社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

映像・音・センサー情報の統合によるレスキュー犬の 1 人称行動認識

井出　佑汰†　荒木　勇人†　濱田龍之介††　大野　和則††,†††柳井　啓司†

† 電気通信大学大学院情報理工学系研究科
†† 東北大学 NICHe
††† 理研 API
E-mail: †{ide-y,araki-t,yanai}@mm.inf.uec.ac.jp, ††{hamada,kazunori}@rm.is.tohoku.ac.jp

あらまし　災害が発生した際に被災地では要救助者の迅速な探索や情報収集が必要となる。その際に救援活動を補助するレスキュー犬 (災害救助犬) が参加することがある。レスキュー犬活用の課題として、周辺の環境情報やトリアージ (緊急度に従って優先順位をつけること) のための情報が不足してしまうこととハンドラーの主観によって情報が記録、伝達されてしまうため客観性が損なわれてしまうことが挙げられる。それらの課題を解決するために従来手法として映像、音声を用いたレスキュー犬の 1 人称動画認識手法が存在する。カメラや慣性センサーなどの計測装置をつけたレスキュー犬のことを特にサイバーレスキュー犬といい、サイバーレスキュー犬からは、映像データ、音声データ、センサーデータを得ることができる。従来手法はこれらのデータを対象として扱っている。本研究では、従来手法の改良を試み、主に 2 つの実験を行った。1 つ目は音声の特徴抽出についての実験である。音声の特徴抽出するにあたって最適なネットワークを検証し、その後適切な音声の長さの選択を行った。2 つ目はセンサー情報の特徴抽出についての実験である。音声と同様にして、最適なネットワークを検証し、その後適切なセンサー情報の長さの選択を行った。以上の二つの実験から得られた結果をもとに映像、音声、センサー情報を統合したネットワークを提案手法として従来手法との比較を行い約 7% の精度向上を実現した。
キーワード　1 人称動画認識、Two-Stream CNN、レスキュー犬

Ego-centric Action Recognition of Cyber Rescue Dogs by Integrating Video, Sound and Sensor Information

Yuta IDE†, Tuyohito ARAKI†, Ryunosuke HAMADA††, Kazunori OHNO††,†††, and Keiji YANAI†

† Department of Informatics,The University of Electro-Communications,Tokyo
†† NICHe Tohoku University
††† RIKEN API
E-mail: †{ide-y,araki-t,yanai}@mm.inf.uec.ac.jp, ††{hamada,kazunori}@rm.is.tohoku.ac.jp

1. はじめに

災害が発生した際に被災地では要救助者の迅速な探索や情報収集が必要となる。その際に、レスキュー犬（災害救助犬）の助けを借りることがある。レスキュー犬とは自らが救助活動を行う犬のことではなく、人間にはない高い踏破能力や鋭い嗅覚を活用することにより探索や情報収集の手助けしてくれる犬のことである。レスキュー犬を指示する人のことをハンドラーという。ハンドラーは言葉を持たない犬の代わりにレスキュー犬が集めた情報を手動で記録して、情報を口頭にて災害指令本部に伝えるという役割がある。

人とレスキュー犬が共同で救助活動を行うなかではいくつかの課題が存在する。ひとつは、周辺の環境情報やトリアージのための情報が不足してしまうこと、もうひとつは、現状の情報伝達の方法ではハンドラーの主観的判断による部分が多くなっているため情報の客観性という面が損なわれてしまうということが挙げられる。一刻も早く要救助者を救出しつつ、救助する側も安全かつ迅速に救助活動を行うにあたってこれらの問題を解決することは必要不可欠なことであると言える。

政府による総合科学技術・イノベーション会議が研究開発を促進している ImPACT というプログラムがある。その一環として、タフ・ロボティクス・チャレンジという災害救助を目的

としたロボット開発研究があり、災害救助用サイボーグ犬の開発の足掛かりとしてサイバーレスキュー犬が研究されている。サイバーレスキュー犬の技術的達成目標として” 救助犬の行動と状態の計測・伝送・認識・マッピング (運動・映像・声・生体信号) と制御による、救助活動支援”(注1) を掲げている。それを可能にするために大野、濱田らによって開発されたレスキュー犬の行動を観察するための装着型記録装置がある [8]。この装置には、映像を取得するカメラ、音声を取得するマイク、慣性センサーが搭載されていてリアルタイムでレスキュー犬からの情報を受け取ることができる。サイバースーツを装着したレスキュー犬を図 1 に示す。

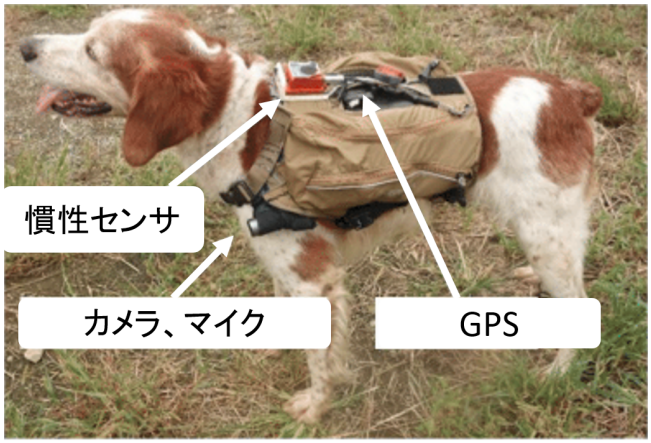

図 1　大野らによって開発されたサイバースーツ

サイバーレスキュー犬から得られた映像、音声を用いたレスキュー犬 1 人称動画認識の手法として [1] [12] が存在する。本研究では、従来手法の改良を試み、主に 2 つの実験を行った。

1 つ目は音声の特徴抽出についての実験である。音声の特徴抽出するにあたって最適なネットワークを検証し、その後適切な音声の長さの選択を行った。

2 つ目はセンサー情報の特徴抽出についての実験である。音声と同様な実験を行った。以上の二つの実験から得られた結果をもとに映像、音声、センサー情報を統合したネットワークを本研究の提案手法とする。

2. 関 連 研 究

本研究では、映像、音声、センサー情報からレスキュー犬の 1 人称動画認識を行う。この章では映像から得られる情報をもとに行動認識をする研究や音声による行動認識の研究などを関連研究として挙げる。動き特徴を考慮して動作認識を行う手法として Simonyan らの研究による、Two-Stream Convolution Networks [10] がある。ネットワークに一つの動画から RGB 画像と OpticalFlow 抽出を行った画像とをそれぞれ分割して入力している。RGB の入力のみでは考慮することができていなかった動き特徴について Oprical Flow も同時に入力することにより動き特徴を考慮できるようになり高性能な動作認識を実現した。音声による動作認識を行う研究として Aytar らによる Sound Net [2] がある。映像データを教師ネットワークに入力し、音声波形を生徒ネットワークに入力してそれぞれの出力分布が近くなるように学習させることで音声認識を行っている。音声認識において画像認識モデルを活用することの有効性を示している。Sound Net では音声の入力は音声波形となっているが、本研究では STFT(短期間フーリエ変換) を利用して音声認識を行う点で異なっている。

犬の 1 人称動画認識の研究として Ehsan らの研究 [4] がある。動画像を入力として犬の行動のモデリングを行い、犬の行動予測をする研究となっている。犬の 1 人称動画像データセットとして DogCentric Activity Dataset [6] があるがレスキュー犬のデータセットとクラスが異なることや音声、センサ情報がないことから本研究では使用しない。1 人称視点動画の認識に音声を組み合わせた研究として Arsha らの Temporal Binding Network [7] がある。1 人称動画認識に音声を組み込むことにより従来手法である映像のみからの動作認識では困難であった動作に対して区別をつけることを可能にしている。そのため、従来手法よりも高い認識精度を達成している。Arsha らは音声の活用が重要であることを主張しており、マルチモーダルな手法の有効性を示している。RGB 画像、OpticalFlow 画像、音声の情報をマルチモーダルに活用した、1 人称動作認識の研究として荒木らの研究 [1] [12] がある。サイバースーツを装着したレスキュー犬から得られたデータセットを題材としている。Two-stream CNN [10] と Sound Net [2] をベースとした音声特徴抽出ネットワークを組み合わせた Sound/image-based Three-Stream Network を提案している。図 2 にて手法の概要図を示す。この研究においても画像情報や音声情報のみで認識する場合よりも画像情報と音声情報を組み合わせた手法のほうが認識精度が向上することを示している。加えてマルチクラスラベル推定となっている。本研究ではこの手法をベースにして荒木らの手法に慣性センサー情報を識別するネットワークを追加して実験を行う。

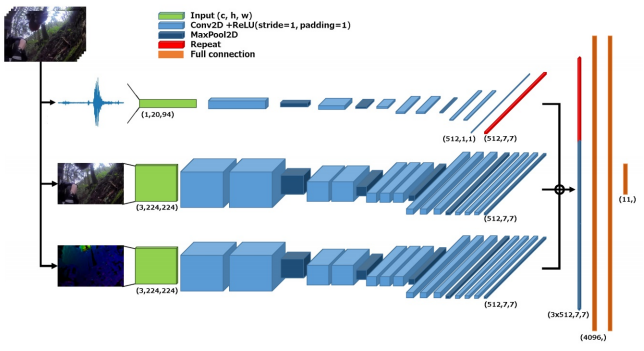

図 2　ネットワークの概要図 ([1] より引用)

3. データセット

本研究は、東北大学の大野らから提供されたレスキュー犬の訓練時に取得されたデータを用いる。このデータは訓練されている最中のレスキュー犬に、サイバースーツを装着させ収集し

(注1) : https://www.jst.go.jp/impact/program/07.html

たデータ群になっている。東北大学の大野らによって現在も作成途中となっていて本研究ではその一部の提供を受けた。そのため本章では今回使用したデータセットについて説明する。提供されたデータには、レスキュー犬 1 人称視点動画、ハンドラー視点動画、第三者視点動画が横一列につながった動画がありこの中でレスキュー犬 1 人称視点動画のみを切り出して使用した。データセットの例を図 3 に示す。この動画をフレーム

図 3　提供されたデータセットの例。左から、レスキュー犬 1 人称視点、ハンドラー視点、第三者視点となっている。

分割して RGB 画像作成し、切り出したフレームとその直後のフレームで計算した OpticalFlow 画像を作成した。音声データは元の動画からフレームごとに切り出した画像に対応するように作成した。ある動画の 1 フレームを F_t として前後 24 フレーム分を音声データの長さとした。音声データを S_t とおくと $S_t = f(F_{t-24}, F_{t+24})$ と表せる。動画がおよそ 30fps であることから、音声データの長さはセンサー情報はおよそ 1.6 秒間になる。Xsens 社製 Mti-300 というおよそ 0.005 秒おきに加速度、地磁気、姿勢、気圧、気温を取得可能なセンサーユニットによって取得したものである。本研究ではこれらのうち角速度、加速度、姿勢のデータを使用する。センサー情報の長さについても音声データの長さと同様にして作成した。荒木らの手法 [1] の使用したデータセット (合計約 55 分) の中には、センサー情報がないものが存在した。そのため、本研究ではセンサー情報のある部分をミニデータセットとして扱い、センサー情報がない部分も含めたものをフルデータセットとして使うことにする。ミニデータセットに含まれるのは 2016 年 7 月 10 日 (約 12 分) と 2016 年 11 月 11 日分 (約 5 分) となっている。どちらのデータセットを扱う際にも、学習データを前半の 8 割、評価データを後半 2 割とした。データセットの取得日を以下の表にまとめる。動作クラスは bark、cling、command、

表 1　データセットの取得日と動画時間

日付	動画時間
2015 年 8 月 1 日	7 分 44 秒
2016 年 7 月 10 日	20 分 12 秒
2016 年 11 月 11 日	4 分 48 秒
2016 年 11 月 27 日 (午前)	5 分 15 秒
2016 年 11 月 27 日 (午後)	7 分 17 秒
2017 年 6 月 19 日	1 分 49 秒

eat-drink、handler(look-at-handler)、run、see-victim、shake、sniff、stop、walk(walk-trot) の全部で 11 クラスとなっている。各動作クラスは動画にアノテーション付けされている。その際には、同じフレームに対して複数の動作クラスが重なっている。そのため本研究で扱うデータはマルチラベルデータとなる。マルチラベルデータの例としては歩きながら臭いを嗅いでいるとなる sniff と walk の 2 つのラベルが付くものや、被災者を見つけて立ち止まって吠えているとなる bark、see-victim と stop の 3 つが付けられているものなどが挙げられる。各データセットにおける各クラスの出現フレーム数を表 2 に示す。

表 2　データセットごとの各クラスの出現フレーム数

クラス	bark	cling	command	eat-drink	handler	run	see-victim	shake	sniff	stop	walk
フル	9299	3631	8035	876	7066	414	7601	1349	16573	31533	45915
ミニ	2281	1936	4903	145	3718	0	1894	405	8208	13194	14154

4.　手　　法

RGB 画像、OpticalFlow 画像、音声特徴を抽出したものをそれぞれネットワークに入力する。各ネットワークに入力後、concat して Fully Connected 層 (FC 層) にて最終的な出力は 11 クラスとなっている。本研究は、マルチクラスラベル推定となっている。提案手法のネットワークの概要を図 4 に示す。

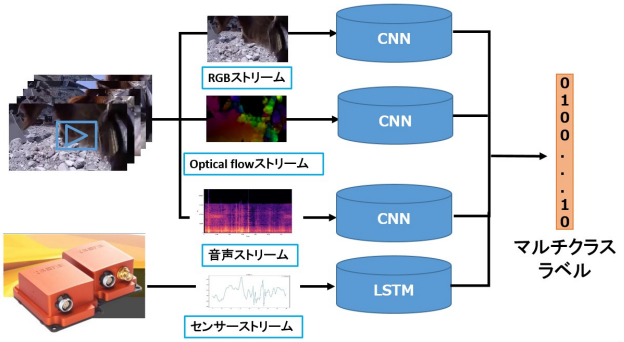

図 4　提案するネットワークモデル概要

4.1　RGB、Optical Flow の特徴抽出ネットワーク

荒木らの手法と同様に ImageNet で pretrain 済みの VGG-16 を使用した。224 × 224 × 3 次元にリサイズしたものを入力とし 2048 次元の出力とした。

4.2　音声の特徴抽出ネットワーク

音声の特徴抽出には ImageNet で pretrained 済みの ResNet-101 を使用した。音声を周波数帯域の STFT(短時間フーリエ変換) を使用して、それを対数スペクトログラム表現に変換することにより 256 × 350 の 2D スペクトログラム行列を生成した。それを 256 × 350 × 1 次元の画像として入力をする。入力した画像を 224 × 224 × 1 にリサイズし convolution 層にて 224 × 224 × 3 にチャネル数を変更後 Resnet-101 に入力し、その後 FC 層にて 2048 次元にしている。

4.3　センサー情報の特徴抽出ネットワーク

センサー情報の特徴抽出には Bidirectional LSTM(Bi-LSTM) [5] を使用した。センサー情報は慣性センサーから得られる x,y,z 軸方向の角速度、加速度、姿勢情報については、最大値、最小値がそれぞれ、1、 -1 となるように正規化した。1.6 秒間のセンサー情報を入力とすると 320 × 9 次元が得られる。

その後、センサー情報 10 個の移動平均を計算したものをネットワークに入力して、レスキュー犬の行動クラスラベルを出力とする。移動平均を求める式は以下のように表される。

$$m_i = \frac{1}{n}(x_i + 0.5 * (x_{i-k} + x_{i+k}) + \sum_{j=1}^{k-1}(x_{i-j} + x_{i+j})) \quad (1)$$

計算後、930(310×3) 次元を Bi-LSTM に入力した。LSTM は隠れ層の次元数を 64、層の数を 2 とした。LSTM に入力後、FC 層で 2048 次元にした。センサー情報のネットワーク図を図 5 に示す。

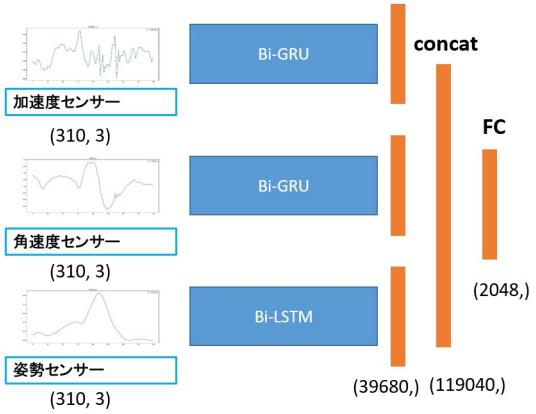

図 5　センサー情報のネットワーク

4.4　映像、音声、センサー情報の統合

上記のネットワークを統合したものが本研究が提案するネットワークとである。各ネットワークから得られる 2048 次元を結合し 2048×4 次元を得る。その後 FC 層を通して最終的な出力は分類クラス数の 11 次元になる。シングルラベルを推定するのであれば SoftMaxCrossEntropyLoss を使用するのだが、本研究ではマルチラベルデータを扱うことからマルチラベル推定を行う。そのため各クラス独立にクラスが生起するとみなして損失関数には MultiLabelSoftMarginLoss を用いた。入力を x、出力を y、クラス数を C とすると MultiLabelSoftMarginLoss は式 2 のように定義される。

$$loss(x, y) = -\frac{1}{C} * \sum_i y[i] * \log((1 + \exp(-x[i]))^{-1}) + (1 - y[i]) * \log(\frac{\exp(-x[i])}{(1 + \exp(-x[i]))}) \quad (2)$$

閾値は 0.5 に設定し、閾値以上の値が出力された場合を推定するクラスとした。以上のことをまとめたネットワークを図 6 に示す。

5.　実　　験

本章では、手法を提案するにあたってどのような実験を行ったのかについて述べ、その考察を行う。実験は音声、センサー情報の特徴抽出にはどのネットワークが最適であるのかについて行い、その後音声の長さについての実験を行った。その後、音声、センサー情報の最適なネットワークと画像の特徴抽出手法との統合した本研究の提案手法と荒木らの従来手法との比較を行う。

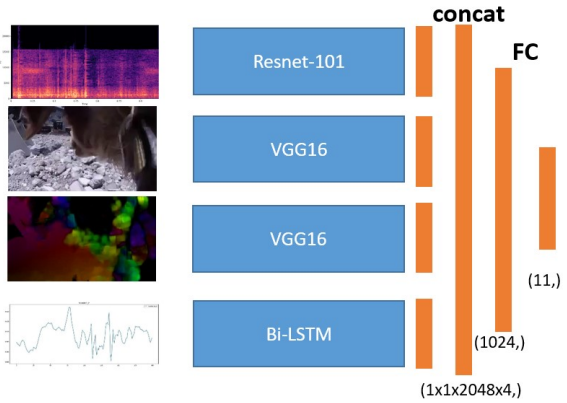

図 6　提案するネットワークモデル

5.1　音声特徴抽出のネットワークについて

ここでは、どのような音声ネットワークが本研究において最適であるのどうかの比較実験を行う。まず、どのようなネットワークで音声特徴を抽出するのかを検証する。その後、音声特徴の適切な長さの選択を行う。

5.1.1　音声特徴抽出ネットワークの選択

荒木らの手法 [1] では音声データからメル周波スペクトラム係数を用いて特徴抽出を行っていた。これにより約 1 秒間分の音声、48kHz から 20×94 の次元を得ている。得られた特徴を Sound Net [2] を参考にした 2D Convolution Network に $20 \times 94 \times 1$ の 1 次元の画像として入力をしている。本研究では、使用されなかった 1D Convolution Network との比較を行う。さらに Arsha らの [7] で使用されていた短時間フーリエ変換での特徴抽出の方法の検討を行う。Arsha ら [7] はます 1.28 秒のオーディオを抽出し、シングルチャンネルに変換後に 24kHz にリサンプリングしている。その後、ウィンドウ長 10 ミリ秒、ホップ長 5 ミリ秒、および 256 の周波数帯域の STFT(短時間フーリエ変換) を使用して、それを対数スペクトログラム表現に変換することにより 256×256 の 2D スペクトログラム行列を生成している。生成した行列の対数を計算したものを入力として特徴抽出を行っている。本研究でも同様な処理を行い、1 秒間の音声から 256×219 の次元を得て実験を行った。スペクトラム行列を可視化したもの例を図 7 に示す。

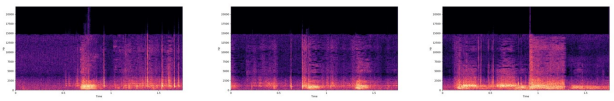

図 7　スペクトラム行列を可視化した例

STFT で生成した行列を入力するネットワークは ImageNet で pretrain 済みの VGG-16、ResNet-50、ResNet-101 を使用した。ミニデータを対象に実験を行った。実験の評価指標に

は Jaccard 係数を用いた。これ以降、本研究において精度が良いとは Jaccard 係数が大きい値のものを指しているものとしすべての実験の評価指標にする。True Positive を TP、False Positive を FP、False Negative を FN とすると Jaccard 係数は式 3 で表される。

$$\frac{TP}{TP+FP+FN} \tag{3}$$

Jaccard 係数で各動作クラスごとの精度と、動作クラス全体での精度を出した。動作クラス全体の精度では各動作クラスの TP、FP、FN をそれぞれ足し合わせたものを使用した。

実験の結果は表 3 の様になった。

表 3　音声のみでの精度の比較 [%]

クラス	bark	cling	command	eat-drink	handler	run	see-victim	shake	sniff	stop	walk	全体
1D	**78.18**	0.00	5.81	0.00	2.48	-	19.60	28.77	17.20	56.54	66.63	40.03
2D [1]	71.12	0.00	**9.86**	0.00	**13.44**	-	**19.92**	7.69	10.78	57.64	65.02	37.66
VGG	64.29	0.00	3.58	0.00	6.51	-	15.25	15.25	17.02	**60.63**	70.48	39.44
ResNet-50	62.79	2.17	3.55	0.00	10.77	-	12.28	35.20	18.78	58.61	68.90	40.99
ResNet-101	66.31	**4.80**	1.76	0.00	11.06	-	13.49	**50.54**	**19.35**	58.66	**71.90**	**42.43**

実験の結果 ResNet-101 に入力して音声特徴を出す手法が最も高い精度を出した。よってこの後の音声の長さを変更した実験では ResNet-101 のネットワークを使用する。

5.1.2　抽出する音声の長さについて

荒木らの手法 [1] では使用する音声は 1 秒間であった。本研究では、さまざまな長さの音声で精度を比較する。音声の長さを 1.0 秒、1.2 秒、1.4 秒、1.6 秒、2.0 秒としたときの音声の精度と荒木らが提案した手法と比較する。

表 4　音声のみでの精度の比較 [%]

クラス	bark	cling	command	eat-drink	handler	run	see-victim	shake	sniff	stop	walk	全体
2D [1]	71.12	0.00	9.855	0.00	**13.44**	-	**19.92**	7.692	10.78	57.64	65.02	37.66
1.0 秒	66.31	4.80	1.758	0.00	11.06	-	13.49	50.54	19.35	58.66	71.90	42.43
1.2 秒	65.24	0.00	2.885	0.00	6.48	-	7.845	55.62	14.74	59.91	66.63	38.97
1.4 秒	79.02	**5.24**	0.25	0.00	12.24	-	15.43	21.35	17.44	59.63	67.92	40.57
1.6 秒	70.51	3.44	**4.19**	0.00	6.37	-	14.96	**82.94**	**21.26**	**63.83**	73.83	**43.74**
2.0 秒	**77.50**	0.00	0.11	0.00	12.87	-	14.14	76.84	11.48	59.73	**74.64**	42.58

実験の結果より、1.6 秒間の長さのときにもっとも高い精度が出たため音声の長さを 1.6 秒にして今後の実験を行う。

5.2　センサー情報のネットワークについて

センサー情報のネットワークについても音声と同様に実験を行った。どのようなネットワークで特徴抽出を行うのかを検証後、適切な長さの選択を行っていく。

5.2.1　センサー特徴抽出ネットワークの選択

センサー情報は、約 0.005 秒おきに情報を取得する。本研究で使用したセンサーは角速度 x,y,z 軸方向、加速度 x,y,z 軸方向、姿勢 x,y,z, 軸方向の 9 次元となっている。1 秒間のセンサー情報を利用すると 200×9 の特徴量を得る。これを 1D Convolution Network、$200 \times 9 \times 1$ 次元を入力として 2D Convolution による特徴抽出を行った。さらに Grave らによる Bidirectional LSTM(Bi-LSTM) [5] に 200×9 次元を入力としての実験も行った。通常の LSTM では順方向の入力のみを学習して時系列予測を行っているのに対して、Bi-LSTM では順方向からの学習に加えて逆方向の学習も行って時系列の予測を行っている。そのため、通常の LSTM よりも高い精度で時系列予測を行うことが可能になっている。本研究においても Bi-LSTM が最も高い精度を得ることができたので、Bi-LSTM を用いて今後の実験を行う。

実験の結果は表 5 のようになった。

表 5　センサー情報のみでの精度の比較 [%]

クラス	bark	cling	command	eat-drink	handler	run	see-victim	shake	sniff	stop	walk	全体
1D	0.00	0.00	3.18	0.00	0.00	-	0.00	0.00	5.949	22.95	50.18	24.57
2D	0.00	0.00	**9.28**	0.00	16.21	-	8.84	0.00	16.16	26.85	47.88	25.26
LSTM	0.00	**2.01**	8.25	0.00	13.06	-	**15.97**	9.63	16.41	28.12	48.01	25.84
Bi-LSTM	0.00	1.43	8.88	0.00	**13.18**	-	15.63	**10.93**	**18.29**	**28.03**	**51.46**	**27.31**

5.2.2　センサーの長さについて

音声の実験と同様にして、センサーの長さを 1.2 秒、1.4 秒、1.6 秒、1.8 秒、2.0 秒と長さを変更して実験を行った。センサーの特徴抽出のネットワークは 5.2.1 節の実験より Bi-LSTM を用いて行った、実験の結果は表 6 のようになった。

表 6　センサー情報の長さを変更したときの精度の比較 [%]

クラス	bark	cling	command	eat-drink	handler	run	see-victim	shake	sniff	stop	walk	全体
1.0	0.00	1.43	8.88	0.00	13.18	-	15.63	10.93	18.29	28.03	51.46	27.31
1.2	0.00	4.06	8.91	0.00	13.14	-	15.84	11.89	16.70	27.00	**51.98**	27.17
1.4	0.00	2.48	10.22	0.00	13.89	-	17.39	11.17	17.86	29.52	49.36	27.14
1.6	0.00	2.47	10.71	0.00	13.94	-	16.45	10.16	16.05	**34.28**	51.12	**28.21**
1.8	0.00	3.78	**11.25**	0.00	**11.79**	-	17.30	**13.99**	**17.02**	33.76	48.62	27.60
2.0	0.00	**6.61**	9.37	0.00	14.78	-	**19.93**	2.67	14.76	34.09	47.21	26.82

センサーの長さが 1.6 秒のときに最も精度が高くなったため、今後の実験においてはセンサー情報を扱う際には長さ 1.6 秒間で扱う。

5.3　提案手法と荒木らの手法との精度比較

ここでは、これまでの実験で得られた音声ネットワークとセンサーネットワークと ImageNet で pretrained 済みの VGG-16 を画像特徴ネットワークを統合した提案手法と荒木らの研究で提案された手法との精度比較を行う。センサー情報が存在する部分についてはミニデータセットでの比較を行った。フルデータセットにおいても、音声ネットワークを Resnet-101 に変更したネットワークとの比較を行う。

5.3.1　ミニデータセットでの比較

提案したネットワークと荒木らの手法との精度比較を行う。荒木らが提案した Three stream network と画像の特徴抽出を荒木らの手法のままにして、音声特徴抽出を ResNet-101 でおこなったもの (VGG-ResNet) と VGG-ResNet にセンサー情報を加えたものを提案手法として精度の比較を行った。LSTM の派生として Chung らの Gated Recurrent Unit (GRU) [3] がある。GRU は LSTM の計算量が膨大であるという問題点を解決するためにモデルを簡素化しつつ性能を同等に維持している。Chung らの研究 [3] では LSTM と GRU どちらのほうが優れているというのでなくタスクに応じてどちらのほうが良いのかが変わることが述べられている。

そのため、センサー情報の特徴抽出ネットワークに使用した Bi-LSTM を Bidirectional GRU(Bi-GRU) に置き換えたものでも実験を行った。ミニデータセットで実験した結果を表 7 に示す。

表 7　手法の比較 [%]

クラス	bark	cling	command	eat-drink	handler	run	see-victim	shake	sniff	stop	walk	全体
荒木ら	16.60	0.26	**9.07**	0.00	14.30	-	36.68	39.32	14.53	45.29	72.04	41.32
VGG-ResNet	62.64	1.17	5.08	0.00	15.74	-	**47.99**	55.68	12.33	57.90	73.42	46.01
本研究 (Bi-LSTM)	73.86	**7.99**	3.29	0.00	14.40	-	41.03	**80.06**	20.36	**64.70**	72.55	**48.05**
本研究 (Bi-GRU)	**78.68**	1.16	3.71	0.00	**18.38**	-	44.77	48.86	**21.49**	59.08	**73.61**	47.51

センサー情報の特徴が効いているのは sniff、stop、になっている。表 7 よりセンサー情報を入れる前と入れた後では、sniff が、8.03% と stop が 6.8% がセンサー情報を入れたときに精度が向上している。また、全体として 2.04% の精度が向上しておりセンサー情報の活用も必須と言える。しかしながら、音声のみの特徴抽出の全体の精度が 43.01%(表 4) に比べてセンサー情報のみの特徴抽出は全体として 28.21%(表 6) となっており十分な精度が得られているとは言い難いためセンサー情報のネットワークの見直す必要があると言える。センサー情報での認識精度の向上により統合したときのさらなる精度向上が期待できる。

5.3.2　フルデータセットでの比較

荒木らの提案した手法と VGG-ResNet についてはフルデータセットでの比較を行った。ここでは、音声の手法の効果についてミニデータセットだけでなくフルデータセットにて検証する。そのためこの実験ではセンサー情報は使用しなかった。実験の結果を表 8 に示す。

表 8　フルデータセットでの比較 [%]

クラス	bark	cling	command	eat-drink	handler	run	see-victim	shake	sniff	stop	walk	全体
荒木ら [1]	57.70	**13.50**	**18.60**	**6.60**	**18.30**	**2.60**	43.40	40.90	**53.00**	**77.90**	**72.50**	51.80
Vgg-ResNet	**64.57**	6.25	4.58	0.00	14.45	0.00	**45.96**	**51.67**	12.59	74.18	71.48	**54.42**

表 8 より映像、音声、音声を統合したときにおいて、荒木らの手法と比較して bark は 7.05% 、shake については 10.77% の精度向上が見られた。フルデータセットにおいて荒木らの手法よりも全体で 2.62% 精度が向上し、音声の手法の有効性を示すことが出来た。

6. おわりに

既存手法である荒木らの手法の音声の特徴抽出手法の見直しを行い、慣性センサーから得られる情報を利用したレスキュー犬 1 人称視点動画の認識を行った。音声の特徴抽出の見直しを行ったことにより従来手法と比べて音声に特徴のある bark、shake のクラスにおいて精度が向上し全体の精度も向上した。さらにセンサー情報の特徴抽出ネットワークの追加により映像、音声のみの場合よりも精度が向上したことからセンサー情報の活用は有効であると言える結果となった。

今後の課題としては、音声のみの場合と比べてセンサー情報のみで十分な精度が出ていないことから、センサー情報のネットワークの見直しが必要であると言える。動画認識の研究として Wu らの Long-Term Feature Banks [11] や Qui らの Learning Spatio-Temporal Representation with Local and Global Diffusion [9] がある。これらの研究では 3D Convolution を活用しながら動画の長期依存性の問題の解決を試みている。より長い範囲の情報を活用することができるようになることで、音声の長さとセンサーの長さを長くした効果をより得られるようになると考えられるためそれらのネットワークを参考にしながらシステムに組み込んでいく必要があるといえる。精度が低い動作クラスについては、センサー情報のネットワークの見直しと画像の特徴抽出ネットワークの見直しを行うことで改善すると考えられるため、改善することが必要になっている。また、センサー情報についての検証がミニデータセットのみを対象としたが、ミニデータセットには run の動作カテゴリが存在しないため、run の動作カテゴリのセンサー情報を含んだデータセットが必要となっている。

謝辞

本研究は JSPS 科研費 15H05915、17H01745、19H04929、17H06100 の助成を受けたものです。

文　献

[1] T. Araki, R. Hamada, K. Ohno, and K. Yanai. Dog-centric activity recognition by integrating appearance, motion and sound. In *Proc. of ICCV Workshop on Egocentric Perception, Interaction and Computing (EPIC)*, 2019.

[2] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Adavances in Neural Infomatoin Processing Systems*, pages 892–900, 2016.

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Adavances in Neural Infomatoin Processing Systems Workshop on Deep Learning*, 2014.

[4] Bagherinezhad H. Redmon J. Mottaghi R. Ehsani, K. and A. Farhadi. Who let the dogs out? modeling dog behavior from visual data. In *Peoc. of IEEE Computer Vision and Pattern Recognition*, 2019.

[5] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Proc of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[6] Takamine A. Kurazume R. Iwashita, Y. and M. S. Ryoo. First-person animal activity recognition from egocentric videos. In *Proc. of International Conference on Pattern Recognition*, 2014.

[7] K.Evangelos, N.Arsha, Z.Andrew, and D.Dima. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[8] Y. Komori, K. Ohno, T. Fujieda, T. Suzuki, and S. Tadokoro. Detection of continuous barking actions from search and rescue dogs' activities data. In *Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 630–635, 2015.

[9] Z. Qiu, T. Yao, C. W. Ngo, X. Tian, and A. T. Mei. Learning spatio-temporal representation with local and global diffusion. In *Peoc. of IEEE Computer Vision and Pattern Recognition*, 2019.

[10] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Adavances in Neural Infomatoin Processing Systems*, pages 568–576, 2014.

[11] C. Y. Wu, C. F. Haoqi, F. K. He, P. K. Ross, and G. . Long-term feature banks for detailed video understanding. In *Peoc. of IEEE Computer Vision and Pattern Recognition*, 2019.

[12] 荒木 勇人 柳井啓司. **レスキュー犬の一人称動画を用いた動作推定**. 画像の認識・理解シンポ ジウム (MIRU), 2019.