H-035

宅内画像を用いた行動推定のための学習モデル構築に関する一検討
A study of constructing a learning model for human action recognition using images inside the house

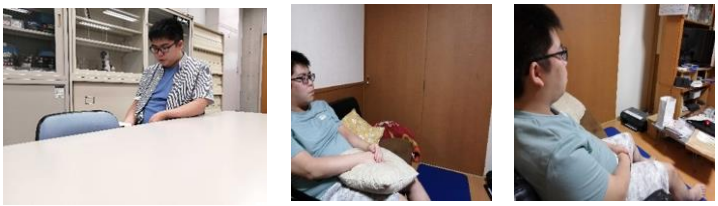
渡邉 奨悟† 橋本 真幸†
Shogo Watanabe Masayuki Hashimoto

1. はじめに

近年，一人暮らしの高齢者の増加が問題になっており，高齢者向けコミュニケーションロボットの開発が行われている．コミュニケーションロボットからの能動的話しかけにより利用者の日常会話量の向上につながることが期待されるが，そのためには利用者に話しかけてよいかどうかの状況の判別が必要である．コミュニケーションロボットのカメラで利用者の行動推定を行うことにより，この状況判別ができると考えられる．

カメラ画像を用いた行動検出や，行動推定技術に関する研究は行われているが，先行研究の多くは屋外にある防犯カメラを用いたものや，部屋の高い位置に置かれた監視カメラを用いたものである[1][2]．コミュニケーションロボットに搭載された位置の低いカメラの画像から行動推定を行う場合の問題点として，撮影範囲が制限され，行動推定に必要なキーアイテム（テレビを見ている場合の「テレビ」など）が映像内に写らないことが考えられる．図 1 に位置の低いカメラで撮影した画像を示す．図 1(a)では実際には本を読んでいるが，画像では机によってキーアイテムである本が隠れてしまっている．また，図 1(b)では実際にはテレビを視聴しているが，人とテレビが同時に写らない．このような画像に対しては行動推定の精度が大幅に低下する可能性がある．

本論文では，撮影位置が低いなどの理由から，撮影範囲が制限されたカメラを用いた行動推定の精度を改善することを目的とし，既存の学習モデルに対する追加学習手法について提案する．



(a)本を読んでいる画像の例 (b)テレビを視聴している画像の例
図 1 キーアイテムが写らない画像の例

2. 提案手法

撮影範囲が制限されたカメラを用いた行動推定においては，映っている人物の一部からでも行動推定が行えることが望ましい．ここでは，キーアイテムを含まない頭部周辺の画像だけから行動を推定することを考える．キーアイテムを含まない画像を教師画像とし，既存の行動推定モデルに追加学習を行う手法を提案する．

初めに，Human action recognition(HAR) dataset とインターネット上から画像を取得し，キーアイテムが写っている画像に対してキーアイテムを含む範囲の矩形座標と対応するクラスを付与するアノテーションを行い教師画像とした[3]．そして，この教師画像を用いて YOLO の学習機能を用いて学習を行い，行動推定を可能にするためのモデルを作成した[4]．（作成したモデルを「従来手法モデル」とする．）

次に，加工等をしてキーアイテムが含まれない画像を用意し，各画像に頭部のみの範囲でアノテーションを行うことでキーアイテム無しのデータセットを作成する．そして，従来手法モデルにこれらキーアイテム無しデータセットを追加学習したモデルを作成する．（「提案モデル」とする．）

本論文で推定を行う行動クラスは，宅内での行動を想定して，「電話をする」，「飲み物を飲む」，「イヤホンをする」，「本を読む」，「食べる」，「寝る」，「パソコンを見る」，「スマホを見る」の 8 クラスとした．

図 2 提案手法の概要図

3. 評価方法

本論文では，前説に述べたように，YOLOv8 の学習モデル作成機能を用いて各モデルを生成し，その精度を検証する．

3.1 従来手法のモデル作成

キーアイテムが含まれるアノテーションを行った教師画像 4263 枚から，従来手法モデルを作成した．表 1 に，使用した教師画像の各クラスの枚数を示す．YOLOv8 では，学習時に train と valid のデータを使用し，test データは学習済モデルの検証に用いる．なお，train データは学習の効率化のためデータ拡張を行っている．

表 1 キーアイテム有り教師画像の枚数

クラス	合計	電話をする	飲む	イヤホンをする	食べる	本を読む	寝る	PCを見る	スマホを見る	Null
train	14985	428	297	414	263	286	332	478	457	533
valid	819	128	105	133	81	89	103	133	139	0
test	436	55	39	60	51	32	65	67	76	0

従来手法モデルの検証の結果得られた行動推定精度を表 2 に示す．表 2 より，全クラス平均の Precision が 0.714，Recall が 0.652，F 値が 0.675 であることが分かる．

† 東洋大学 大学院 理工学研究科 電気電子情報専攻
Electrical, Electronic and Information Engineering Major, Graduate School of Science and Engineering, Toyo University

表 2　従来手法モデルの行動推定精度
(キーアイテム有り画像でテスト)

class	Precision	Recall	F-1 score
電話をする	0.702	0.55	0.617
飲む	0.765	0.667	0.713
イヤホンをする	0.590	0.655	0.621
食べる	0.742	0.46	0.568
本を読む	0.703	0.813	0.754
寝る	0.770	0.723	0.746
PCを見る	0.736	0.791	0.763
スマホを見る	0.700	0.553	0.618
全クラス平均	**0.714**	**0.652**	**0.675**

3.2　提案手法のモデル作成

撮影範囲が制限されたカメラで撮影した画像を想定し，キーアイテムが含まれない頭部周辺の画像を 630 枚用意し，データセットを作成した．作成したデータセットの画像の枚数を表 3 に示す．これは従来手法モデルの作成に使用した表 1 の画像とは別のものである．なお，train データは学習の効率化のためデータ拡張を行っている．

表 3　キーアイテム無し画像の枚数

クラス	合計	電話をする	飲む	イヤホンをする	食べる	本を読む	寝る	PCを見る	スマホを見る	Null
train	1880	66	66	76	41	44	69	77	65	26
valid	80	16	12	13	8	5	10	20	16	0
test	80	12	7	16	5	6	9	21	13	0

4.　結果

3.1 で作成した従来手法モデルの表 3 の test データ(キーアイテム無しの画像)に対する行動推定精度を表 4 に示す．表 2，表 4 を比較すると，従来モデルではキーアイテム無しの画像に対して，Precision が 0.714 から 0.426，Recall が 0.652 から 0.222，F 値が 0.675 から 0.270 と行動推定精度が劣化することが分かる．なお，「PC を見る」クラスについては test データで正解の画像がなかったため Precision，Recall が 0 になっている．これは PC が顔から距離が離れているアイテムのため行動推定が難しいと考える．

表 4　従来手法モデルの行動推定精度
(キーアイテム無し画像でテスト)

class	Precision	Recall	F-1 score
電話をする	0.476	0.224	0.305
飲む	0.712	0.233	0.351
イヤホンをする	0.591	0.661	0.624
食べる	0.534	0.392	0.452
本を読む	0.278	0.0391	0.069
寝る	0.618	0.174	0.272
PCを見る	0	0	0
スマホを見る	0.200	0.0556	0.087
全クラス平均	**0.426**	**0.222**	**0.270**

次に，従来手法モデルに表 3 のキーアイテム無しのデータセットを追加学習させた提案モデルの精度を表 5 に示す．

表 5　提案モデルの行動推定精度
(キーアイテム無し画像でテスト)

class	Precision	Recall	F-1 score
電話をする	0.769	0.909	0.833
飲む	0.563	0.818	0.667
イヤホンをする	0.588	0.667	0.625
食べる	0	0	0
本を読む	0.455	0.833	0.589
寝る	0.833	0.556	0.667
PCを見る	0.333	0.200	0.250
スマホを見る	0.286	0.308	0.297
全クラス平均	**0.478**	**0.536**	**0.491**

表 5 より追加学習を行った提案モデルでは，Precision が 0.478，Recall が 0.536，F 値が 0.491 となり，追加学習前(表 4)と比較して高い精度で行動推定が行えることがわかる．

表 6 に，従来手法モデル及び提案モデルで，それぞれキーアイテム有り，キーアイテム無しの test 画像に対して行動推定した際の精度を示す．表 6 より提案モデルは従来手法モデルのキーアイテム無しの画像に対する推定精度よりも向上していることが分かる．しかし，提案モデルではキーアイテムがある場合の精度が低下している．

表 6　各モデルの行動推定精度

モデル	テストデータ	Precision	recall	F-1 score
従来手法モデル	アイテム有り	0.714	0.652	0.675
	アイテム無し	0.426	0.222	0.270
提案モデル	アイテム有り	0.391	0.180	0.235
	アイテム無し	**0.478**	**0.536**	**0.491**

5.　まとめ

キーアイテムが含まれない画像に対しての行動推定について，追加学習によってより推定精度の高いモデルの作成ができることを確認できた．ただし，クラス別にみると精度が落ちているクラスがある．また，提案モデルではキーアイテム有り画像に対しての精度が低下してしまっている．これらに対する分析と精度改善が今後の課題である．

参考文献

[1] 長山，“ひったくり発生を直前に予測する超次世代型知的防犯カメラシステムに関する研究”，電気学会論文誌 D，103 巻 2 号，2023.
[2] 荻野，田中，“2 台の監視カメラ映像を用いた深層学習による勤務者の行動認識と行動管理システムの構築”，甲南大学紀要．知能情報学編，15 巻 2 号，2023.
[3] Human Action Recognition (HAR) Dataset (オンライン) ,入手先〈https://www.kaggle.com/datasets/meetnagadia/human-action-recognition-har-dataset〉(参照 2024-06-11).
[4] GitHub - ultralystics/ ultralystics(オンライン) ,入手先〈 https://github.com/ultralytics/ultralytics 〉 (参 照 2024-06-11).