Vol.2016-CVIM-200 No.2
2016/1/21

手の動作に基づく複数一人称視点作業映像のアライメント

樋口 啓太[1,a)]　米谷 竜[1,b)]　佐藤 洋一[1,c)]

概要：本論文ではウェアラブルカメラにより撮影した一人称視点作業映像おいて，作業を構成する動作単位での解析を実現するために，複数の映像をアライメントする手法を提案する．ウェアラブルカメラは装着者の体験を主観的な視点から記録することができるため，実世界での作業映像のアーカイブとその利活用が期待されている．しかし，同一の作業を映している場合でも，作業順序や作業時間が異なるといったことが起こりえる．本研究では手の動作を基本単位とした，一人称視点作業映像間のアライメント手法を提案する．本アライメント手法では，見本となる作業映像に手動で基本動作のラベル付けをし，見本映像と他の同一の作業を映した伝搬対象の映像をフレーム毎に対応づけることにより，基本動作のラベルを割り当てる．本論文では，提案手法の初期評価を実施し，今後の課題を議論した．

キーワード：作業映像アライメント，一人称視点映像

First Person Video Alignment Based on Hand Activity

Keita Higuchi[1,a)]　Ryo Yonetani[1,b)]　Yoichi Sato[1,c)]

1. はじめに

一人称視点映像は人間の頭部や視線位置に装着したウェアラブルカメラにより撮影した映像である．装着した人物の主観的な体験を記録することができため，装着者の視点を中心として，装着者の身体や周囲の人物，物体がどこに配置しているのかを撮影することができる．作業中の人物がウェアラブルカメラを装着した場合，作業者の視点から見てどの物体を，どのように操作しているのかを鮮明に映し出すことができる．そのため，実環境中での作業のアーカイブや，コンピュータによる作業支援のために使われることが期待されている．

一人称視点映像を基に装着者の行動を解析するための研究が盛んに行われている [1], [2], [3], [4]．映像解析の結果から，装着者が誰と会話をしているか，装着者がどのような作業をしているのかといった情報を得ることができる．また，それらの情報を利用することによって，人物間のコミュニケーション解析や個人の技能解析などへ応用することが考えられる．

本研究では，一人称視点作業映像をより詳細に解析可能にし，その利活用を促すために，複数作業映像間のアライメント実現を目指す．作業映像間のアライメントとは，特定の動作 (物を取る，移動させるなど) が作業映像中において，いつ映っているのか対応付けすることである．そのため，アライメントを取ることにより，作業中における動作単位での映像の解析をすることができる．さらに，アライメントされた情報を基にした映像の検索や，複数映像の同時閲覧支援，さらに作業者への実時間支援などへ活用することができる．しかしながら，一人称視点作業映像では，同一の作業を映している場合でも異なる環境で作業をしている場合や，作業時間や作業順序が異なるといった場合が起こりえるため，アライメントを実現するためにはこれらの課題を解決しなくてはならない．

本研究では手による動作を基本単位とした，一人称視点作業映像間のアライメント手法を提案する．本手法では，見本となる作業映像（リファレンス）に基本動作のラベル

[1] 東京大学
The University of Tokyo, Bunkyo, Tokyo, Japan
[a)] khiguchi@iis.u-tokyo.ac.jp
[b)] yonetani@iis.u-tokyo.ac.jp
[c)] ysato@iis.u-tokyo.ac.jp

を手動で付け，そのリファレンスを基に他の動画にもラベルを伝播させる．それにより，同一の作業を映している伝搬対象となる映像（クエリ）に，リファレンスと同様の基本動作ラベルをフレーム単位に割り当てる．作業映像間の比較に伸縮マッチングを取り入れることにより，映像毎の作業時間の変化に対応できる映像比較を可能とする．また，それぞれの映像をフレーム毎に比較するための，手の動きを記述する特徴抽出方法を提案する．

本論文では，最初に一人称視点作業映像のアライメントの目的と基本方針について説明し，その応用シナリオを提示する．そして，複数映像間アライメントを実現するための提案手法を説明する．その提案手法を用いた評価実験を実施し，今後の課題を議論する．

2. 一人称視点作業映像のアライメント

一人称視点作業映像のアライメントとは，複数の映像を，作業を構成する基本動作基づき対応付けをすることである．基本動作とは，物を取る，移動させるといった作業の基本単位となるような短時間中に行われる動作である．一連の作業は基本動作の繰り返しとなるため，同一の作業を撮影している場合には，基本動作の順序は映像間で近似している．そのため，時系列情報を考慮した基本動作の対応付けにより，映像間のアライメントを実現できるのではないかと考える．

本研究では手による基本動作に着目し，ウェアラブルカメラで撮影された手を使った作業映像のアライメント実現を目指す．手を使った作業には製造や工作，料理といった多くの種類がある．アライメントにより基本動作単位での作業映像の解析ができるようになるため，個人技能の理解と撮影された映像の利活用を促進できる可能性がある．これまでの一人称視点映像を用いた作業認識に関する研究により，どのような作業をしているのかを判別することが可能となっている．本研究ではこれらの作業認識の結果から得られた同じ作業を撮影した映像に対して，作業がどのような基本動作から構成されているのかや，それぞれの基本動作にどのくらいの時間がかかっているかといった情報を付与することを目的としている．

2.1 リファレンス映像を用いた複数映像のアライメント

本研究ではアライメントの実現方法として，ユーザ自身の手によって基本動作のラベル付けをされたリファレンス映像を作成し，他の映像にも基本動作ラベルを伝播させる手法を提案する．そのため，一つの映像にラベル付けをするだけで，同一カテゴリーの映像に自動で同様のラベルを割り当てることができる．図 1 に本アライメント手法のコンセプトを示す．リファレンス映像とクエリ映像では，同一の作業を映しているため，手の動作に基づき映像をフレームレベルで対応付けすることでラベルの伝播をしている．しかし，それぞれの動作は映像により長さが異なる．そのため，時間の伸長を許容するフレーム間マッチングの手法を採用する．

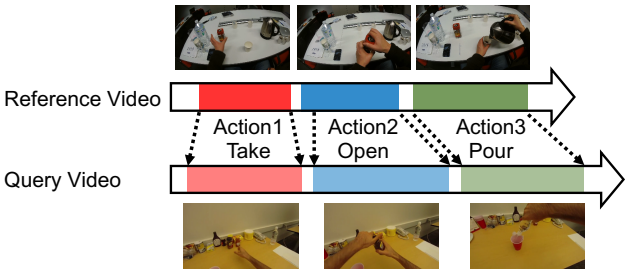

図 1 本アライメント手法のコンセプト：基本動作のラベルが付けられたリファレンス映像を作成し，クエリとなる同様の作業を映した映像と対応づけることによりラベルを伝播させる．

本論文では作業映像アライメントの初期検討として，手の動作にのみ着目して基本動作を定義する．そのため，異なるオブジェクトへの操作であったとしても，同様の動作をしていれば同一のラベルを定義する．

2.2 応用シナリオ

手の基本動作に基づく作業映像アライメントが実現することにより，以下のような応用シナリオが考えられる．

2.2.1 作業工程の解析

複数映像間での基本動作のアライメントをすることで，個人技能や作業工程の詳細な解析ができる可能性がある．アライメントの結果から，幾つかの基本動作からなる作業工程にかかった時間を測定することができる．そのため，それぞれの工程に費やした時間の自動計測ができる．熟練者の作業映像をリファレンスとし，非熟練者の作業映像とアライメントすることで，熟練者と比較した際の作業速度の差を明らかにし，教育に応用することが期待できる．また，高度な技術を持つ専門家の作業を撮り溜め解析することで，個人特有の技能の特徴を抽出できるかもしれない．

2.2.2 動作に基づく作業映像の検索

本アライメント手法を用いることにより，映像間の対応付けが成功するかを判別することができるため，一人称視点作業映像の検索に応用できる．ウェアラブルカメラは作業の様子を容易に撮影することができるため，今後様々な作業映像がインターネットを通してアクセスできるようになると予想できる．そのとき，検索したい作業映像中の動作と，同様の動作をすることによって検索クエリを作成するという直接的な検索方法が実現できる．

2.2.3 映像の同期再生による閲覧支援

本アライメント手法により，映像を動作ごとに同期させて再生をするような閲覧支援が可能となる．ウェアラブルカメラにより撮影した作業映像は，同じ作業を映している場合でも，それぞれの動作が異なる速度で行われることがあるため，2 つ以上の映像を同時に閲覧し比較することは

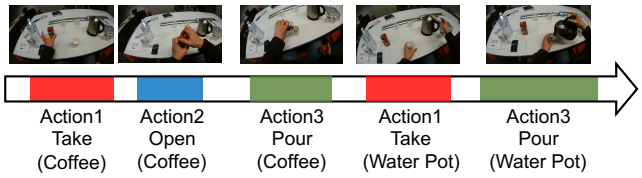

図 2 リファレンス映像へのラベル付け：最初に作業映像中に含まれる基本動作に基づきリファレンス映像へラベル付けをする．本手法では，手の動作にのみ着目をして基本動作を定義する．そのため，異なる操作対象物であっても，同様の操作であれば同一のラベルを付ける．

容易ではない．本手法により，Wang らの Videosnapping [5] のような映像間の時間対応付けを，基本動作単位できる．そのため，複数映像の同期再生によって，それぞれを動作単位で比較することができるため，技能の習熟度解析を支援できる．

2.2.4 作業の実時間支援

本手法を実時間処理として実現することができれば，作業中に教師映像をリアルタイムフィードバックすることによる作業支援が可能となる．ヘッドマウントディスプレイ (HMD) を使った作業支援に関する研究が盛んに行われている [6]．それらの研究では，作業者に HMD を通して作業手順を指示することにより，作業支援を実現している．本アライメント手法を用いることにより，教師映像を基に現在の作業工程を推定することができるため，作業状況に応じた映像を提示するような作業支援が可能となる．

3. アライメント手法

本アライメント手法は，1) リファレンス映像へのユーザによる基本動作のラベル付け，2) 作業映像から基本動作の特徴を記述，3) 伸縮マッチングによる映像間の対応付けによるラベルの割り当てから構成される．

3.1 リファレンス映像へのラベル付け

最初に，ユーザ自身の手にってリファレンス映像への基本動作ラベルの割り当てをする．リファレンスとなる映像を選択し，映像を閲覧しながら各フレーム中にどのような動作が行われているかを割り当てる．図 2 に，リファレンス映像とラベルの例を示す．本手法では操作対象となる物体の区別をせず，手がどのような基本動作をしているかのみに着目をし，映像中において同様の動作には同一のラベルを付ける．

3.2 作業映像からの特徴記述

作業映像間の対応付けをするために，作業映像の各フレームへの特徴記述をする．本研究では，手が基本動作をしている間，どのような動きをしているのかを抽出するために，一定フレーム区間における特徴点の移動方向ヒストグラムを各フレームにおける特徴量とする．図 3(A-B) に

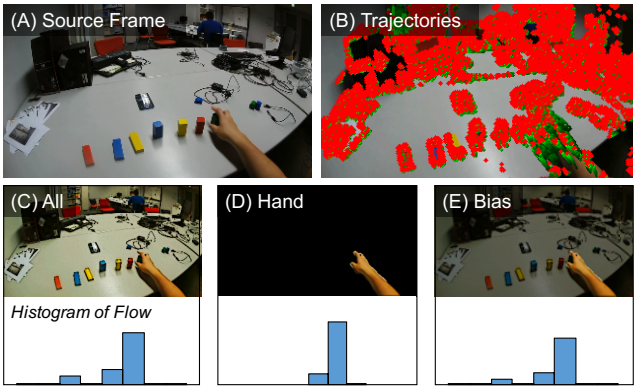

図 3 映像中の各フレームへの特徴記述（特徴点の移動方向ヒストグラム）：(A) 作業映像中の 1 フレーム (B) 特徴点追跡結果 (C) 画像全体からの特徴抽出 (D) 手領域のみからの特徴抽出 (E) 手領域とそれ以外の領域に重み付けをした特徴抽出

示すように，Dense Trajectories [7] により，15 フレーム間の特徴点を追跡する．各フレームにおける特徴点の始点・終点位置から，各特徴点の移動方向と距離を計算する．特徴点の移動方向に基づきヒストグラムを生成し，正規化したものを各フレームの特徴量とする．

本研究では図 3(C-E) のように，映像中から特徴量を記述する領域として映像全体 (All)，手領域のみ (Hand)，手領域と他領域への重み付け (Bias)，の 3 つを提案する．映像全体では，Dense Trajectories によるすべての特徴点追跡結果から移動方向ヒストグラムを作成する．手領域のみでは Li らの手法 [8] を用いて手領域を抽出し，手領域内の特徴点のみからヒストグラムを作成する．手領域と他領域への重み付けでは，ヒストグラム作成の際に，領域毎に特徴点の距離にバイアスを与える．本手法では手領域の重みを大きくしている（手領域の重みを 1.0 としたとき，それ以外の領域では 0.2 とした）．

3.3 伸縮マッチングによる対応付け

伸縮マッチングにより，記述された各フレームの特徴量から映像間の対応付けをする．作業時間の差異を考慮した対応づけをするために．本研究では伸縮マッチングの一手法である動的計画法によるマッチング（DP マッチング）を用いて対応付けし，ラベルの割り当てをする．それぞれの映像の特徴量からコスト表を作成し，DP マッチングにより最短経路を計算する．得られた経路から，クエリ映像のそれぞれのフレームに，対応付けられたリファレンス映像のフレームの基本動作ラベルを割り当てる．図 4 に伸縮マッチングによりフレームを対応付けし，ラベルを割り当てる模式図を示す．

4. 評価

本手法による基本動作ラベル割り当ての精度を評価するための初期実験を実施した．本実験では評価対象として

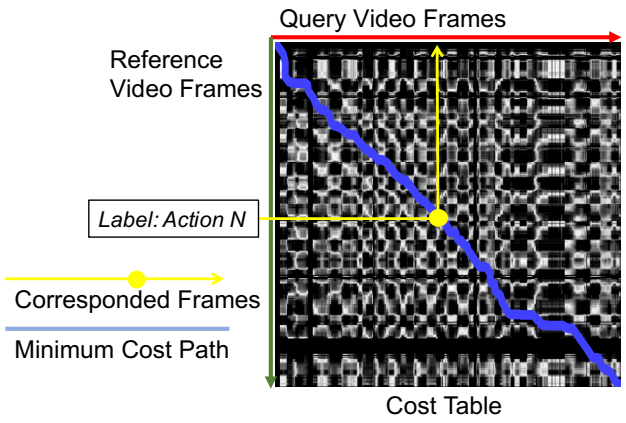

図 4 DP マッチングを用いたフレーム間の対応付け：リファレンス映像のフレームに付けられた基本動作ラベルを，対応付けられたクエリ映像のフレームに伝播させる．

特徴を記述する領域が異なる映像全体 (All)，手領域のみ (Hand)，手領域と他領域への重み付け (Bias) の 3 つを比較した．

4.1 評価対象のデータ

本実験では 1) ブロックを積む，2) インスタントコーヒーを作るという 2 種類の作業映像を評価対象とした（図 5）．それぞれの作業に対し，4 本ずつ映像を撮影した．そのうちの 1 本をリファレンス映像とし，手動によるラベル付けをした（表 1）．精度を評価するために，残りの 3 本のクエリ映像にもリファレンスと同様のラベル番号を用いてラベルを割り当てた．映像中に置いて手が動いていない，もしくは手が映っていないフレームに関しては，動作なしのラベルを割り当てた．

「ブロックを積む」映像は，2 名の作業者が 2 本ずつ作業映像を撮影した．すべての映像で，同じ形状を同じ数のブロックで形成させている．ブロックを設置する順番に関しては，入れ替わりが発生している．本作業では，すべて右手でブロックを操作している．

「インスタントコーヒーを作る」映像も，2 名の作業者が 2 本ずつ作業中の映像を撮影した．しかし，同じ作業者により撮影された 2 本の映像のうち，1 本は作業の中に砂糖を入れるという工程が含まれており，もう 1 本には含まれていない．リファレンス映像には砂糖を入れる工程が含まれているものを選択した．そのためクエリ映像のうち 2 本には，作業工程の抜け落ちが存在する．また，本映像では作業者は両手を使い作業している．

4.2 評価方法

本手法の評価尺度として，アライメントによりクエリ映像に割り当てられた基本動作ラベルと，事前に手動で割り当てたラベルが一致する割合を評価した．クエリ映像中の全フレーム数を F_{all}，動作なしが割り当てられたフレーム数を F_{non}，正しい基本動作ラベルが割り当てられたフレーム数 $F_{correct}$ としたとき，精度 $accuracy$ を 式 (1) のように計算した．

$$accuracy = \frac{F_{correct}}{F_{all} - F_{non}} \quad (1)$$

最適な対応付けとラベル割り当てができた場合，すべてのクエリ映像において $accuracy$ は 100%となる．また，映像の長さを線形に伸縮して，リファレンスとクエリ映像の長さを合わせたときの $accuracy$（Linear Matching Accuracy）は「ブロックを積む」映像で平均 31.2%，「コーヒーを作る」映像で平均 24.2%である．

表 1 実験用映像に割り当てた基本動作ラベル一覧

ラベル	映像 1:ブロックを積む	映像 2:コーヒーを入れる
1	ブロックを取る	物を取る
2	ブロックを手前に移動	物を手前に移動
3	ブロックの位置を修正	物を元の場所に戻す
4	撮影終了ボタンを押す	開ける
5	—	注ぐ
6	—	まぜる
7	—	飲む
8	—	撮影終了ボタンを押す
0	動作なし	動作なし

4.3 結果

図 6 に精度評価の結果を示す．「ブロックを積む」映像では，3 つのクエリへのアライメント精度の平均が，All:50.8%，Hand:69.5%，Bias:56.5%という結果であり，特徴を記述する領域としては Hand が最も精度が高かった．一方で，「コーヒーを作る」映像では精度の平均が，All:76.5%，Hand:65.9%，Bias:79.1%であり，Bias が最も精度が高く，Hand が最も精度が低かった．全体の平均精度は，All:64.5%，Hand:67.7%，Bias:67.8%であり，Hand と Bias の間で差はなかった．

4.4 考察

本手法によるアライメント精度は適した特徴記述領域を選択した場合，「ブロックを積む」映像では 70%程度，「コーヒーを作る」映像では 80%程度であり，初期実験として十分な結果を得られた．特に，コーヒーを作るでは 8 種類 (+動作なし) の基本動作ラベルが付けられていたにも関わらず精度が高かった．この理由としては，映像中に多様な動作があったことにより，同一の基本動作同士が対応付けられやすくなったためであると考えられる．一方で，砂糖を入れる映像と入れない映像をアライメントしたときに，砂糖を入れる工程とポットからお湯を入れる工程が対応付けられてしまうこともあった．これを防ぐためには，操作対象となる物体を考慮した動作の対応付けが必要になる．

ブロックを積む映像では，ブロックを取る，ブロックを移動という 2 つの基本動作ラベルが交互付けられたことに

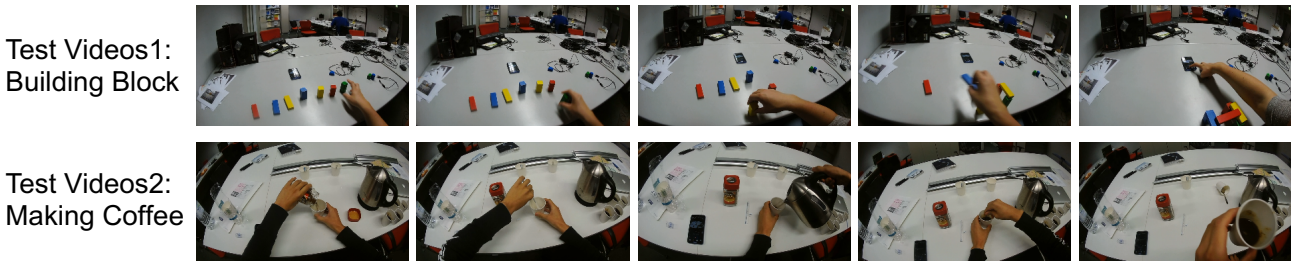

図 5 実験に使った作業映像の例：「ブロックを積む」では右手を使って，ブロックの操作をしている．「コーヒーを作る」では，両手で作業している．それぞれ 4 本の映像を，そのうち 1 本をリファレンス映像とした．アライメントの精度評価のために，クエリ映像に対しても基本動作ラベルを付けた．

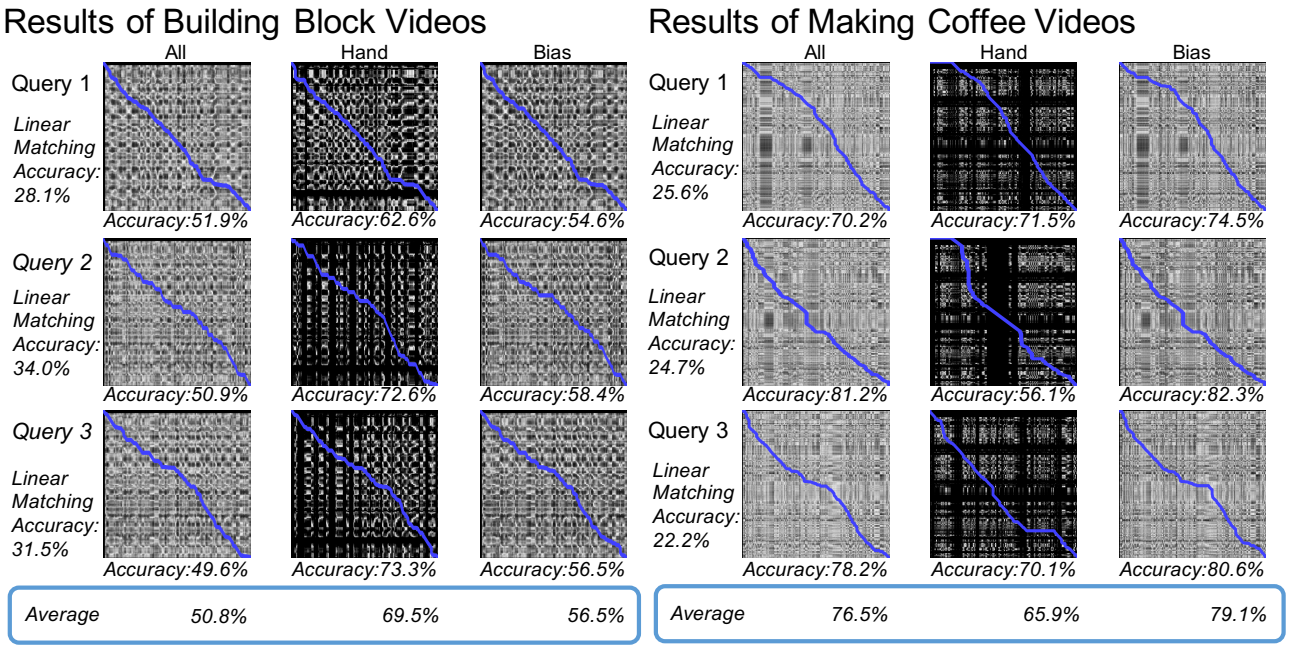

図 6 実験の結果：図左側は「ブロック積み」映像のアライメント精度である．図右側は「コーヒーを作る」のアライメント結果である．ブロック積み映像では，手領域のみからの特徴記述 (Hand) が平均 69.5%と最も精度高くなった．コーヒーを作る映像では，手領域とそれ以外への重み付けによる特徴記述 (Bias) が平均 79.1%で最も精度が高かった．

よって，ラベルの切り替わる境界で誤ったラベルが割り当てられることが多かった．この原因のひとつに，現在のリファレンス映像へのラベル付け方法が完全手動によるものであるということが考えられる．手動だけでは，動作の切り替わるタイミングの判断が曖昧になるため，ラベルの割り当てにおいて動作間の境界で精度が落ちてしまう．

アライメントの精度はブロックを積む映像では特徴記述領域として Hand が最も高く，コーヒーを作る映像では Bias が最も高かった．そのため，今回の実験からは提案した 3 つの特徴記述領域の中で，どの領域が最も手の動作対応付けに向いているかを判断することができないと言える．現在は手の動きのみを記述する特徴を利用しているが，それ以外の特徴とも組み合わせることにより，本アライメント手法に適した特徴記述方法を発見できる可能性がある．

5. 関連研究

一人称視点映像中における手による動作を解析するための研究が盛んに行われている．Li らの研究では一人称視点映像中の手領域を，ピクセル単位で抽出する方法を提案している [8]．Fathi らは，手が操作している物体を抽出する手法実現した [9]．また，映像中における手の握り方に着目し，物体の保持の仕方を分類する研究も行われている [10], [11]．本研究では，手の動作に着目しリファレンスとなる作業映像から，複数の映像に基本動作のラベルを割り当てる手法を提案している．これまでの研究で実現されている操作物体検出や握り方分類を，動作の特徴記述に取り入れることにより，本アライメント手法の精度を向上させることができると考える．

三人称視点映像やモーションキャプチャデータから得ら

れて動作の時系列データを，伸縮マッチングなどを用いて対応づけする研究が行われている [12], [13]．本研究では一人称視点映像中に映る，手動作の時系列データを基に，映像間のアライメントをしている．これまでの三人称視点映像における動作のマッチング手法の知見を取り入れることで，より正確なアライメントが実現できる可能性がある．

6. 今後の課題

6.1 高次の特徴量記述による精度の向上

本論文では，手の動きに着目し，その特徴を記述するために特徴量追跡から得られた移動方向ヒストグラムを利用した．今後は手の形状や位置にも着目した特徴量を利用することにより，より作業映像間の対応づけに適した特徴量があるかを試行したい．また，多様な作業映像で実験をすることにより，作業のカテゴリ毎に適した特徴記述の方法を明らかにしていきたい．

また，今後は操作対象物体の検出と，その特徴記述に取り組みたい．作業映像のアライメントにおいて，同一の物体に対する同様の操作のみが対応付けられるようになれば，ラベル割り当ての精度が高められる可能性がある．そのために，一人称視点映像中における操作対象物体の検出 [9] から得られた結果を，特徴記述領域として利用することを検討している．

6.2 動作切り替わりの自動検出

作業映像中における動作が切り替わるタイミングを，手の動きなどから自動検出をすることによって，基本動作のラベル付けを支援したい．現在，ラベル付けはすべて手動により行っているため，動作切り替わりの判断が曖昧になってしまい，ラベル割り当ての精度が低くなってしまう．自動で正確に動作を切り分けることができれば，提示された動作区間へのラベル付けとなるため，リファレンス映像に一貫したラベルの割り当てができる可能性がある．また，動作区間の切り分けと，そこに割り当てられたラベルが蓄積することにより，動作区間への自動ラベルを割り当ても期待される．

6.3 作業順序の入れ替わりへの対応

実験で使った映像中には，作業工程の抜け落ちは存在していたが，作業工程の入れ替わりは存在していなかった．順序が入れ替わってしまった場合には，動作の対応が取れずアライメント精度が落ちてしまう可能性がある．そのため今後は，基本動作の遷移モデル作成などによる，作業順序の入れ替わりに対応できるようなアライメント手法に取り組みたい．また，作業工程をいくつかの基本動作の集合として定義することによって，工程中の動作入れ替わりを許容するような方法も検討したい．

7. おわりに

本研究では手による動作を基本単位とした一人称視点作業映像のアライメント手法を提案した．本手法では最初に，リファレンス映像への基本動作のラベル付けをすることにより，クエリとなる映像とフレーム間の対応づけを行い，基本動作ラベルを割り当てる．2 種類の作業映像を対象にした初期実験を実施したところ，適切な特徴記述をすることにより，70–80%程度の精度でラベルの割り当てができることがわかった．今後は，アライメント精度の向上や，動作区間の自動検出，作業順序の入れ替わりへの対応などに取り組みたい．

謝辞 本研究は JST CREST の支援を受けたものである．

参考文献

[1] Kris M. Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast Unsupervised Ego-Action Learning for First-person Sports Videos. CVPR 2011.
[2] Pirsiavash, Hamed, and Deva Ramanan, Detecting activities of daily living in first-person camera views. CVPR 2012.
[3] Ryo Yonetani, Kris Kitani and Yoichi Sato, Ego-Surfing First-Person Videos. CVPR 2015
[4] Fathi, Alireza, Jessica K. Hodgins, and James M. Rehg, Social interactions: A first-person perspective. CVPR 2012.
[5] Oliver Wang, Christopher Schroers, Henning Zimmer, Markus Gross, and Alexander Sorkine-Hornung, Videosnapping: interactive synchronization of multiple videos. ACM Transactions on Graphics (TOG), 33(4), 77.
[6] Xianjun Sam Zheng, Cedric Foucault, Patrik Matos da Silva, Siddharth Dasari, Tao Yang, and Stuart Goose, Eye-wearable technology for machine maintenance: Effects of display position and hands-free operation. CHI 2015.
[7] Wang, Heng, and Cordelia Schmid, Action recognition with improved trajectories. ICCV 2013.
[8] Cheng Li and Kris M. Kitani, Pixel-level Hand Detection for Ego-centric Videos. CVPR 2013.
[9] Fathi, Alireza, Xiaofeng Ren, and James M. Rehg, Learning to recognize objects in egocentric activities. CVPR 2011.
[10] De-An Huang, Wei-Chiu Ma, Minghuang Ma, Kris M. Kitani, How Do We Use Our Hands? Discovering a Diverse Set of Common Grasps. CVPR 2015.
[11] Minjie Cai, Kris M. Kitani, Yoichi Sato, A Scalable Approach for Understanding the Visual Structures of Hand Grasps. ICRA 2015.
[12] Gong, Dian, and Gerard Medioni, Dynamic manifold warping for view invariant action recognition. ICCV 2011.
[13] Kulkarni, Kaustubh, et al. "Continuous action recognition based on sequence alignment." International Journal of Computer Vision 112.1 (2015): 90-114.