

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

指導教員 柳井 啓司 教授

平成31年1月30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。

# 目次

<b>第1章 はじめに</b>	<b>1</b>
<b>第2章 関連研究</b>	<b>3</b>
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 who . . . . .	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
<b>第3章 提案手法</b>	<b>9</b>
3.1 動画像平均画像クラス分類 . . . . .	10
3.2 オプティカルフロー画像クラス分類 . . . . .	10
3.3 動画像からのマルチクラス推定 . . . . .	10
3.4 オプティカルフロー画像からのマルチクラス推定 . . . . .	10
3.5 音声からのマルチクラス推定 . . . . .	10
3.6 音声と動画のマルチモーダル情報クラス分類 . . . . .	10
3.6.1 Sound based Two-stream . . . . .	10
3.6.2 Sound based Three-stream . . . . .	10
<b>第4章 データセット</b>	<b>12</b>
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 . . . . .	13
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 . . . . .	14
4.2.1.10 stop . . . . .	14
4.2.1.11 walk-trot . . . . .	15
4.2.2 データ整形 . . . . .	15
4.3 実験 . . . . .	15
<b>第 5 章 実験結果</b>	<b>19</b>
5.1 予備実験 . . . . .	19
5.2 シングルクラス分類 . . . . .	20
5.2.1 動画像平均画像クラス認識 . . . . .	20
5.2.2 オプティカルフロー画像クラス認識 . . . . .	20
5.2.3 音声クラス認識 . . . . .	20
5.2.4 音声と動画のマルチモーダル情報クラス認識 . . . . .	20
5.3 マルチクラス認識 . . . . .	20
5.3.1 動画像平均画像クラス認識 . . . . .	20
5.3.2 オプティカルフロー画像クラス認識 . . . . .	20
5.3.3 音声クラス認識 . . . . .	20
5.3.4 音声と動画のマルチモーダル情報クラス認識 . . . . .	20
<b>第 6 章 まとめ, 今後の課題</b>	<b>21</b>

# 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ざるを得ず、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観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 [?] 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 [?]。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 [?] 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 [?], [?]。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。

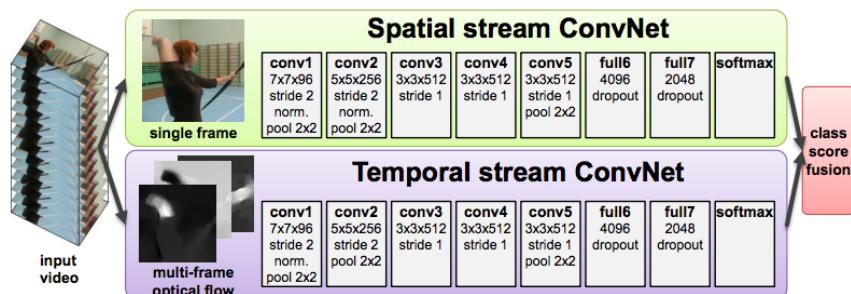


Figure 1: Two-stream architecture for video classification.

図 2.2: Two-stream convolutional networks アーキテクチャ ([?] 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 [?]。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 who

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 [?]。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。

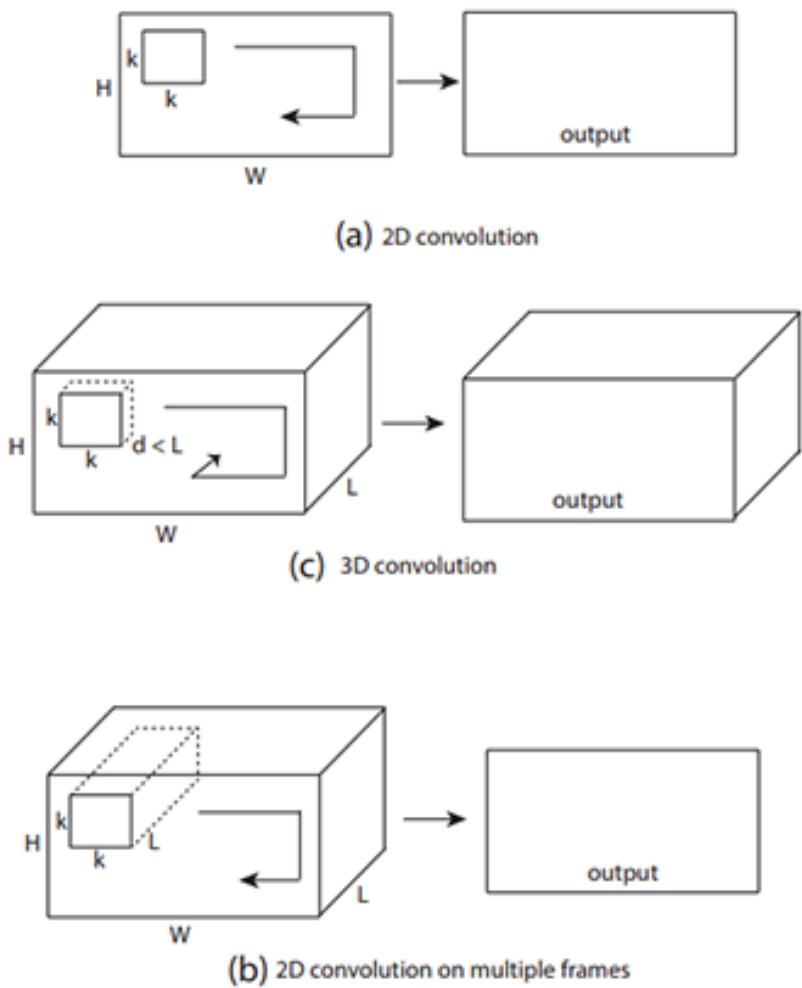


図 2.3: 3D Convolution の詳細 ([?] より引用) . 2D Convolution では縦横方向の畳み込みを行っており , 3D Convolution では加えて時間方向への畳み込みを行っている .

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 [?]。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。

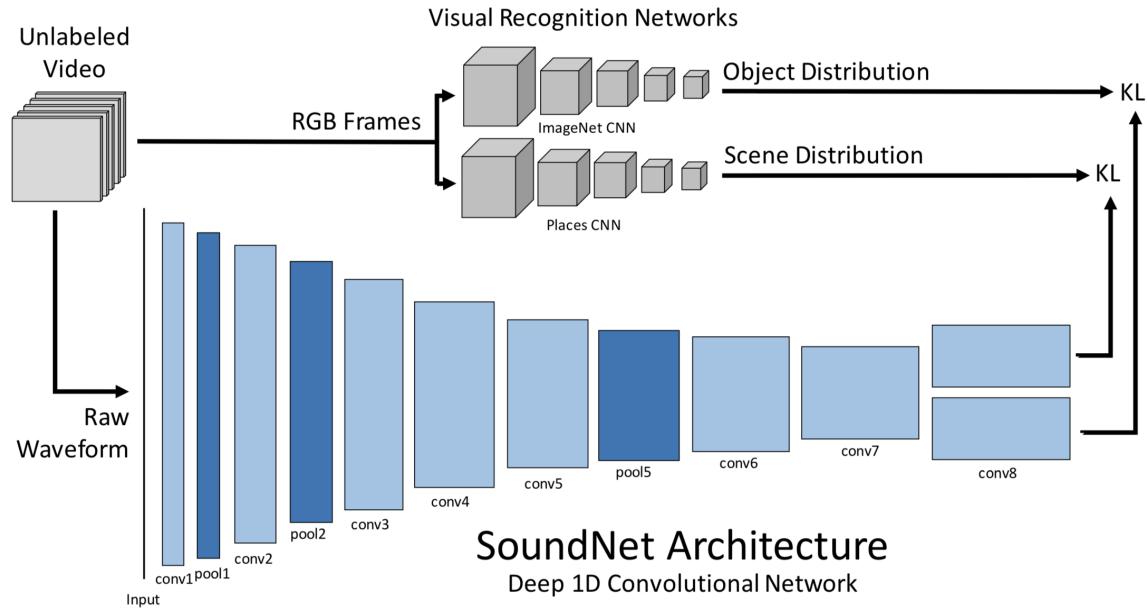


図 2.4: Sound Net のアーキテクチャ ([?] より引用) . 動画から映像と音声を切り分け , 音声に対して 1 次元の畳み込みを行なっている .



図 2.5: Audio-Visual Scene Analysis ([?] 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のフレーム ( $F_t$ ) を取り出し、直後の  $F_{t+1}$  間とのオプティカルフロー画像 ( $O_t$ ) 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 3 つの特徴量を結合し、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 動画像平均画像クラス分類
- 3.2 オプティカルフロー画像クラス分類
- 3.3 動画像からのマルチクラス推定
- 3.4 オプティカルフロー画像からのマルチクラス推定
- 3.5 音声からのマルチクラス推定
- 3.6 音声と動画のマルチモーダル情報クラス分類
  - 3.6.1 Sound based Two-stream



図 3.1: Sound based Two-stream のアーキテクチャ

- 3.6.2 Sound based Three-stream



図 3.2: Sound based Three-stream (提案手法) のアーキテクチャ

## 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

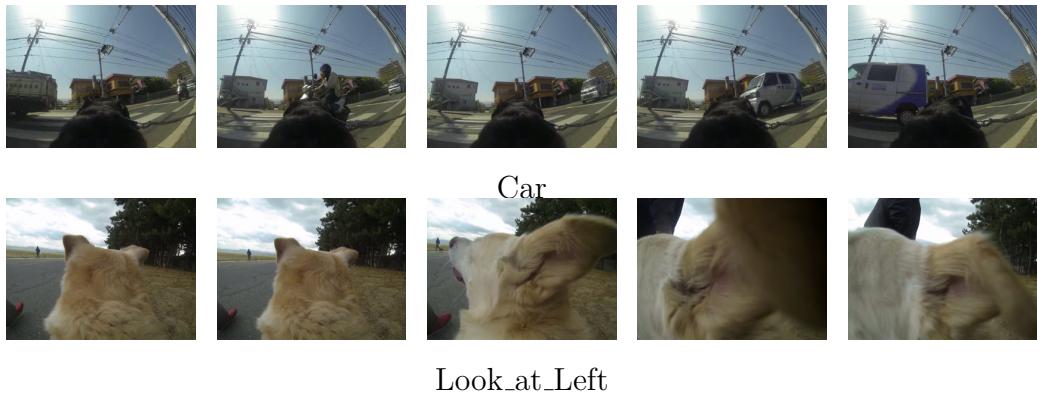


図 4.1: DogCentric Activity Dataset

## 4.2 サイバーレスキュードog 訓練データセット

サイバーレスキュードog訓練データセットは、訓練されているレスキュードogに、専用の計測スーツを着用させ収集したデータセットである。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、dogの一人称視点動画に加えてハンドラー視点動画、レスキュードog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dog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マルチラベルデータセットである。dog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ているが、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。

### 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。

#### 4.2.1.5 look at handler

犬がハンドラーを見ている状態。

#### 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。

#### 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。

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

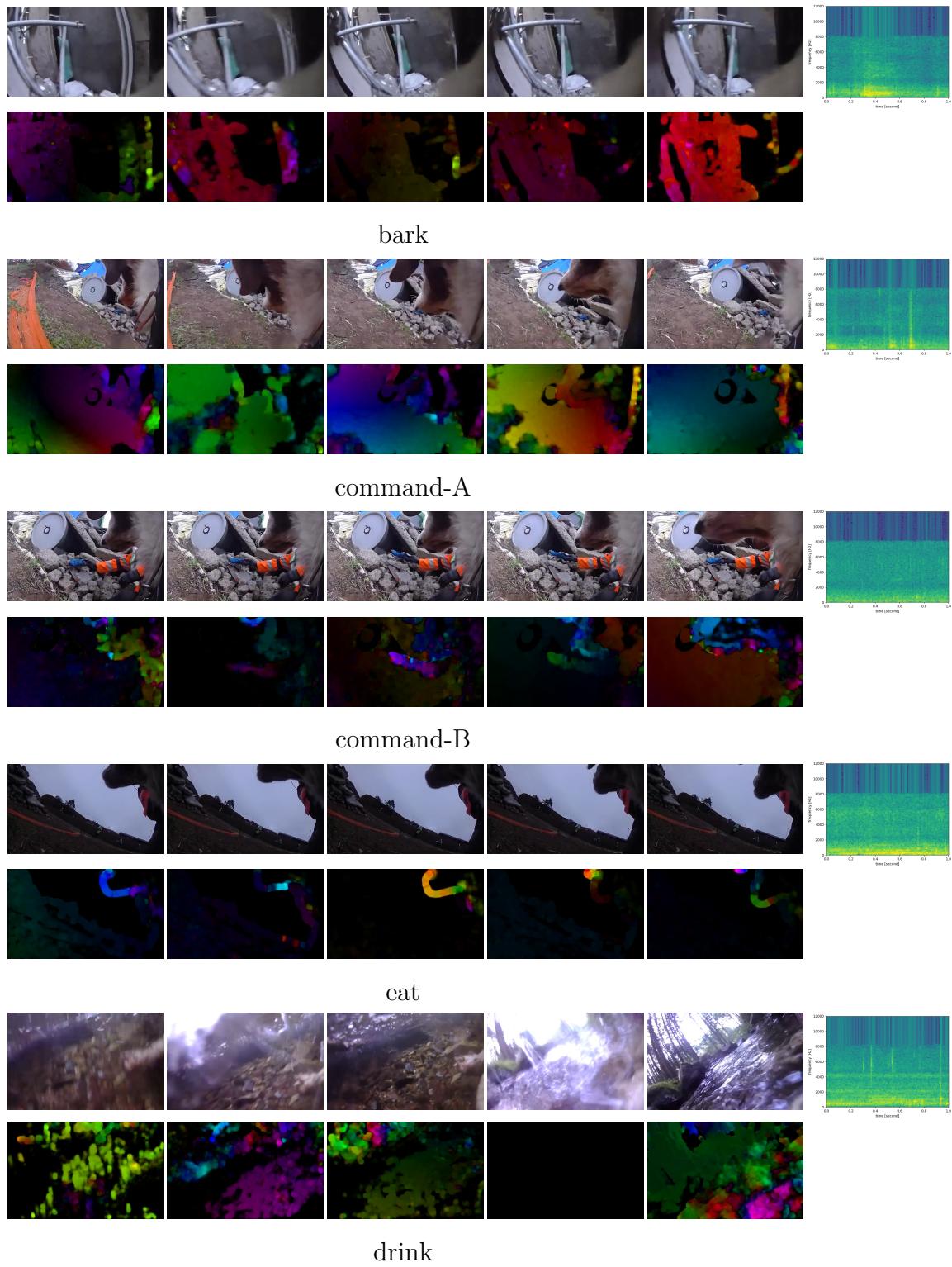
#### 4.2.1.11 walk-trot

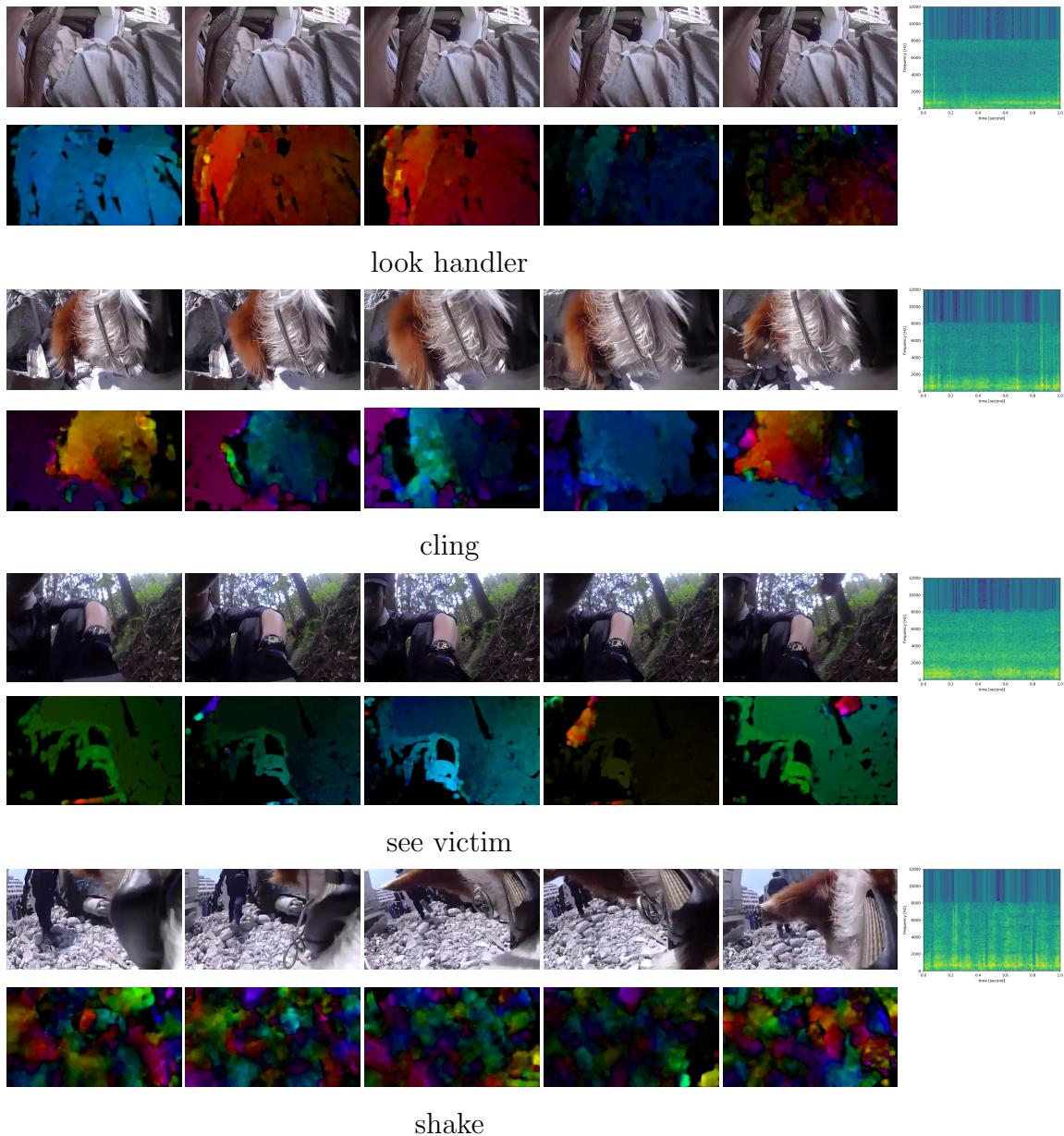
歩いており、run ではない状態。

#### 4.2.2 データ整形

本研究では、フレーム毎にラベルを対応づけた際のラベルのないフレームなどを排除し、過学習を防ぐ目的で 6fps として整形したデータを用いた。そのため、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。画像のみでのクラス的特徴の図示は困難であるため、整形した動画的特徴、音声的特徴とをそれぞれ 図 4.2 に示す。

### 4.3 実験





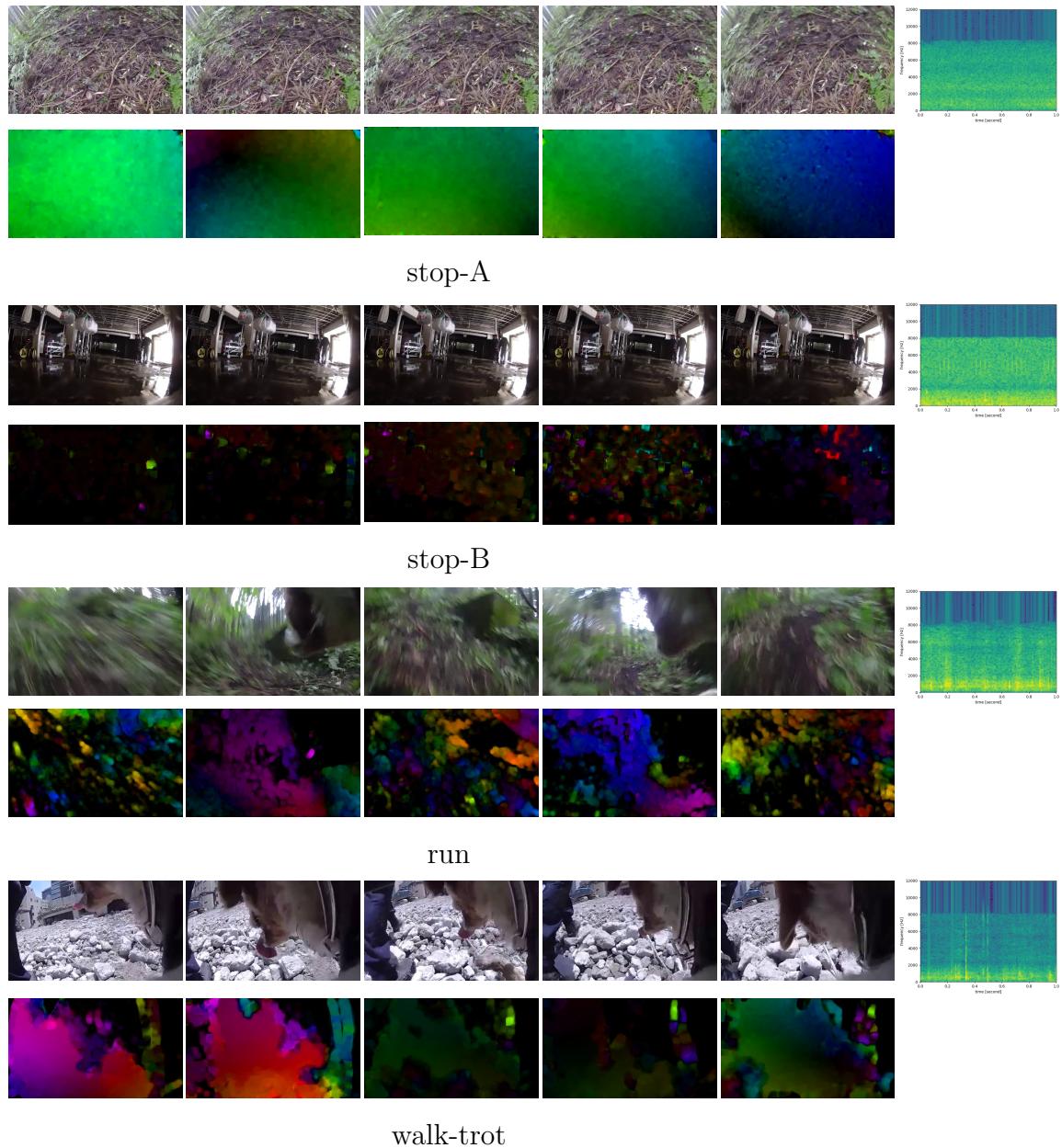


図 4.2: サイバーレスキュー犬訓練データセット

# 第5章

## 実験結果

### 5.1 予備実験

予備実験の結果を(図5.1,5.2)にそれぞれ示す。分類率は、VGG16モデルを利用したものが64.3%,ResNetモデルを利用したものが59.5%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	0
Drink	0	1	0	0	0	0	0	0	2	1
Feed	0	0	1	0	0	0	0	0	0	0
Left	1	0	0	1	0	0	0	0	2	0
Right	0	0	0	0	0	0	0	0	1	0
Pet	0	0	1	0	0	3	1	0	2	0
Ball	0	0	0	0	0	0	5	0	0	0
Shake	0	0	0	0	0	0	1	1	2	0
Sniff	0	0	0	0	0	0	1	0	3	0
Walk	0	0	0	0	0	0	0	0	0	6

図5.1: VGG16 pretrained modelによるfinetuningの結果

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

図 5.2: ResNet pretrained model による finetuning の結果

## 5.2 シングルクラス分類

### 5.2.1 動画像平均画像クラス認識

### 5.2.2 オプティカルフロー画像クラス認識

### 5.2.3 音声クラス認識

### 5.2.4 音声と動画のマルチモーダル情報クラス認識

## 5.3 マルチクラス認識

### 5.3.1 動画像平均画像クラス認識

### 5.3.2 オプティカルフロー画像クラス認識

### 5.3.3 音声クラス認識

### 5.3.4 音声と動画のマルチモーダル情報クラス認識

## 第6章

### まとめ,今後の課題

動画の各フレームの平均を取り, 画像として識別した。データの少ないクラスは精度が低いため, データを補う必要がある。予備実験では簡易的な方法を用いたが, 今後は最新手法による分類を検討している。またレスキュー犬の行動を認識する際には複数クラスの出力にする必要がある。

今後の課題として, 時系列情報を特徴量抽出に使う。また, 音声データから特徴量を抽出し, 動画特徴量と併せたマルチモーダルな特徴量を利用し, レスキュー犬の行動分類を行う。

リアルタイム