2R-06

CNN 特徴を用いた作業映像の時間的アライメント

佐藤 寛 †　　谷口 行信 †
† 東京理科大学

1　はじめに

近年，急速なグローバル化に伴い，製造業や飲食店などのサービス業を中心に異なる言語を用いる人々の間で技術・技能を伝承する必要性が高まっている．効率的な技術・技能伝承を支援する一つの手段として，熟練者と非熟練者の作業映像を見比べる方法がある．しかし，作業者ごとに作業速度が異なるため，二つの動画を単純に先頭から再生すると作業シーンが徐々にずれてしまい，同時に同じ作業を比較することが難しくなる．

本研究では，作業映像を効率的に比較するために，複数の作業映像の中から同一の作業シーンを見つけ，自動的に対応付ける映像アライメント手法を提案する（図1）．具体的には時間的アライメントの精度向上のため，畳み込みニューラルネットワークを導入した結果について報告する．また，本研究では，アライメントを行う映像の作業順序は同一であると仮定している．

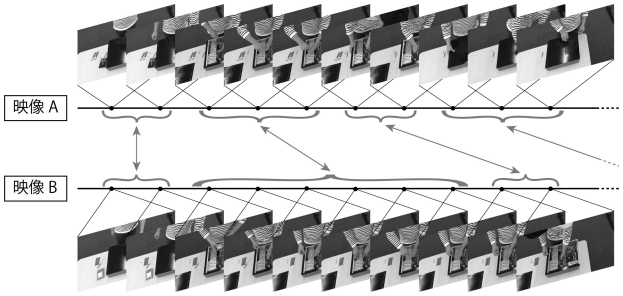

図 1: 映像アライメントの概念図

2　関連研究

2.1　映像アライメント

佐藤ら [1] は，SIFT 特徴を用いた Bag-of-Featurs（以下，BoF），improved Dense Trajectories（以下，iDT），色ヒストグラム（以下，CH）の 3 つの特徴量を手元領域に限定して取得したものを統合し DP マッチングを行うことでアライメントを行っている．特徴量の統合には彼らが適応的重み付けと呼ぶ手法が用いられている．佐藤らの手法では，アライメントを行う映像内の道具の配置が異なる場合，同一作業同士であっても，手の動きの方向が変化し，うまくアライメントが行えないという問題点がある．

2.2　深層学習による動作認識

人物の動作認識において畳み込みニューラルネットワーク（以下，CNN）の有効性が示されている．特に動作認識においては，3 次元 CNN が高い精度を達成している [2].

2 次元 CNN が見えの情報のみを取得しているのに対して，3 次元 CNN は見え情報に加えて，動きの情報も取得していると考えられている．CNN は学習過程で学習タスクに適応した特徴表現を獲得していると考えられているので，動作認識タスクを学習させた CNN は，動作認識に適応した特徴表現を獲得していると考えられる．本研究では，CNN の中間層の出力を特徴量と捉え（以下，CNN 特徴），映像から特徴抽出を行う．

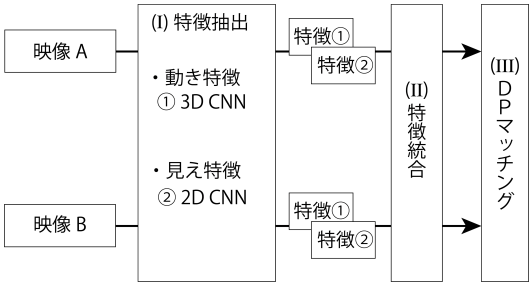

図 2: 提案手法の処理の流れ

3　提案手法

本研究では，従来手法の問題点を解決し，特に手を伸ばす方向が定まっていない作業に対応できるよう改良した作業映像の時間的アライメント手法を提案する．提案手法の処理の流れを図 2 に示す．(1) CNN 特徴を導入し，(2) 手の動きを 10 クラスに分類するタスクをファインチューニングにより学習させた 3 次元 CNN を用いる．具体的な手順として，(I) 特徴抽出ステップでは，比較映像と参照映像から 2 種類の CNN 特徴を抽出する．(II) 特徴統合ステップでは，複数の特徴を統合した特徴ベクトルを生成する．(III) DP マッチングステップでは，フレーム間距離の総和が最小となるように DP マッチングを用いてフレーム間の対応付けを求める．

3.1　特徴量抽出

映像の各フレームから特徴量を抽出する．本研究では，特徴抽出器として，2 種類の CNN を用いる．1 つは Resnet-152（以下，2D CNN），もう 1 つは 3D ResnXt-101[3]（以下，3D CNN）である．2D CNN は ImageNet 事前学習済みのモデルを使用する．3D CNN は Kinetics 事前学習済みモデルと，それをファインチュー

表 1: ファインチューニングに用いたラベル

ラベル	動作
1	両腕を左右に振る
2	片腕を左右に振る
3	片腕を上下に振る
4	片手の把持運動
5	両手の把持運動
6	片手の平を返す
7	両手の平を返す
8	円を描く
9	ねじを回す
10	タイピング

Temporal Video Alignment with CNN Features
†Hiroshi Sato　†Yukinobu Taniguchi
†Tokyo University of Science

Copyright ©2019 Information Processing Society of Japan.
All Rights Reserved.

ニングしたもの（以下，Fine-Tuned）を使用する．3D CNN は 16 フレームごとに 1 つの特徴ベクトルを出力するため，入力した 16 フレームごとに共通の特徴量が与えられる．それぞれ最終出力層の 1 つ前の層の出力を特徴量として抽出した．

3D CNN のファインチューニングには，機器の組立作業のような，主に，手先を使った作業のアライメントを想定して，表 1 のラベルを用いた．各ラベルには 10 本程度の動画が含まれており，各ラベルに含まれる動画の再生時間の合計は 4~5 分である．

4 実験・考察

CNN 特徴を導入することで精度が向上すること，特に手を伸ばす方向が決まっていない作業でも時間的アライメントが可能であることを確認するために，2 つの実験を行った．

4.1 実験条件

本実験ではパソコンパーツ増設作業動画（以下，PC 動画），サラダ調理動画（以下，調理動画）の 2 つのデータセットを用いた．PC 動画では道具の配置が共通であるが，調理動画では野菜やボウル，包丁の配置が動画によって異なっている．動画の各フレームには工程ごとにラベルを人手で付与した．PC 動画の作業工程ラベルを表 2 に示す．

表 2: PC 動画の工程ラベル

ラベル	工程
1	PC のふたのネジをはずす
2	PC のふたを開ける
3	HDD の装着
4	HDD のネジをはめる
5	ケーブルの装着
6	メモリの装着
7	PC のふたを閉める
8	PC のふたのネジをはめる
0	その他の動作

調理映像には作業順序が異なる映像が存在するが，提案手法では，作業順序は同一であることを仮定しているため，作業順序が同一であるペアにのみアライメントを行う．精度比較のため，BoF，CH，iDT によるアライメントを行うが，各パラメータは佐藤らの実験と同じ値を用いた．

4.2 評価指標

映像 A の t 番目のフレームに割り当てられたラベルを l_t^A，映像 A の総フレーム数を T^A とする．映像 A の t 番目のフレームが映像 B の t' 番目のフレームに対応付けられるとき，アライメント性能を評価する指標として正答率を

$$
\begin{aligned}
\text{正答率} &= \frac{F_{correct}}{T^A}, \\
F_{correct} &= \sum_{t=1}^{T^A} \delta(l_t^A - l_{t'}^B), \quad \delta(x) = \begin{cases} 1 & (x = 0), \\ 0 & (x \neq 0). \end{cases}
\end{aligned}
$$

と定義する．

本実験では PC 動画 12 本から 6 本を，調理動画 12 本から 5 本を無作為に選ぶ操作を繰り返し行い，選択した映像全ての組合せに対して正答率を求め，PC 映像に対して 30 パターン，調理映像に対して 14 パターン分の平均正答率を求めた．

4.3 実験結果・考察

各特徴量単体，及び複数特徴量を統合した場合の平均正答率を表 3 に示す．

表 3: 平均正答率

特徴量	PC 映像	調理映像
BoF	69.69%	50.27%
iDT	70.57%	55.35%
CH	64.07%	53.64%
2D CNN	73.03%	52.84%
3D CNN	70.00%	52.43%
3D CNN (Fine-Tuned)	72.12%	53.51%
BoF + iDT + CH[1]	74.17%	57.09%
2D CNN + 3D CNN（Fine-tuned）	**78.58%**	**66.63%**

単一特徴量での結果を見ると，PC 映像では 2D CNN が最も精度が高い．これは，作業工程の中で見えの情報の変化が大きく，2D CNN が見えの特徴を効率良く抽出していることによると考えられる．

それに対し，調理映像では iDT が最も精度が高かった．これは，食材の形状が変化するため見え情報がアライメントの手がかりにならなかったものと考えられる．iDT が同じく動き情報を取得している 3D CNN より精度が高い理由としては，(1) 3D CNN が見え特徴も同時に抽出しており，見えの特徴が悪い影響を与えていること，(2) 16 フレームごとに共通な特徴量になっているので，フレーム単位の細かい変化に対応できないことが考えられる．

また，3D CNN のファインチューニングを行うと精度の向上がみられることから，本研究でのファインチューニングが，作業映像アライメントの精度向上に有効であることが示された．

複数特徴量を統合した結果については，3D CNN 特徴と 2D CNN 特徴を統合した場合に，どちらのデータセットについても佐藤ら [1] の手法の精度を超える結果となった．これは，2D CNN，3D CNN それぞれの特徴が適応的重み付けにより，DP マッチングに適した特徴に統合されているからであると考えられる．

5 おわりに

本研究では，作業映像の時間的アライメントにおいて CNN 特徴を導入した場合の結果について報告した．2D CNN 特徴と，3D CNN 特徴を，統合することで，従来手法を上回る精度が得られることを示した．今後は，作業順序が入れ替わる場合にも対応可能な時間的アライメント手法の検討が課題である．

本研究は JSPS 科研費 JP17K06608 の助成を受けたものである．

参考文献

[1] T. Sato, Y. Shimada and Y. Taniguchi, “Temporal Video Alignment Based on Integrating Multiple Features by Adaptive Weighting,” *In IWAIT2018,* pp.1–5, 2018.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks,” *In ICCV2015*, pp. 4489–4497, 2015.

[3] K. Hara, H. Kataoka and Y. Satoh, “Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?,” *In CVPR 2018*, pp. 6546–6555, 2018.

Copyright ©2019 Information Processing Society of Japan.
All Rights Reserved.