レスキュー犬の一人称動画を用いた動作推定

荒木 勇人[1,a)]　井出 佑汰[1]　濱田 龍之介[2]　大野 和則[2,3]　柳井 啓司[1]

概要

被災地での災害救助を補助する犬をレスキュー（災害救助）犬といい, カメラなどの計測装置を装備したレスキュー犬をサイバーレスキュー犬と言う. 本研究では, 犬にとりつけたセンサからサイバーレスキュー犬活動の識別を目的とする. 犬一人称動画像と音声を CNN を用いて動作毎に分類する sound/image-based three-stream CNN の提案と, 提案手法をマルチクラス推定に用いた実験を行なった. Sound/image-based three-stream CNN は動画から得られた静止画像・optical flow 画像・音声を入力とする動画識別ネットワークである. 本提案手法の有効性を示すため,3 つの入力それぞれ単体とその組み合わせパターン（静止画像単体,optical flow 画像単体, 音声単体, 静止画像+optical flow 画像, 静止画像+音声,optical flow 画像+音声）を用いた識別との比較実験を行なった. 結果は,51.8%で提案手法で最も高い精度が得られた. 推定には, レスキュー犬の訓練の様子を撮影したデータセットを用いた. このデータセットは現在も作成中であり, まだ学習のデータ量が十分とは言えない. 一人称視点動画からのマルチクラス推定というタスクと, レスキュー犬訓練データセットの複雑さ（激しい揺れ, ノイズ, マルチラベル）のあいまったチャレンジングなタスクであることを踏まえても不十分な結果であり, 今後さらに手法の改良とデータセットの拡充が必要である.

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. また, ハンドラーによる記録はどうしても主観的になるので客観性に欠け, さらにそれが口頭伝達されることで正確性がより欠落する. 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（図 1）. 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個人の主観に基づくことなく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.

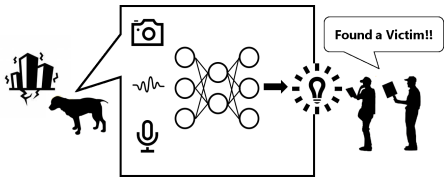

図 1 提案手法の概要. 入力画像から複数データを抽出し, それぞれ別のストリームへ入力する. 各ストリームの出力をもとに, 動画のレスキュー犬行動推定結果を最終出力とする.

2. 関連研究

動画分類の研究に two-stream CNN [6] 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 [4]. 図 2 にレスキュー犬に装着可能な軽量な行動計測スーツを示す. 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.

[1] 電気通信大学 大学院情報理工学研究科 情報学専攻
[2] 東北大学 NICHe
[3] 理研 AIP
[a)] araki-t@mm.inf.uec.ac.jp

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 [2]. 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.

犬一人称の動画像データセットには DogCentric Activity Dataset [3] がある. これは, 背中にカメラを搭載した犬の散歩データセットである. 犬一人称視点動画からの犬行動分類に利用される. しかし, レスキュー犬特有のクラスなどには対応しておらず, 音声データも収録されておらず, シングルクラスで分類されているため本研究では利用しない.

音声に焦点をあてた動画分類の研究には SoundNet [1] がある. 動画から音声と画像を取り出し, 画像を教師データとし, 音声は生徒データとして出力が等しくなるように学習している. この手法は音響シーン分類, 物体分類の標準ベンチマークにおいて教師あり学習の最高精度を達成した. 本研究では音声のみからの行動推定は目的としない. しかし, 音声の意味的情報は動画認識に重要であることが明らかにされている. SoundNet と two-stream CNN 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. SoundNet が音声波形を入力とするのに対し, 本研究では音声から MFCC 特徴を抽出してからストリームへの入力とする点で異なっている. 音声と動画像だけでなく, 加速度と速度情報を用いたマルチモーダルな学習の研究には丹野らの [7] がある. これは, 情報を追加することによっての認識精度の向上や低下を報告しており, 特に音声によって分類精度が向上するとされている. マルチモーダルな深層学習という点で本研究と同様のテーマだが, マルチクラス推定という点で本研究とは異なっている.

3. データセット

学習には, サイバーレスキュー犬訓練データセットを利用した. これは, 訓練されているレスキュー犬に専用の計測スーツを着用させ収集したデータ群である. 東北大学の大野らからその一部を提供された. 提供されたデータは 7 本からなる動画で, 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 11 種 (bark:吠える, cling:

図 2 装着型計測・記録装置 [4] より引用.

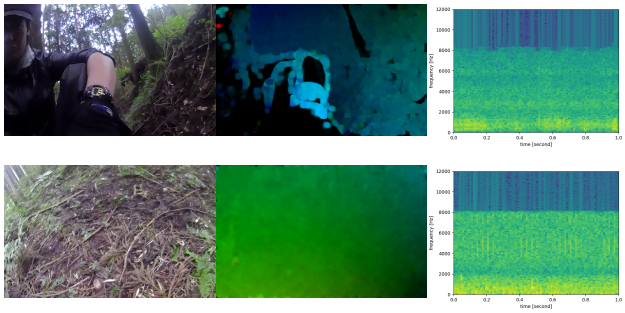

図 3 サイバーレスキュー犬訓練データセット see victim クラス（上段）と stop クラス（下段）. 左から静止画像, optical flow 画像, MFCC スペクトラムで可視化した音声である.

匂いに執着する, command:ハンドラーからの働きかけがある, eat-drink:飲食する, look_at_handler:ハンドラーを見る, run:走る, see_victim:要救助者を発見した, shake:体をブルブルと震わせる, sniff:匂いを嗅ぐ, stop:脚を止める, walk-trot:歩く) あり, その一部を図 3 に示す. 総時間は 57 分 40 秒, 秒間フレーム数は 29.97, 総フレーム数は 103,696 枚である. 分類クラスそれぞれについて時間範囲を指定する形で動画にアノテーションがされており, 同時刻に複数のクラスが重なるためマルチラベルデータとして取り扱う. これを 6fps にサンプリングして整形したものを学習と評価に用いた. クラス毎の出現頻度が 表 1 の値である. 表記のためクラス名を簡略した. 各クラス毎に約 100 回から 9000 回のサンプルがあり数値上では十分である. しかしこれは連続したシーンを含めての数値であり, その実は多様性に欠ける中身となっている. 例えば,run クラスは 98 サンプルあるが, 独立したシーンとしては 3 シーンにすぎず, bark,eat-drink,see_victim,shake クラスを合わせても 100 シーンに満たない.

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.

4. 提案手法

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-

表 1 サイバーレスキュー犬訓練データセット 利用範囲内で計測した出現回数

クラス	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk
出現回数	1744	1127	2439	343	2011	98	1549	239	7719	6384	8764

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図 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）スペクトログラム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 Σ 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.

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

学習は全てにおいて 100 エポック行った. 学習率は 1e-03〜1e-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.

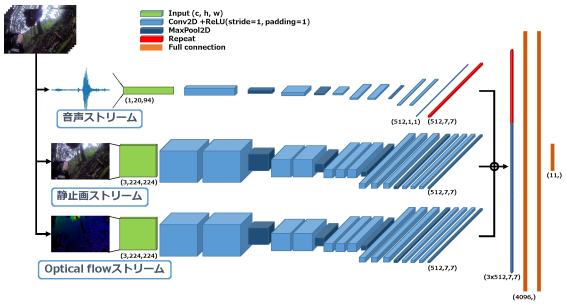

図 4 Sound/image-based three-stream アーキテクチャ. 音声, 静止画像, optical flow 画像それぞれを入力とする 3 ストリームからなり, それぞれの出力を結合してクラスを推定している.

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図 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.

表 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)	○	×	×	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)	×	×	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)	×	×	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)	○	○	×	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)	○	×	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)	×	○	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**

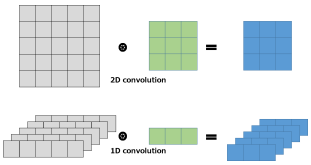

図 5 1D convolution と 2D convolution の詳細. データは同様でも入力の形と処理が異なり, 出力の結果も異なっている.

6. 考察

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, 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. [5] 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, 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.

また音声からの特徴抽出について, 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が残っている. 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. 今回は MFCC スペクトログラムを採用したが,[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. 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.

参考文献

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

[2] Ehsani, K., Bagherinezhad, H., Redmon, J., Mottaghi, R. and Farhadi, A.: Who Let The Dogs Out? Modeling Dog Behavior From Visual Data, *Proc.of IEEE Computer Vision and Pattern Recognition* (2018).

[3] Iwashita, Y., Takamine, A., Kurazume, R. and Ryoo, M. S.: First-Person Animal Activity Recognition from Egocentric Videos, *Proc.of International Conference on Pattern Recognition (ICPR)*, Stockholm, Sweden (2014).

[4] 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).

[5] Minghuang, M., Haoqi, F. and Kris, M. K.: Going Deeper into First-Person Activity Recognition, *Proc.of IEEE Computer Vision and Pattern Recognition* (2016).

[6] Simonyan, K. and Zisserman, A.: Two-stream convolutional networks for action recognition in videos, *Advances in Neural Information Processing Systems*, pp. 568–576 (2014).

[7] 丹野良介, 小澤 暖, 伊藤浩二.：危険運転シーン抽出のためのマルチモーダル深層学習技術, 第 11 回データ工学と情報マネジメントに関するフォーラム (DEIM) (2019).