

平成30年度 修士論文

レスキュー犬の一人称動画
を用いた動作分類

電気通信大学大学院 情報理工学研究科

情報学専攻 メディア情報学コース

1730010 荒木 勇人

主任指導教員 柳井 啓司 教授

指導教員 橋本 直己 教授

平成31年1月28日

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の活動を識別した。光学センサと音声センサから得られたデータと CNN を用いて動画を分類する Sound based Three-stream 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 based Three-stream 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。本提案手法の精度を示すため、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。マルチクラス推定と言う難しいタスクと、レスキュー犬訓練データセットの複雑さ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、次に繋がる十分な結果が得られたと言える。

目次

第1章 はじめに	1
第2章 関連研究	3
2.1 タフ・ロボティクス・チャレンジ	3
2.1.1 サイバー救助犬	3
2.2 動画認識	4
2.2.1 動作認識	4
2.2.1.1 Two-stream	5
2.2.1.2 3D Convolution	5
2.2.1.3 Modeling Dog Behavior From Visual Data	5
2.2.2 音声分類	7
2.2.2.1 Sound Net	7
2.2.2.2 Audio-Visual Scene Analysis	7
第3章 提案手法	9
3.1 音声と画像を用いた Sound based Two-stream network	9
3.2 音声・静止画像・optical flow 画像を用いた Sound based Three-stream network	10
第4章 データセット	12
4.1 DogCentric Activity Dataset (DCAD)	12
4.2 サイバーレスキュー犬 訓練データセット	13
4.2.1 分類クラス詳細	13
4.2.1.1 bark	13
4.2.1.2 cling	14
4.2.1.3 command	14
4.2.1.4 eat-drink	14

4.2.1.5	look at handler	14
4.2.1.6	run	14
4.2.1.7	see victim	14
4.2.1.8	shake	14
4.2.1.9	sniff	15
4.2.1.10	stop	15
4.2.1.11	walk-trot	15
4.2.2	データ整形	15
4.2.2.1	動画整形	15
4.2.2.2	クリップセット	16
4.2.2.3	フレーム毎切り抜き画像セット・optical flow 画像 セット	18
4.2.2.4	音声セット	18
第 5 章 実験		22
5.1	予備実験：クラス分類	22
5.1.1	DCAD 動画像平均画像クラス分類	22
5.1.2	レスキュー犬訓練データセット動画像平均画像クラス分類 .	23
5.1.3	オプティカルフロー動画平均画像クラス分類	23
5.2	マルチラベル推定	25
5.2.1	静止画像からのマルチクラス推定	25
5.2.2	optical flow 画像からのマルチクラス推定	25
5.2.3	音声からのマルチクラス推定	26
5.2.3.1	1D Convolutional network	26
5.2.3.2	2D Convolutional network	26
5.2.4	Two-stream network	27
5.2.5	Sound based Two-stream network	28
5.2.5.1	音声データと静止画像からのマルチクラス推定 .	28
5.2.5.2	音声データと optidcal flow 画像からのマルチクラ ス推定	28
5.2.6	Sound based Three-stream network	28
5.3	考察	29
第 6 章 まとめ, 今後の課題		31

第1章

はじめに

被災地での救助活動を行う際に、訓練されたレスキュー犬（災害救助犬）が人間の補助として探査を行う場合がある（図1.1）。災害救助犬を育成し、現場に派遣する団体は日本国内に複数存在し、必要に応じて現場に派遣される。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。レスキュー犬にはがれきの隙間などの狭い空間、倒壊した建築物など人間には踏破困難な環境でも探査可能であったり、またその発達した嗅覚を頼りにした探査が可能である。このように、人間では探査が困難あるいは不可能な環境においても人間の能力をレスキュー犬が補うことで効果的な救助活動が期待される。しかし、彼らレスキュー犬は人間に向けた言語を持たない。そのため、人間はレスキュー犬の行動をよく観察し、彼らが収集した情報を彼らの様子から推察、理解しなくてはならない。現状では、レスキュー犬を直接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犬の周辺環境の情報収集と理解に努めている。収集された情報は消防などのハンドラーらを統括する指揮命令者に口頭伝達され、現場の把握に活かされる。このレスキュー犬と人間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）のための災害現場周辺環境情報や、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らるので客觀性に欠け、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。レスキュー犬によって収集された情報を個人の主觀に基づくことなく分類し、整理された情報を共有できれば災害救助活動の効率化がより期待される。

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用いてレスキュー犬の行動を分類すること目的とする。深層学習を用いた画像識別にある既存手法を予備実験として行った。予備実験をもとに、動画からのレスキュー犬行動分類

を行う。本研究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な動画分類である。本研究により、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。



図 1.1: 被災地におけるレスキュー犬らの救助活動 [1] より引用

第2章

関連研究

本研究では犬の一人称視点動画からの犬の活動分類を行う。人間のライフログとしての一人称動画の分類や、車載映像からの車の行動推定、第三者視点での動画分類、音声を用いた動画分類などについて紹介し、本研究との関連を述べる。

2.1 タフ・ロボティクス・チャレンジ

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、“革新的研究開発推進プログラム ImPACT”というプログラムがある[?]。“ImPACTは研究開発を促進し、持続可能な発展性のあるイノベーションシステムの実現を目指したプログラム”であり、複数の研究開発プログラムを包括している。タフ・ロボティクス・チャレンジはそのプログラムのうちの一つであり、遠隔自律ロボット、屋外ロボットサービス事業の実現を目指したプログラムである。このプログラムでは首都圏直下型地震などを想定し、刻々と変化する厳しい環境下でも実用性を保つ災害救助を目的としたロボットの研究開発が行われている。倒壊家屋や配管内を探索するロボット、悪天候でも飛行するドローンなどを用いての計測や認識、マッピング、活動支援などが達成目標として掲げられる。

2.1.1 サイバー救助犬

サイバー救助犬の研究はタフ・ロボティクス・チャレンジの一つである。災害救助用サイボーグ犬の開発を見据え、その足がかりとして研究されている。サイバー救助犬の技術的達成目標は“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”とされており、

レスキュー犬の行動をモニタリングするために、濱田、大野らによって装着型計測・記録装置が開発された [2]。図 2.1 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。サイバー救助犬は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の無線配信が可能である。そのため、人の目の及ばない範囲でレスキュー犬が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が把握可能である。



図 2.1: 装着型計測・記録装置 [2] より引用

2.2 動画認識

犬一人称視点映像の動きや音声の特徴は、レスキュー犬の周辺環境を知るための重要な手掛かりの 1 つである。レスキュー犬の一人称動画に限らず、動画から特徴を取得してその内容を分類する類の研究は行われている。

2.2.1 動作認識

映像から動き特徴を抽出する手法は大きく分けて 2 つある。1 つはあらかじめ動画を複数枚の画像に分割してから特徴量を抽出する手法である。もう 1 つは動画から直接特徴量を抽出する手法である。前者は既存の画像認識の技術を簡単に流用でき、入力データが比較的小さいので学習コストが低い。対して後者はフレー

ム間の情報を考慮できるが、動画を直接入力データとするため学習コストが非常に高い。

2.2.1.1 Two-stream

事前に動画から静止画を切り出してから特徴を抽出し学習する手法としては Simonyan らによる Two-stream convolutional networks がある [3], [4]。これは、1つの動画から通常の RGB 画像と optical flow 画像を抽出し、それぞれを入力とする個々のネットワークを学習することで動き情報を考慮して動画を分類する手法である。図 2.2 に Two-stream convolutional のネットワーク構造を示す。

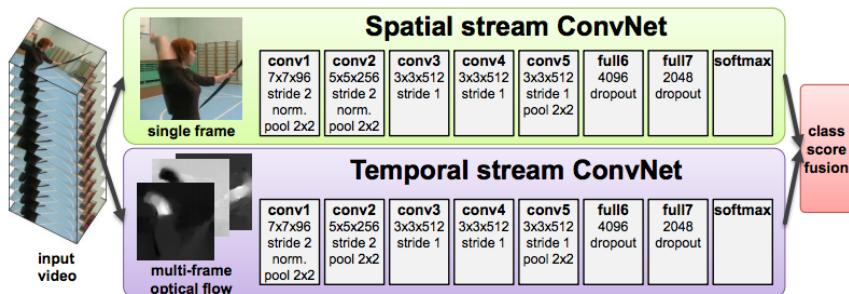


Figure 1: Two-stream architecture for video classification.

図 2.2: Two-stream convolutional networks アーキテクチャ ([3] より引用)。切り出した RGB 画像と optical flow 画像を個々のネットワークに入力し、出力を合わせている。

2.2.1.2 3D Convolution

動画から直接特徴を抽出して学習する手法としては Tran らによる 3D Convolution がある [5]。画像に対して 2 次元であったフィルターを 3 次元形状に拡張することで、縦横の空間以外である時間方向への広がりを持って特徴抽出が可能になった。図 2.3 に 3D Convolutional の詳細を示す。

2.2.1.3 Modeling Dog Behavior From Visual Data

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 [6]。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。

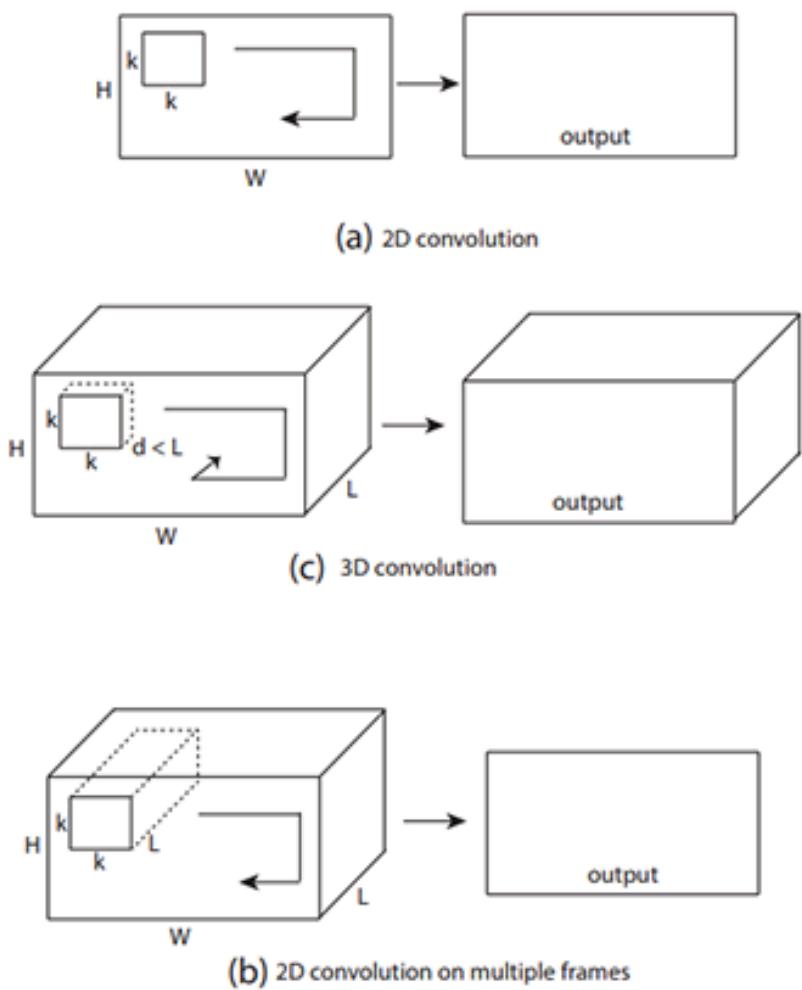


図 2.3: 3D Convolution の詳細 ([5] より引用)。2D Convolution では縦横方向の畳み込みを行っており、3D Convolution では加えて時間方向への畳み込みを行っている。

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。本研究は動画像と音声からなるマルチモーダルな情報を入力とした犬の行動の分類を目的としている。

2.2.2 音声分類

音声と画像から特徴を抽出する研究には以下のようなものがある。

2.2.2.1 Sound Net

音声をクラス分類する研究として Ayter らによる Sound Net がある [7]。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。図 2.4 に Sound Net のネットワーク構造を示す。

2.2.2.2 Audio-Visual Scene Analysis

音声と動画を紐づけて、その関係を明らかにする研究として Owens らによる Audio-Visual Scene Analysis がある ??。映像内の音源特定、音声からの動作認識、複数の話者が個々の画面にいる際の話者の特定を行なっており、音声と映像の関連性を示している。具体的な図を 2.5 に示す。

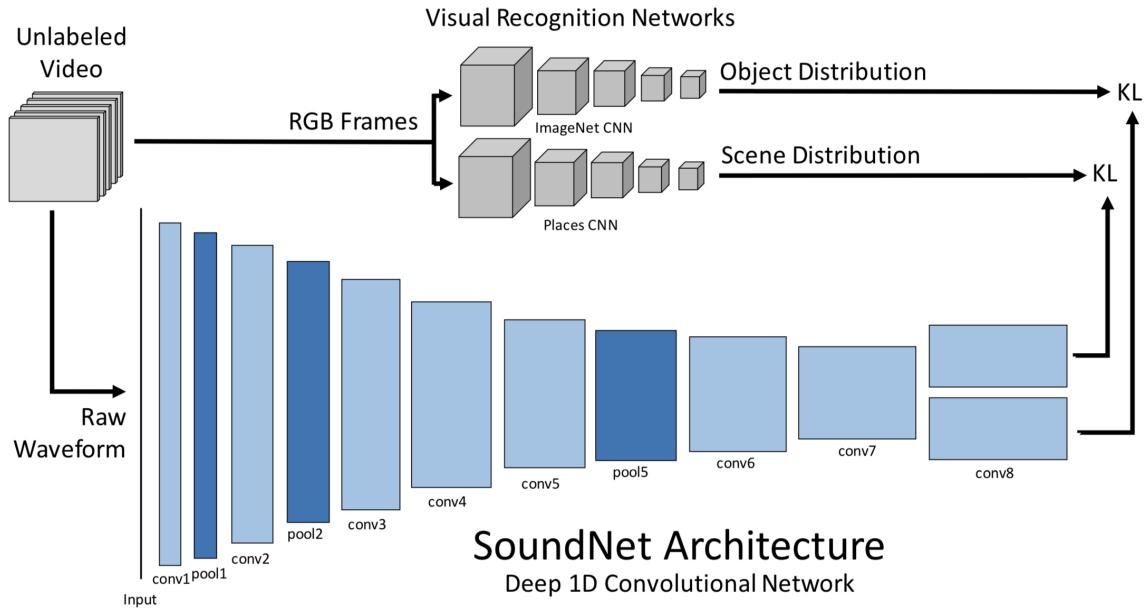


図 2.4: Sound Net のアーキテクチャ ([7] より引用) . 動画から映像と音声を切り分け、音声に対して 1 次元の畳み込みを行なっている .

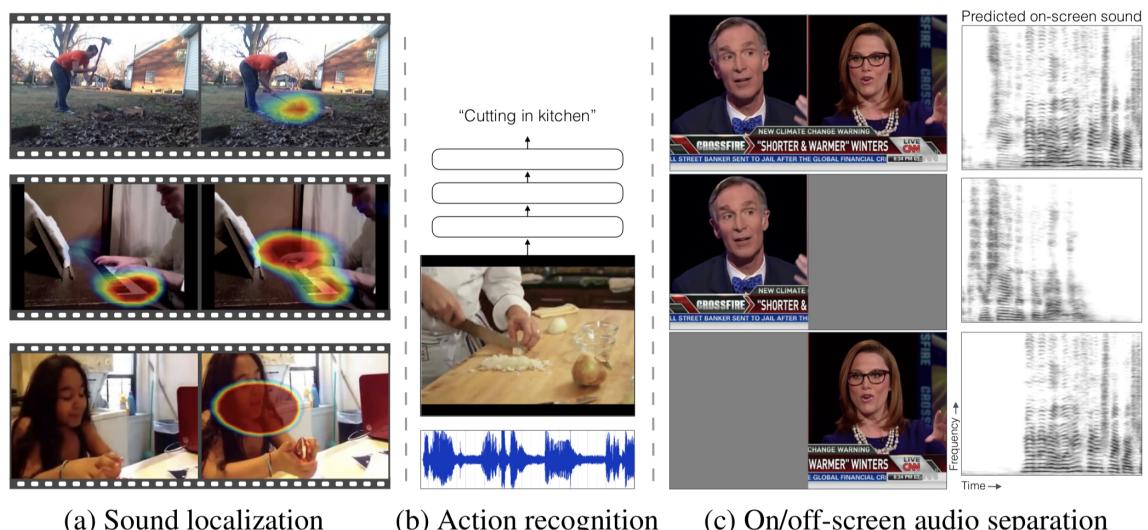


図 2.5: Audio-Visual Scene Analysis ([8] より引用) .

第3章

提案手法

本研究では犬の行動推定のために、動画像・音声のマルチラベル分類を行った。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) からメル周波数ケプストラム係数 (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。

分類はフレーム毎に行った。

3.1 音声と画像を用いた Sound based Two-stream network

1 つ目の提案手法として、音声を用いた Two-stream network を提案する。既存の Two-stream network を改造し、静止画と音声からの動画分類の手法である。Two-stream network と違い音声を用いるため、これを Sound based Two-stream network と呼称する。Sound based Two-stream network のアーキテクチャを図 3.1 に示す。本研究ではこのネットワークをクラス分類ではなくマルチクラス推定に用いる。

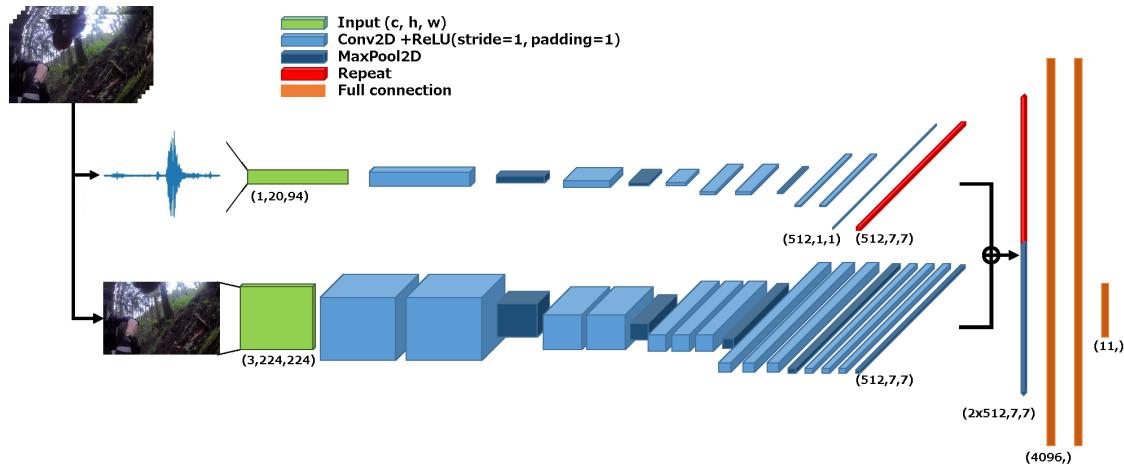


図 3.1: Sound based Two-stream (提案手法) のアーキテクチャ . $(3, 224, 224)$ 次元の画像と $(1, 20, 94)$ 次元を音声をそれぞれ別のネットワークに通し , 得られた特徴を結合したのち FC レイヤを通しクラス数と同じ次元の出力を得る .

3.2 音声・静止画像・optical flow 画像を用いた Sound based Three-stream network

2つ目の提案手法として , 音声 , 静止画像 , optical flow 画像の 3つの情報を用いた Sound based Three-stream を提案する . Sound based Two-stream network に , 画像を入力とするネットワークを加えた 3つの stream を組み合わせている . Sound based Two-stream network のアーキテクチャを図 3.1 に示す .

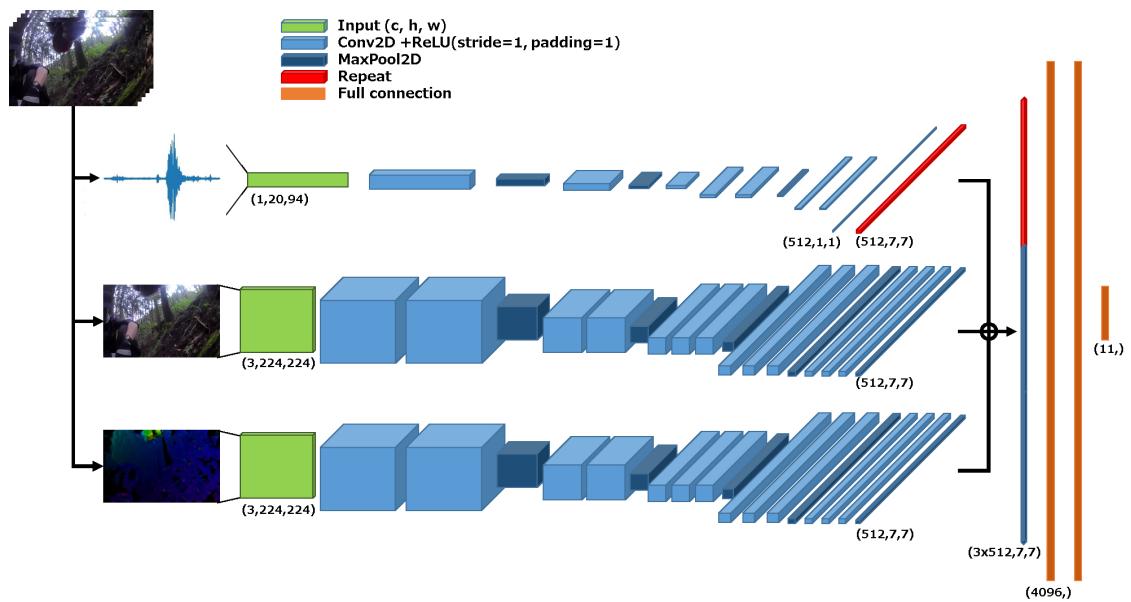


図 3.2: Sound based Three-stream (提案手法) のアーキテクチャ。 $(3, 224, 224)$ 次元の画像と $(1, 20, 94)$ 次元を音声をそれぞれ別のネットワークに通し、得られた特徴を結合したのち FC レイヤを通してクラス数と同じ次元の出力を得る。

第4章 データセット

既存の公開されている犬一人称視点動画データセットに DogCentric Activity Dataset(DCAD) がある。本研究ではレスキュー犬向けにラベル付けされたレスキュー犬訓練動画が必要であるため、本実験ではレスキュー犬の訓練動画を用いた。訓練動画を用いる前に、犬一人称視点動画から行動分類が可能かどうかを確認するため簡易な予備実験を行なった。予備実験には DCAD を用いた。

また、本実験には現在作成中のサイバーレスキュー犬の訓練データセットを用いた。

4.1 DogCentric Activity Dataset (DCAD)

4頭の犬の背中に GoPro カメラを取り付けて散歩をした動画を单一クラス分けしたデータセット 図 4.1。動画は 320 x 240 解像度、48 frames per second で撮影されている。散歩する地域やコースは犬毎に異なり、アノテーションはそれぞれの犬に同じラベルのアクティビティをラベル付けしている。アクティビティは 10 クラス（横断前の待機: Car，水分の摂取: Drink，手渡しでの食事: Feed，左を向く: Look at left，右を向く: Look at right，人間が犬を撫でる: Pet，ボールで遊ぶ: Play with ball，身体をブルブルと振る: Shake，何かの匂いを嗅ぐ: Sniff，歩く: Walk）あり、それぞれ合わせて 209 クリップになる 表 4.1。

表 4.1: DogCentric Activity Dataset 内訳

Activity	Car	Drink	Feed	Left	Right	Pet	Ball	Shake	Sniff	Walk
Clips	26	10	25	21	17	25	14	19	27	25

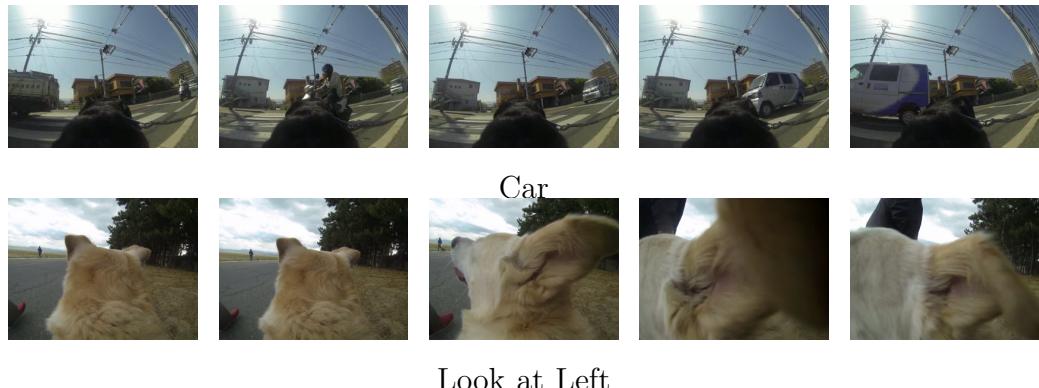


図 4.1: DogCentric Activity Dataset

4.2 サイバーレスキュー犬 訓練データセット

サイバーレスキュー犬訓練データは，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。現在も作成中であり，完成していないデータを含めた全てを使用することは困難である。そのため，本研究ではその一部の提供されたデータをこれとして取り扱うものとする。これは約2分から20分の7本の音声付き動画からなり，犬の一人称視点動画に加えてハンドラー視点動画，レスキュー犬とハンドラーを映す第三者視点動画が含まれる。本研究では犬の一人称視点動画のみを用いて推定を行う。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103696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，複数のクラスが同時刻に重なるためマルチラベルデータとして取り扱う。本研究ではこのデータを整形したものをサイバーレスキュー犬訓練データセットとして実験を行なった。動画は犬一人称視点のみを用いたが，犬一人称，ハンドラー視点，第三者視点毎にそれぞれラベル付けされたアノテーション情報に関しては全てを用いて学習を行った。以下にデータにラベル付けされたクラスと，整形したデータについて詳細を述べる。

4.2.1 分類クラス詳細

4.2.1.1 bark

被災者を発見し，かつ吠えている状態。わかりやすい音声的特徴があり，固有の画面揺れが生じる。

4.2.1.2 cling

臭いに対し、鼻を近づけ嗅いでいる状態。sniff のより詳細な状態であり、cling がラベル付される際は sniff と必ず重複する。

4.2.1.3 command

ハンドラーからの働きかけのある状態。待て/行け等の口頭指示、褒め、指差し指示など状況が多様。

4.2.1.4 eat-drink

何かを食べている/飲んでいる状態。訓練において被災者発見に対する成功報酬に餌が与えられる他、草を食む、地面/川の水を飲むなど状況が多様。以下、eat と表記する。特筆がない場合、eat という表記は eat (食事) と drink (水分攝取) の両方の意、drink という表記は水分攝取の意のみ表現する。

4.2.1.5 look at handler

犬がハンドラーを見ている状態。以下、handler と表記する。

4.2.1.6 run

走っている状態。walk-trot と比較すると、画面に浮遊感があり、揺れや音が激しい。

4.2.1.7 see victim

カメラに被災者が映った状態。以下、victim と表記する。

4.2.1.8 shake

犬が激しく体を震わせている状態。振動に合わせてカラカラカラとカメラの揺れる音がする。

4.2.1.9 sniff

臭いを嗅いでいる状態。探査に対するやる気などを測る一つに指標になる。地面などに鼻を近づけている状態だけでなく、浮遊臭を嗅いでいる際も含む。

4.2.1.10 stop

足を運んでいない状態。その場での足踏みは含む。方向転換は含まない。画面の動き情報が少なく特徴的である。

4.2.1.11 walk-trot

歩いており、run ではない状態。以下、walk と表記する。

4.2.2 データ整形

本研究では、提供されたサイバーレスキュー犬訓練データを整形し、クラス毎に短く切り出したクリップセット/フレーム毎に切り出した画像セット/フレーム毎に計算した optical flow 画像セット/フレームに合わせて一定時間毎に切り出した音声セットをそれぞれ作成した。データを整形して複数セットを作成したのち、さらにラベルのないフレームなどを排除してから、動画に戻した際に 6fps となる量に学習データを間引いた。30fps 近い動画をフレーム毎に学習すると、直近のフレームに対して過度に反応する恐れがあるため、この処理で過学習を防ぐ狙いがある。整形・サンプリングによって、学習および評価に使った総フレーム数は 14581 枚となった。この範囲でラベル付けされた回数をクラス毎に 表 4.2 に示す。静止画のみでの音声付き動画のクラス特徴の表現は困難であるため、整形した動画的特徴、音声的特徴とをそれぞれ 図 4.5 に示す。

4.2.2.1 動画整形

提供されたデータ映像は犬一人称視点映像、ハンドラー視点映像、第三者視点映像が横並びに結合（図 4.2）されていたため、これは犬一人称視点映像のみを切り出した。また、提供された 7 本の動画にはハンドラー視点がないものや、加えて、結合時にノイズが入ったと思われるものが混在したもののがそれ存在したため画像サイズが不揃いであった。ノイズの入った第三者視点のない動画につい

て、図4.3にその例を示す。これについて、ノイズ部分は切り取ったが比率は揃えなかった。ノイズは学習に悪影響を与えること、微細なアスペクト比のずれは学習の際に吸収されること、アスペクト比をそろえる際に必要性の有無を判断できない情報が切り取られることをその判断理由とする。このようにして整形した動画を V とする。



図4.2: 提供時のデータ。左から、犬一人称視点、ハンドラー視点、第三者視点の動画が結合されている。合わせて1280x240pixelで、犬一人称視点だけ切り出しても通常カメラで撮影される比率にはならない。

4.2.2.2 クリップセット

V からラベル毎にクリップ群を切り出すと、アノテーションが单一のフレームと複数被るフレームが同じクリップを介して存在する。ただし、同じ箇所から別々のラベルを切り出した際に全く等しく被ることはない。本研究において、クリップ群を作成する目的はそのフレーム間の平均画像を作成することにある。フレーム間の平均をとる際にそのフレームの多少の前後で結果が異なるため、その差異の



図 4.3: 犬一人称視点 , 第三者視点の結合された動画 . ハンドラー視点がなく , 犬一人称視点の動画の画面左端一帯に黒いノイズが入っているパターン . ノイズを合わせて 1280x360pixel で , 犬一人称視点を切り出すと比率は 16:9 だがノイズを取り除くと比率は崩れる .

表 4.2: サイバーレスキュードッグ 訓練データセット 利用範囲内の出現回数

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

学習を期待して別クリップに同じフレームが重複して現れる問題を無視した。このクリップセットから計算したフレーム間平均画像を図 4.4 に示す。



図 4.4: クリップセットから計算したフレーム間平均画像。

4.2.2.3 フレーム毎切り抜き画像セット・optical flow 画像セット

フレーム毎に切り出した画像には、そのフレームの時間点にアノテーションされているラベルをそのまま用いた。

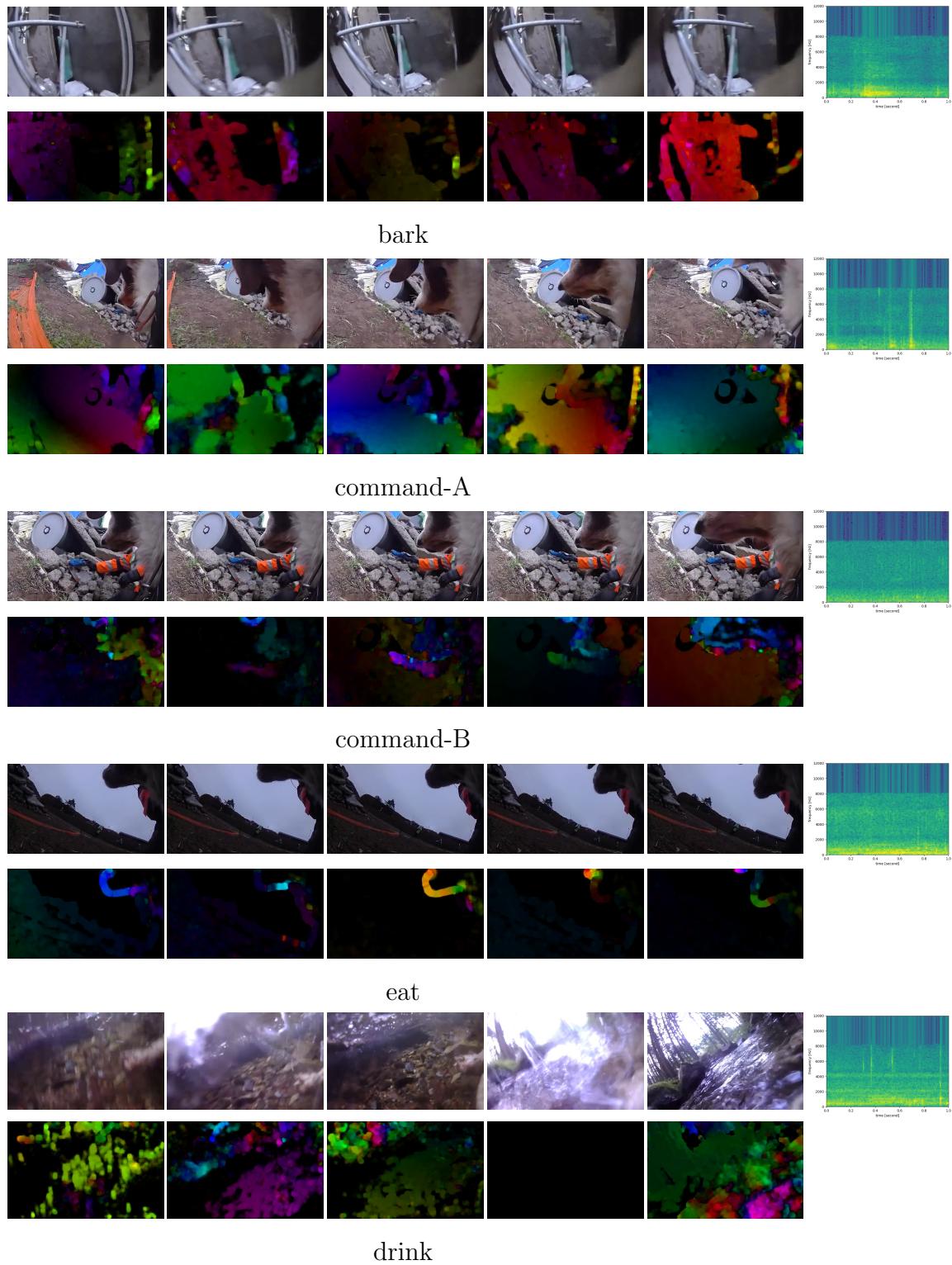
optical flow の計算には Python のオープンソースライブラリである OpenCV を用いた。なお、利用したバージョンは python3.6, OpenCV3.3.1 である。元の RGB 動画のフレーム F_t と F_{t+1} から求まる optical frame 画像を O_t とした際に、 O_t には F_t と同じラベルを紐付けた。

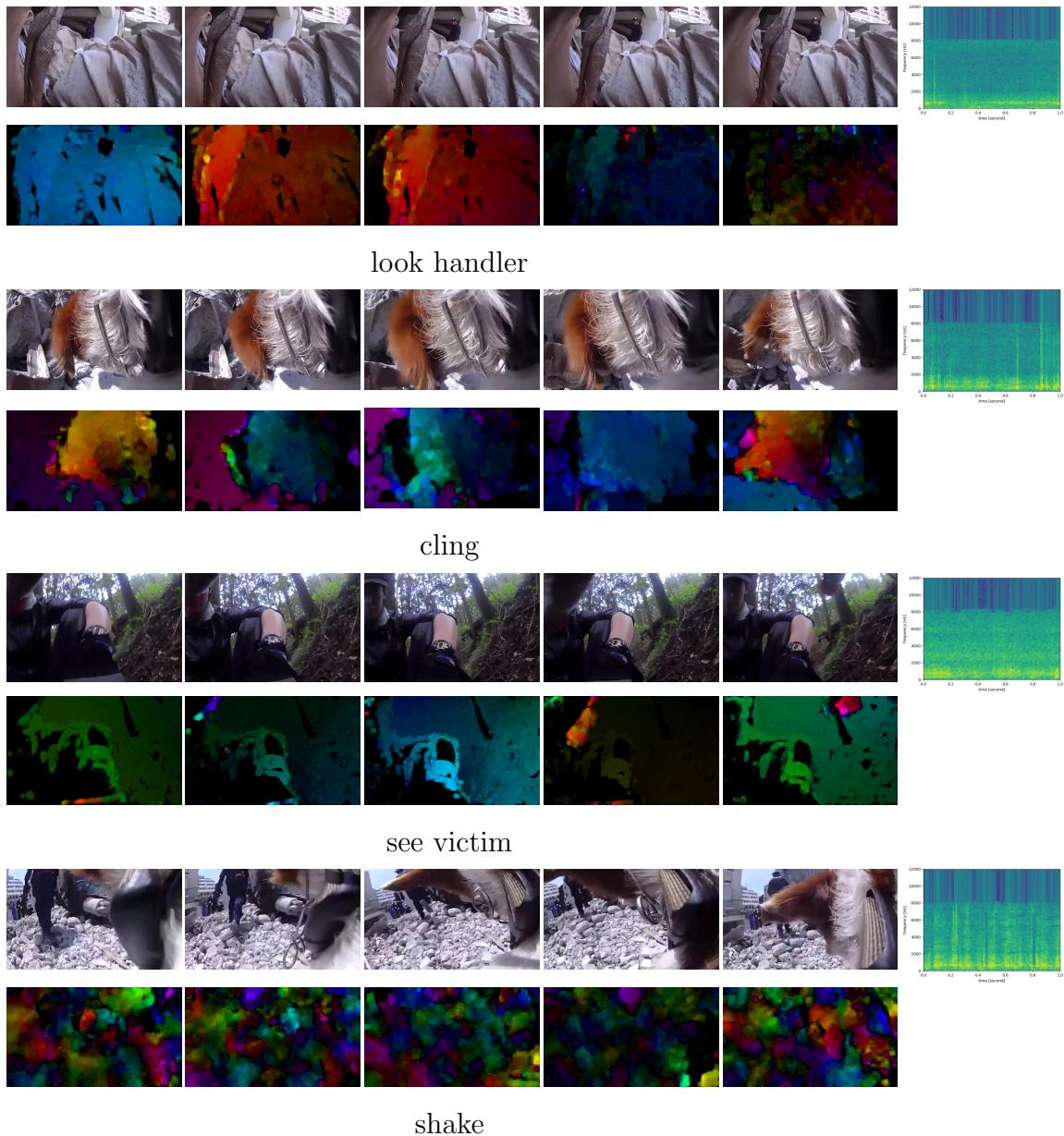
4.2.2.4 音声セット

フレーム毎切り抜き画像セットに対応する音声を切り出した。音声データ S_t は、 F_t を中央において前後 15 フレーム分の長さとした。

$$S_t = f(F_{t-15}, F_{t+15})$$

動画は 29.97fps であるため、31 フレームのこれは約 1 秒に当たる。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すると、1 秒は短すぎず現実的で取り扱いやすい時間であり、特徴を取り出すにもコストが低いため実験にも適している。





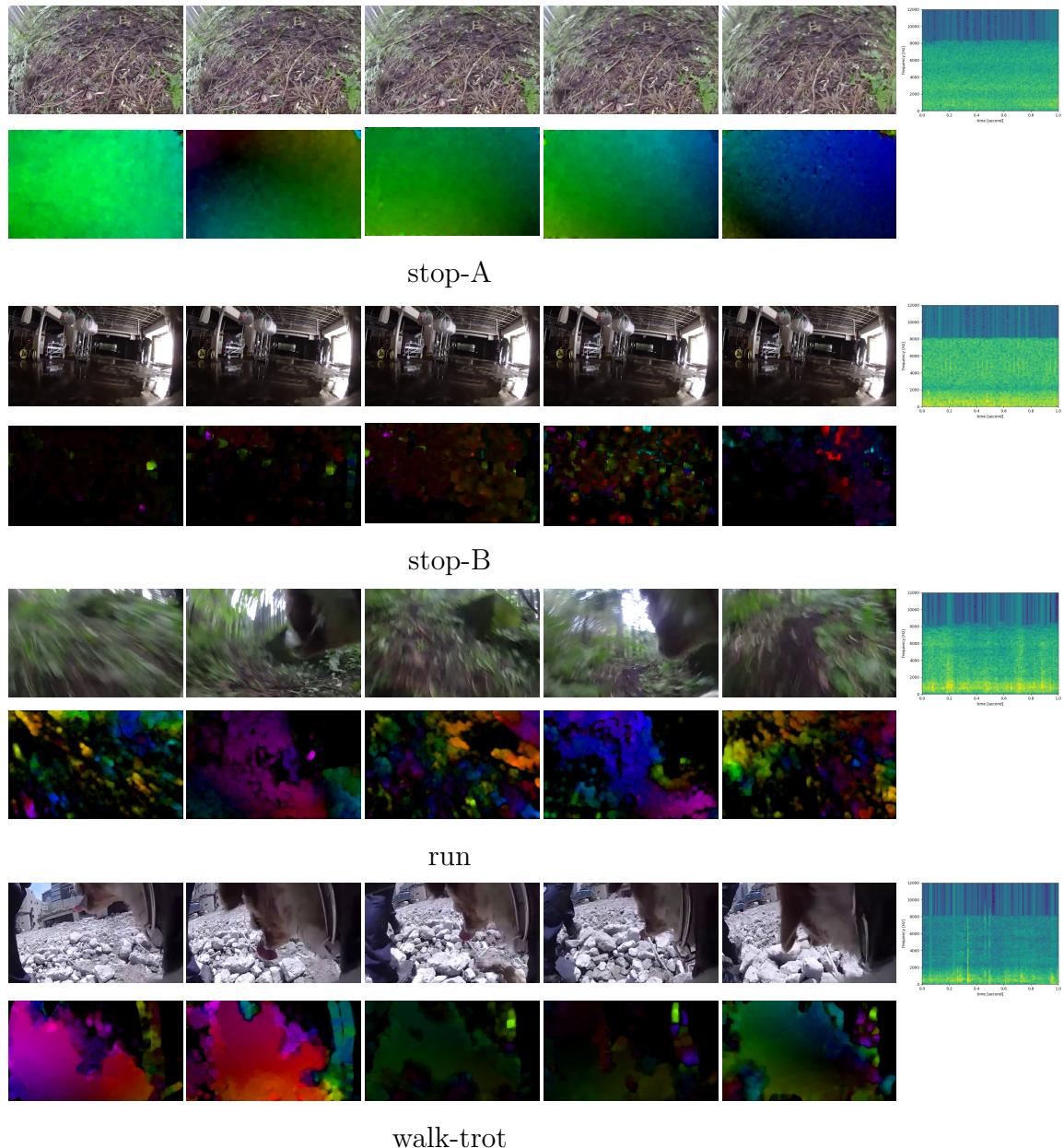


図 4.5: サイバーレスキュー犬訓練データセット

第5章

実験

予備実験を含め、行なった実験は9種類ある。

予備実験として、DCADとレスキュー犬訓練データセットの2種類のクラス分類を行なった。

本実験では、静止画像からのマルチラベル推定、optical flow画像からのマルチラベル推定、音声データからのマルチラベル推定2種、Sound based Two-stream networkを用いた音声データと静止画像からのマルチクラス推定、同じく Sound based Two-stream networkを用いた音声データとoptical flow画像からのマルチクラス推定、そして Sound based Three-stream networkを用いた音声データと静止画像とoptical flow画像からのマルチクラス推定の7種類を行なった。

5.1 予備実験：クラス分類

DCADとレスキュー犬訓練データセットについて、それぞれクラス分類を行なった。DCADはクリップ毎にフレーム間の平均をとり、画像として扱ってVGG16のpretrained modelを用いてfinetuningを行なった。レスキュー犬訓練データセットは動画をラベル毎に切り出して短いクリップ群を作り、そのクリップ毎に同様にフレーム間の平均を取った画像を作成しVGG16のpretrained modelを用いてfinetuningした。

5.1.1 DCAD 動画像平均画像クラス分類

予備実験の結果を(図5.1に示す。分類率は64.3%であった。全般的に、データの多いクラスは精度が高い傾向にあるが、データの少ないクラスは精度が低い傾

向にある。加えて、*Car* クラスは道路の進行方向に対して垂直に待機している 10 クラスの中で特殊なクラスであり、車などの写ったフレームの影響で分類精度が上昇していると考えられる。*Feed* クラス、*Pet* クラス、*Play_with_ball* クラスは、それぞれフレーム内を人間が占める割合が多いクラスと言え、そのため混同が起こりやすいと考えられる。

	Car	Drink	Feed	Left	Right	Pet	Ball	Shake	Sniff	Walk
	Car	6	0	0	0	0	0	0	0	0
	Drink	0	1	0	0	0	0	0	2	1
	Feed	0	0	1	0	0	0	0	0	0
	Left	1	0	0	1	0	0	0	2	0
	Right	0	0	0	0	0	0	0	1	0
	Pet	0	0	1	0	0	3	1	0	2
	Ball	0	0	0	0	0	0	5	0	0
	Shake	0	0	0	0	0	0	1	1	2
	Sniff	0	0	0	0	0	0	1	0	3
	Walk	0	0	0	0	0	0	0	0	6

図 5.1: VGG16 pretrained model と DCAD による finetuning の結果

5.1.2 レスキュー犬訓練データセット動画像平均画像クラス分類

レスキュー犬訓練データセットでの予備実験の結果を図 5.2 に示す。

データ数の多い walk クラスや stop クラスだけでなく、shake クラスや eat クラスなどのデータ数の少ないクラスも大まかに分類できていることが分かる。この結果によって、レスキュー犬訓練データセットからクラス分類・推定が可能であることが示された。

5.1.3 オプティカルフロー動画像平均画像クラス分類

動画像のフレーム間の平均を取った手法と同様に、クリップ毎に optical flow の平均画像を作成し VGG16 の pretrained model を用いて finetuning を行なった。結果を図 5.3 に示す。

やはりデータ数の影響を受けているものの、通常の動画のフレーム平均画像とは異なる傾向が得られた。この結果によって、optical flow 画像から得られる特徴の有用性が示された。

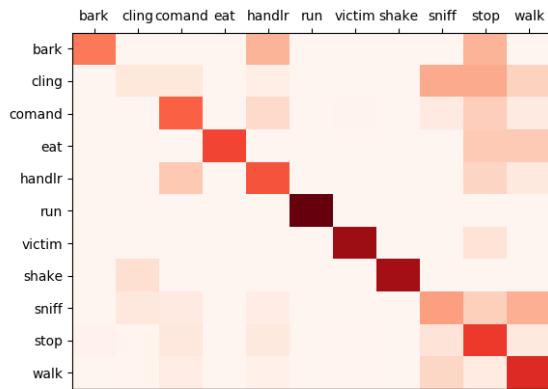


図 5.2: ResNet pretrained model とレスキュー犬訓練データセットによる finetuning の結果

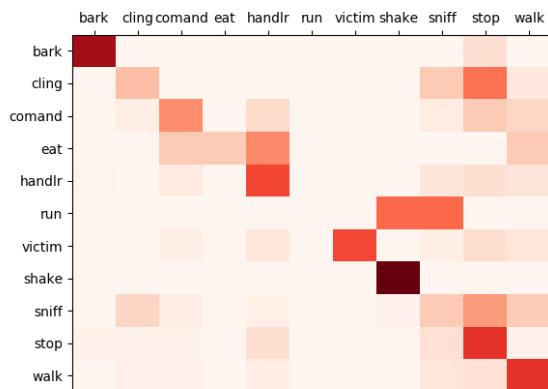


図 5.3: ResNet pretrained model とレスキュー犬訓練データセットによる finetuning の結果

表 5.1: 静止画像を用いた VGG16 の finetuning 結果

クラス	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk	全体
Precision	0.475	0.148	0.0	0.377	0.611	nan	0.37	nan	nan	0.74	0.636	0.565
Recall	0.333	0.108	0.0	0.025	0.059	0.0	0.313	0.0	0.0	0.742	0.72	0.656
Jaccard	0.244	0.066	0.0	0.024	0.057	0.0	0.204	0.0	0.0	0.588	0.51	0.436

5.2 マルチラベル推定

レスキュー犬訓練データセットのマルチラベル推定を行なった。動画で見たデータセットの前半 70% を学習、後半 30% を評価に用いた。モデル毎にそれぞれチューニングを行い、モデル内で最も精度の良いものを示す。各表ではクラス毎の精度と全体を合計しての精度を Precision (適合性) , Recall (再現率) , Jaccard 係数で示している。なお、Jaccard 係数とは

$$tfrac{TP}{TP + FN + FP}$$

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。

5.2.1 静止画像からのマルチクラス推定

静止画像からのマルチクラス推定では、ImageNet で学習した VGG16 の pre-trained model を用いての finetuning を行なった。推定精度を表 5.1 に示す。

eat クラス、run クラスが特に精度が低く、データ数の少なさと関係していると考察できる。データ数が少ないにも関わらず Precision の高い shake クラス、反対にデータ数が十分にも関わらず精度の低い sniff クラスからは、図 4.5 に示したように画像特徴の取りやすさ、取りにくさに依存していることが確認できる。

5.2.2 optical flow 画像からのマルチクラス推定

optical flow 画像からのマルチクラス推定では、静止画像からのマルチラベル推定と同じように ImageNet で学習した VGG16 の pretrained model を用いての

表 5.2: optical flow 画像を用いた VGG16 の finetuning 結果

クラス	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk	全体
Precision	0.265	0.0	nan	0.0	0.938	nan	0.169	nan	nan	0.79	0.604	0.51
Recall	0.232	0.0	0.0	0.0	0.017	0.0	0.018	0.0	0.0	0.695	0.693	0.664
Jaccard	0.141	0.0	0.0	0.0	0.017	0.0	0.017	0.0	0.0	0.586	0.476	0.406

finetuning を行なった。推定精度を表 5.2 に示す。

静止画像と比較して、shake クラス、stop クラスなどの動き特徴の現れやすそうなクラスの精度が高くなることを期待していたが、それらを含め全体的に精度が下がった。画像特徴が失われたため、推定が困難になった様子がうかがえる。

5.2.3 音声からのマルチクラス推定

動画の音声データのみを用いてマルチクラス推定を行なった。ネットワークは [7] の音声分類ネットワークを参考に構成した。メル周波数スペクトラム係数を用いて音声データから特徴を取り出し、その値をネットワークへの入力とした。29.97fps の動画 31 フレーム分の音声を 48000Hz として扱い、(20, 94) 次元の入力を得た。

5.2.3.1 1D Convolutional network

図 ??に示したネットワークの 2D Convolution レイヤを 1D Convolution レイヤに置き換えたネットワークを用いて推定を行なった。推定結果を表 5.3 に示す。

全体を通して、静止画像からのマルチクラス推定よりも精度が上昇した。特に、bark クラス、command クラス、shake クラス、sniff クラスの精度上昇が顕著であり、音特徴がクラス推定に重要であることが示された。

5.2.3.2 2D Convolutional network

音声データから得た (20, 94) 次元の入力にチャネルを追加し、(20, 94, 1) の画像としてネットワークへ入力した。1D Convolutional network と比較し特徴量が増え、精度がわずかに上昇した。推定精度を表 ??に示す。

((まだです))

表 5.3: 音声データを用いた 1d Convolutional network の推定結果

クラス	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk	全体
Precision	0.909	0.13	0.361	0.141	0.245	0.0	0.419	0.708	0.583	0.919	0.759	0.699
Recall	0.717	0.161	0.361	0.026	0.24	0.0	0.442	0.538	0.781	0.798	0.907	0.656
Jaccard	0.669	0.078	0.22	0.023	0.138	0.0	0.274	0.44	0.502	0.745	0.704	0.512

表 5.4: 音声データを用いた 2d Convolutional network の推定結果

クラス	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk	全体
Precision	0.844	0.094	0.357	0.013	0.192	nan	0.407	0.794	0.588	0.917	0.808	0.639
Recall	0.628	0.064	0.285	0.002	0.079	0.0	0.284	0.33	0.83	0.797	0.898	0.721
Jaccard	0.563	0.04	0.188	0.001	0.059	0.0	0.201	0.304	0.524	0.744	0.74	0.512

(??? 全体としての精度は上昇したものの、クラス別に比較した際に 1D Convolutional network で学習したモデルに劣るクラスが半数近くを占める。学習コストなどを含めるなど評価指標を変えた際には 2D Convolutional network が劣る可能性がある。???)

5.2.4 Two-stream network

静止画像と optical flow 画像を学習していない VGG16 ネットワークにそれぞれ入力し、得られた 2 つの出力を結合した結果からマルチクラス推定を行なった。推定結果を表 5.5 に示す。

静止画像単体、optical flow 画像単体からの推定と比較して精度がわずかに上昇している。特に sniff クラスなどが劇的に精度が上昇している。動画から得られた静止画像の画像特徴に加えて、optical flow から得られる動き特徴をそれぞれ用いた学習ができていると考えられる。ただし、shake クラスなど、静止画像と optical flow 画像がお互いに悪影響を与え全く分類できなくなってしまったクラスも存在している。

表 5.5: Two-stream network の推定結果

クラス	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk	全体
Precision	0.522	0.04	0.315	0.0	0.395	nan	0.478	nan	0.472	0.848	0.771	0.571
Recall	0.122	0.033	0.047	0.0	0.204	0.0	0.36	0.0	0.813	0.807	0.833	0.646
Jaccard	0.11	0.018	0.043	0.0	0.155	0.0	0.259	0.0	0.426	0.705	0.668	0.435

表 5.6: 音声と静止画像からのマルチクラス推定結果

クラス	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk	全体
Precision	0.909	0.05	0.312	0.051	0.249	0.042	0.42	0.56	0.592	0.885	0.787	0.661
Recall	0.709	0.077	0.341	0.028	0.177	0.002	0.537	0.589	0.758	0.802	0.855	0.673
Jaccard	0.662	0.031	0.195	0.018	0.115	0.002	0.308	0.402	0.498	0.726	0.694	0.5

5.2.5 Sound based Two-stream network

音声データに静止画像 , optical flow 画像を個別に組み合わせ , マルチクラス推定を行なった . 音声の特徴を取り出すネットワークには 2D Convolutional network を用いた .

5.2.5.1 音声データと静止画像からのマルチクラス推定

音声と静止画像を組み合わせたマルチクラス推定の結果を表 5.6 に示す .

((まだです))

5.2.5.2 音声データと optidcal flow 画像からのマルチクラス推定

音声と optical flow 画像を組み合わせたマルチクラス推定の結果を表 5.7 に示す .

((まだです))

5.2.6 Sound based Three-stream network

表 5.8 に示す .

表 5.7: 音声と optical flow 画像からのマルチクラス推定結果

クラス	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk	全体
Precision	0.887	0.071	0.332	0.052	0.245	0.143	0.329	0.692	0.564	0.881	0.791	0.681
Recall	0.729	0.177	0.441	0.019	0.198	0.01	0.409	0.424	0.782	0.845	0.847	0.641
Jaccard	0.667	0.054	0.234	0.014	0.123	0.01	0.223	0.356	0.487	0.759	0.692	0.493

表 5.8: still optic sound

クラス	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk	全体
Precision	0.888	0.165	0.282	0.188	0.313	0.212	0.527	0.708	0.621	0.891	0.822	0.702
Recall	0.623	0.423	0.355	0.092	0.306	0.029	0.709	0.492	0.783	0.861	0.86	0.663
Jaccard	0.577	0.135	0.186	0.066	0.183	0.026	0.433	0.409	0.53	0.779	0.725	0.518

((まだです))

5.3 考察

全体の結果を表 5.9 に示す .

表 5.9: 各実験比較表

	bark	cling	command	eat	handler	run	victim	shake	sniff	stop	walk	全体
静止画像	0.244	0.066	0.0	0.024	0.057	0.0	0.204	0.0	0.0	0.588	0.51	0.436
optical flow	0.141	0.0	0.0	0.0	0.017	0.0	0.017	0.0	0.0	0.586	0.476	0.406
音声 (Conv1D)	0.669	0.078	0.22	0.023	0.138	0.0	0.274	0.44	0.502	0.745	0.704	0.512
音声 (Conv2D)	0.563	0.04	0.188	0.001	0.059	0.0	0.201	0.304	0.524	0.744	0.74	0.512
静止画像+optical	0.11	0.018	0.043	0.0	0.155	0.0	0.259	0.0	0.426	0.705	0.668	0.435
静止画像+音声	0.662	0.031	0.195	0.018	0.115	0.002	0.308	0.402	0.498	0.726	0.694	0.5
optical flow+音声	0.667	0.054	0.234	0.014	0.123	0.01	0.223	0.356	0.487	0.759	0.692	0.493
静止+optical+音声	0.577	0.135	0.186	0.066	0.183	0.026	0.433	0.409	0.53	0.779	0.725	0.518

第6章

まとめ,今後の課題

Sound based Three-stream 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. 提案手法との比較のために行なった実験は以下である.

- 静止画像からの犬行動マルチラベル推定
- optical flow 画像からの犬行動マルチラベル推定
- 音声からの犬行動マルチラベル推定
- 静止画像と optical flow 画像からの犬行動マルチラベル推定
- 音声と静止画像からの犬行動マルチラベル推定
- 音声と optical flow 画像からの犬行動マルチラベル推定
- 音声と静止画と optical flow 画像からの犬行動マルチラベル推定

結果は提案手法が最も高く, 音声・静止画像・optical flow 画像のそれぞれに必要な情報が含まれてあり, 3つのデータにそれぞれ必要な情報が含まれているとわかった. 精度は 51.2% と数値では決して高いとは言えないが, 約 30fps の動画で 1/5 フレーム毎に行なった推定の結果であるため非実用的とも言えない結果となった.

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. また, 実験に対する疑問として音声フレームの長さはどの程度か適しているのか判断できていない. この変更でどの程度影響があるのかを検証し, 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.

参考文献

- [1] ResqueDog Association Japan. resque-dog. <http://buycott.me/10for1/rescue-dog.html>. Accessed: 2019-01-08.
- [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.of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 630–635, 2015.
- [3] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pp. 568–576, 2014.
- [4] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015.
- [5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2014.
- [6] 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.
- [7] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016.
- [8] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *arXiv preprint arXiv:1804.03641*, 2018.