際に現場で判断を任せるにはまだ不十分な結果となった.

精度をより上げるためには, 現在の手法の改良, 新しい手法の取り入れ, データセットの拡張が考えられる. 例えば, 音声について現在は静止画像の前後 1 秒を抽出しているが最適なフレーム長を調べるなどの余地がある. 特徴抽出についても今回は音声の特徴抽出に MFCC 特徴を採用したが, [1] のように波形をそのまま入力する分類手法も存在する. また, 人間の一人称視点映像の分類研究で用いられているような動画分類特有の処理を入れるなどの手法を取り入れることで精度の向上が期待できる. 例えば, [3] で用いられている腕のセグメンテーションネットワークのように, 犬領域を推定するようなネットワークは犬動作の推定にも応用できる. これは犬の状態推定への貢献が期待される. ただし, 腕領域とは異なり犬領域のデータセットは確認できていない. 音声からの特徴抽出は, 音声フレームの長さが実験に対して適しているのかという疑問も残っている. 今回は検証しなかったが, 本来は推定に最適なフレーム長を求めるべきである. より適切な音声フレーム長が判断できればより高い精度も期待できる. 音声の特徴抽出方法についても最適と断言するには至っていない. 今回音声の特徴抽出に MFCC を採用したが, [1] のよう

Table 2 各実験結果比較表.

	静止画像	optical flow 画像	音声	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk	全体
(1)	x	x	x	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)	x		x	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)	x	x	1D	<b>0.669</b>	0.078	0.22	0.023	0.138	0.0	0.274	<b>0.44</b>	0.502	0.745	0.704	0.512
(4)	x	x	2D	0.563	0.04	0.188	0.001	0.059	0.0	0.201	0.304	0.524	0.744	<b>0.74</b>	0.512
(5)		x		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)		x	2D	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)	x		2D	0.667	0.054	<b>0.234</b>	0.014	0.123	0.01	0.223	0.356	0.487	0.759	0.692	0.493
(8)			2D	0.577	<b>0.135</b>	0.186	<b>0.066</b>	<b>0.183</b>	<b>0.026</b>	<b>0.433</b>	0.409	<b>0.53</b>	<b>0.779</b>	0.725	<b>0.518</b>

に波形をそのまま入力する分類手法も存在する。今回は取り扱わなかったが、検討の価値があるだろう。データセットについても本研究で利用した内容は十分ではない。特に、eat, shake, run クラスなどは圧倒的にデータ量が少ない。クラス毎のデータ数だけでなく、慣性センサなどから取得される情報の利用も動作推定の精度向上に対する効果が期待される。レスキュー犬訓練データの増強は必須課題とも言える。

さらに、今回は研究の範囲としなかったが実際の現場での利用を想定した場合、レスキュー犬行動動画の入力に対してリアルタイムに結果を出すことも求められる。

## References

- [1] Aytar, Y., Vondrick, C. and Torralba, A. A.: Soundnet: Learning sound representations from unlabeled video, *Advances in Neural Information Processing Systems* (2016).
- [2] Komori, Y., Fujieda, T., Ohno, K., Suzuki, T. and Tadokoro, S.: Detection of Continuous Barking Actions from Search and Rescue Dogs' Activities Data, *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 630–635 (2015).
- [3] Minghuang, M., Haoqi, F. and Kris, M. K.: Going Deeper into First-Person Activity Recognition, *Proc. of IEEE Computer Vision and Pattern Recognition* (2016).
- [4] Simonyan, K. and Zisserman, A.: Two-stream convolutional networks for action recognition in videos, *Advances in Neural Information Processing Systems*, pp. 568–576 (2014).