H-011

物体のアフォーダンスを用いた一人称映像行動認識
Egocentric Action Recognition with Object Affordances

塩田 宰† 高木 基宏† 熊谷 香織† 近藤 重邦† 瀬下 仁志† 青野 裕司†
Tsukasa Shiota Motohiro Takagi Kaori Kumagai Shigekuni Kondo Hitoshi Seshimo Yushi Aono

1. はじめに

店舗や病院，工場といった様々な現場において，作業者の作業内容を自動で記録・分析する技術が求められている．自動的な記録・分析を実現するためには，作業者の行動を様々なデバイスを用いて認識する必要がある．特に，カメラの映像を用いて行動を認識する手法がコンピュータビジョン・機械学習の分野で盛んに研究されている．近年では，RGB 情報や動作主の姿勢・視線といった身体情報を利用した深層学習による手法が様々提案されている[1][2][3].

しかし，既存手法では映像の見えや身体動作のパターンが類似している行動を誤認識してしまうことがある．例えば，「扉を開く」行動と「扉を閉める」行動は，映像の RGB 情報や身体情報の軌跡が時間軸方向に対して対称となっており誤認識されやすい.

そこで本研究は，動作主が作用する物体の有するアフォーダンス[4]を活用し，行動をより正確に認識する手法を提案する．アフォーダンスとは，周辺環境が動作主に対して与える意味であり，その意味は動作主のとりうる行動に寄与する．例えば，動作主は閉じられた扉に対して「開く」という行動を取ることができる．反対に，その扉はそれ以上物理的に閉じられないため，動作主は「閉じる」という行動を取ることができない．つまり，動作主の行動可能性は，扉の開閉状態といった扉が動作主に与える意味，つまり，扉のアフォーダンスに影響されると捉えることができる．このように，動作主の作用する物体のアフォーダンスは行動を判断するための重要な手掛かりとなる．従って，動作主が行動時に作用する物体のアフォーダンスを考慮して行動を認識するモデルを構築することにより，動作主の行動をより精緻に認識できるという仮説を立てた.

動作主が物体に作用する行動を収録した一人称映像データセットを用いた実験を実施し，物体のアフォーダンスを考慮する手法の性能が既存手法より高くなることを示す.

2. 提案手法

物体のアフォーダンスを行動認識モデルに理解させるために，動作主が作用する物体の情報を行動と同時に学習する手法を提案する．行動認識は認識対象の映像が事前に定義された行動ラベルのいずれに該当するか分類する問題として定式化される．ここで，本研究における行動ラベルとは，動作主が物体に対して実施する行動を動詞と名詞の組み合わせによって表現したものである．詳細は 3.1 節に示す．提案手法では，行動ラベルを分類するタスクに物体のアフォーダンスに関する 2 種類のタスクを追加する．1 つ目は，動作主が作用する物体を認識するタスク（物体分類タスク）である．2 つ目は，動作主が作用する物体の状態を推定するタスク（物体状態推定タスク）である．図 1 に提案手法の概要図を示す.

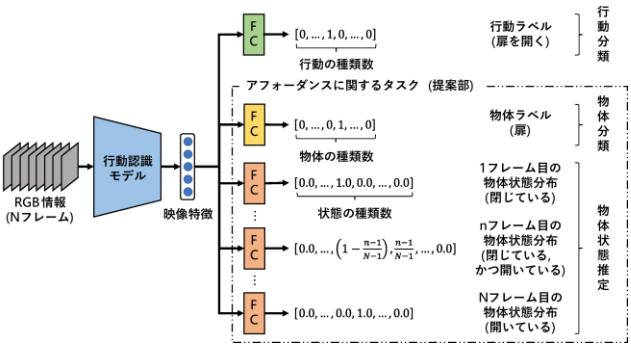

図 1　提案手法概要図

物体分類タスクは，RGB 情報を基に行動認識モデルが出力する映像特徴を用いて，動作主が行動時に作用する物体を認識するタスクである（図 1 中段）．本タスクにおいて正解データとなる物体ラベルは，行動ラベルで名詞として記述されている物体を抽出することで作成でき，人手による追加アノテーションは不要である.

物体状態推定タスクは，映像特徴から各フレームにおける物体の状態を推定するタスクである（図 1 下段）．状態とは，扉であれば「開いている」「閉まっている」といった形容詞等で表現されるものである．つまり，物体のアフォーダンスの理解に必要な物体の状態を RGB 情報から推定するタスクである．本研究では，各フレームにおける物体の状態を物体状態分布として表現し，提案手法ではそれらを推定する．正解データとなる物体状態分布は，動作主が作用する物体の初期状態と最終状態を行動ラベルに対してアノテーションし，それらの状態と映像のフレーム長を利用することで機械的に生成できる．具体的に，フレーム長がNである映像の$n\ (1 \le n \le N)$フレーム目の正解物体状態分布は，初期状態に対応する要素が$\left(1-(n-1/N-1)\right)$，最終状態に対応する要素が$n-1/N-1$，それ以外の要素が0となる分布である．本タスクのために必要な人手作業は，行動前後における物体の状態を行動ラベルに対してアノテーションするのみであり，低コストである.

行動分類タスクと物体分類タスク，および物体状態推定タスクの同時学習を実現するために，損失関数は以下のように定義する.

$$Loss = loss_{act} + \gamma\left(loss_{obj} + \frac{1}{N}\sum\nolimits_{N} loss_{state}\right)$$

ここで，Nは入力される映像のフレーム長，$loss_{act}$は正解行動ラベル分布と予測行動ラベル分布の交差エントロピー損失，$loss_{obj}$は正解物体ラベル分布と予測物体ラベル分布の交差エントロピー損失，$loss_{state}$は正解物体状態分布と予測物体状態分布の KL ダイバージェンス損失，γは物体分類タスクおよび物体状態推定タスクの影響度を調節するハイパーパラメータである.

†NTT 人間情報研究所　NTT Human Informatics Laboratories

表 1　実験結果

	Overall acc.	Mean class acc.	Verb err.	Noun err.	Verb + Noun err.
ベースライン	64.02%	55.81%	8.69%	11.43%	15.86%
提案手法（状）	**65.80%**	**56.98%**	**8.24%**	12.09%	**13.87%**
提案手法（状+物）	**64.96%**	**56.59%**	**8.52%**	12.83%	**13.68%**

3. 実験

既存手法と提案手法を比較し，物体のアフォーダンスに関するタスクの行動認識性能に対する有効性を検証する．

3.1 データセット

実験には，動詞（e.g. open, cut）と名詞（e.g. fridge, tomato）の組み合わせで表現される 106 種類の行動（e.g. Open fridge, Cut tomato）を撮影した映像が計 10,321 件収録されている EGTEA Gaze+ [5] を利用した．訓練とテストの比率を約 8:2 に分割した 3 つの split が定義されている．各手法の性能の評価は，overall acc.と mean class acc.を 3 つの split について算出することで実施した．正解物体ラベルは行動ラベルの名詞とした．正解物体状態分布の作成に向けて，106 種類の行動ラベルの名詞に対して初期状態と最終状態を表現する形容詞を著者らが人手でアノテーションした．定義された形容詞の種類数は計 20 種類となった．

3.2 実装詳細

行動認識モデルには Kinetics [6]で事前学習された SlowFast (8x8 R50, α=4, β=1/8) [1]を利用し，訓練/推論時のフレームの空間方向のサイズはそれぞれ224^2，256^2とした．提案手法の映像特徴のサイズは 1024 次元とした．物体状態推定タスクは slow pathway に入力される 8 フレームに対して実施した．学習の最適化には，学習率を 5e-3，慣性項を 0.9，重み減衰を 1e-4 とした SGD Momentum を利用し，バッチサイズは 16，エポック数は 100 に設定した．損失関数のハイパーパラメータγは 0.5 とした．学習のための検証データは split の訓練データから 1,000 本を無作為抽出した．

3.3 実験結果

表 1 に 3 つの split における実験の各種指標の平均値を示す．ベースラインは SlowFast を用いて行動分類のみを学習した結果，提案手法（状）は行動分類と物体状態推定を同時学習した結果，提案手法（状+物）は行動分類と物体状態推定と物体分類を同時学習した結果である．Verb err.，Noun err.，Verb + Noun err.はそれぞれ推論時に行動ラベルの動詞部分のみを誤認識した割合，名詞部分のみを誤認識した割合，動詞・名詞を共に誤認識した割合を示す．ベースラインと比較して提案手法の各種正解率が向上していることから，学習時に物体のアフォーダンスに関するタスクを同時学習することで，より精緻な行動認識が実現できることを確認した．

3 パターンの誤認識率についてベースラインと提案手法を比較すると，提案手法では動詞部分のみを誤認識する割合と動詞・名詞を共に誤認識する割合が低下していた．具体的な事例を確認すると「Take eating_utensil (put → grasped)」と「Put eating_utensil (grasped → put)」（図 2 上段）のような時間軸方向に見えが対称となる行動，「Wash hand (dirty → clean)」と「Turn on faucet (closed → open)」（図 2 中段）のような時空間方向に見えが類似している行動の誤認識が改善されていた．一方で，提案手法では名詞部分のみを誤認識する割合が増加していた．具体的な事例を確認すると「Cut cucumber (whole → cut)」と「Cut lettuce (whole → cut)」（図 2 下段）や「Open fridge (closed → open)」と「Open fridge_drawer (closed → open)」のような時空間方向に見えが類似しており，かつ物体の状態の変化も同じである行動を誤る傾向にあった．

図 2　推論結果の比較

4. おわりに

本研究では，物体のアフォーダンスに関するタスクを同時学習する行動認識手法を提案した．同時学習によって，入力情報が類似する行動の誤認識が改善され，行動の認識性能を向上することができた．一方で，アフォーダンスに関する情報に差がない行動については誤認識が増加する結果となった．

参考文献

[1] C. Feichtenhofer et. al., “SlowFast Networks for Video Recognition”, in *ICCV*, pp. 6202-6211, 2019.
[2] L. Shi et. al., “Two-Stream Adaptive Graph Convolutional Networks for Skelton-Based Action Recognition”, in *CVPR*, pp. 12026-12035, 2019.
[3] Y. Huang et. al., “Mutual Context Network for Jointly Estimating Egocentric Gaze and Action”, IEEE TIP, vol. 29, pp. 7795-7806, 2020.
[4] J. J. Gibson. “The ecological approach to visual perception”, Houghton Mifflin, 1979.
[5] Y. Li et. al., “In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video”, in *ECCV*, pp. 619-635, 2018.
[6] J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”, in *CVPR*, pp. 6299-6308, 2017.