

レスキュー犬の一人称動画を用いた動作分類

発表者： 情報学専攻 メディア情報学 プログラム 学籍番号 1730010 荒木 勇人
 指導教員： 柳井 啓司 教授

1 はじめに

被災地での救助活動を行う際に、人間の補助として訓練されたレスキュー犬（災害救助犬）が探査を行う場合がある。レスキュー犬は、犬としての特性を生かして人間と協力して被災地の探索を行う。がれきの隙間などの狭い空間や倒壊した建物など人間には踏破困難な環境でも探査可能であり、発達した嗅覚を頼りにした救助活動が可能である。しかし彼らは人間に向けた言語を持たないため、人間はレスキュー犬の行動から彼らが収集した情報を理解しなくてはならない。現状では、レスキュー犬を指揮するハンドラーと呼ばれる人間がレスキュー犬の行動を手動でマーキングしており、その情報を消防などの指揮命令者に口頭伝達している。このレスキュー犬との共同探索の問題点として、トリアージ（緊急度に従った手当の優先順位付け）のための周辺環境情報や、要救助者情報の不足があげられる。

本研究では、レスキュー犬にセンサを装着して得られたデータを用いてレスキュー犬の行動推定を目的とする。本研究の具体的なタスクは、映像だけでなく音声などのデータも活用したマルチモーダル情報を用いたクラス推定である。これにより、レスキュー犬が今何をしているのか明示的に判断することが可能となり、トリアージに必要な情報が整理され、災害救助活動の効率化が期待される。

2 関連研究

動画分類の研究に two-stream CNN [1] がある。これは動画のフレームとフレームから求まる optical flow 画像を個別のネットワークで学習することで動き情報を考慮した動画分類を行っている。

レスキュー犬行動のモニタリングのために、大野、濱田らによって装着型計測・記録装置が開発された [2]。図 1 にレスキュー犬に装着可能な軽量な行動計測ツールを示す。これを着用したレスキュー犬はサイバー救助犬とも呼ばれる。各種センサを用いた計測データを記録し、リアルタイムに映像などのデータを無線配信することが可能である。そのため、レスキュー犬が人の目の及ばない範囲で活動する際にもレスキュー犬の行動やその周辺環境などを把握するのに役立つ。



図 1 装着型計測・記録装置 [2] より引用。

また、Ehsan らによる犬の一人称視点動画からの犬行動予測の研究がある [3]。これは、犬の行動をモデリング

し、犬が次にどのような道をたどり行動するかを予測している。

しかし、これらの研究は犬の行動のモデリングであり、犬の周辺環境の推定などは行っていない。また、入力は動画像のみであり、音声などのデータは利用していない。レスキュー犬の課題には、犬の周辺環境情報や動画像からだけでは判断できない情報の取得が含まれている。例えばレスキュー犬は要救助者を発見するとその場で待機し吠え続けるように訓練されている。このように、動画像データからだけではなく、音声データ、および慣性データ・GPS データなどの情報を複合的に用いてレスキュー犬の状態を判断しなければならない。ただし、本研究では動画像と音声情報のみの提供をうけたため、これらを入力とした犬の行動推定を行う。

音声に焦点をあてた動画分類の研究には Sound Net [4] がある。これと [1] を参考に、本研究では音声識別ネットワークと two-stream ネットワークを統合したアーキテクチャを構築した。

3 データセット

東北大大学の大野らから提供されたレスキュー犬訓練データセットを本研究で整形したものを学習に用いた。提供されたデータは 7 本からなる動画で、犬視点動画、ハンドラー視点映像、第三者視点映像が横並びに結合されており、時間の範囲を指定するように犬行動がラベル付けされている。犬行動は同時刻に複数発生するため、ラベル付けもそのようになっている。ここから犬視点動画のみを切り出し、フレームの静止画像とその直後のフレームから計算した optical flow 画像、および前後 15 フレーム分の長さの音声の 3 データをフレーム毎に抽出した。これを 1/5 の量にサンプリングしたものを学習と評価に用いた。犬行動は 11 種 (bark, cling, command, eat-drink, look_at_handler, run, see_victim, shake, sniff, stop, walk-trot) あり、その一部を図 2 に示す。



図 2 サイバーレスキュー犬訓練データセット see victim クラス。左から静止画像、optical flow 画像、MFCC スペクトラムで可視化した音声である。

4 提案手法

静止画像、optical flow 画像、音声をそれぞれ別のストリームに入力し、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 を提案する。なお、音声信号は MFCC 特徴に変換したものを入力とした。提案手法のアーキテクチャを図 3 に示す。

表 1 各実験結果比較表.

	静止画像	optical flow 画像	音声	bark	cling ⁹	command	eat	handler	run	victim	shake	sniff	stop	walk	全
(1)		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	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)	x	x	2D	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)			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	0.234	0.014	0.123	0.01	0.223	0.356	0.487	0.759	0.692	0.493
(8)			2D	0.577	0.135	0.186	0.066	0.183	0.026	0.433	0.409	0.53	0.779	0.725	0.518

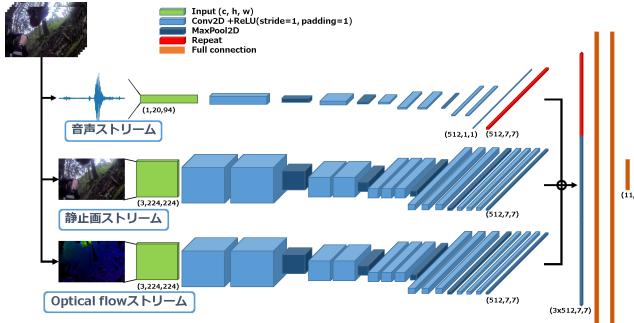


図 3 Sound/image-based three-stream のアーキテクチャ. 音声、静止画像、optical flow 画像それぞれを入力とする 3 つのストリームからなり、それぞれの出力を結合してクラス推定を行っている。

5 実験

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。音声での学習は畠み込み層について 1D と 2D の 2 種類で行った。畠み込み層の次元の差について図 4 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 の畠み込みを用いた。

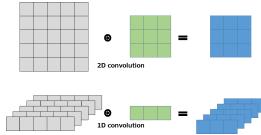


図 4 1D convolution と 2D convolution の詳細。データは同様でも入力の形と処理が異なり、出力の結果も異なっている。

結果のまとめを表 1 に示す。静止画像、optical flow 画像単体では精度が低く、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 ともに精度が高く、静止画像や optical flow 画像と組合せた実験では全体的な精度は低下した。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 畠み込み層で学習したの方のが数値が高い。

これらから、データセットに対する提案手法の有効性が示された。

6 おわりに

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、[4] 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 [5] 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。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、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。

参考文献

- [1] K. Simonyan and A. Zisserman. <https://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.pdf> Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pp. 568–576, 2014.
- [2] 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. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 630–635, 2015.
- [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] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016.
- [5] 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.