修士論文

視線を利用した
二人称視点動作認識

平成 28 年 2 月 提出

東京大学大学院
情報理工学系研究科
電子情報学専攻

48-146451　　村上 晋太郎

指導教員　　佐藤 洋一 教授

Abstract

近年，ウェアラブルカメラの普及に伴いカメラの装着者の視界を記録した一人称視点映像の利用が盛んになっている．一人称視点映像の中では，映像の記録者の視界に他の人物が映り込む場合がある．このようにして得られる映像を二人称視点映像と呼ぶ．ウェアラブルカメラでは，従来利用されてきた固定カメラ等と比較してより詳細な人物映像を得ることができる．そこで，本論文では二人称視点映像中での人物の動作認識に取り組む．

従来の動作認識では，しばしば画面全体の動きから特徴量を生成するという手法が用いられてきた．しかしながら，二人称視点映像中では映像の観測者の頭部運動による背景の動きや第三者の映り込みにより，認識対象の動作とは関係のない動きの情報が特徴量に含まれてしまうことがある．このような問題を回避するためには，映像中で動作認識に重要な部位と重要でない部位を推定し，重要でない部位の影響を抑えつつ特徴量を扱うことが望ましい．

一方，近年ユーザーの視線の向きを計測することができる視線計測機器が従来より安価かつ手軽に利用できるようになった．特にウェアラブルカメラと一体になった視線計測機器を用いることにより，一人称視点映像中でのカメラ装着者の視線位置を計測することが可能である．このような情報は，映像の記録者がどのような場所に注意を向けているのかといった情報を知る手がかりになる．

そこで，本研究ではウェアラブルカメラと一体となった視線計測機器を用い，二人称視点映像の記録者の注視位置に基づいて映像の各位置に現れる動きの重要度を考慮しつつ特徴量を生成することで，動作認識の精度を向上させる手法を開発する．提案手法の評価のにあたって，視線データの付与された二人称視点映像データセットを構築した．実験の結果，視線データを利用することで動作の認識精度が向上することを確認した．

目 次

図 目 次

表 目 次

1　序論

1.1　背景と目的

近年，GoPro[1] や Google Glass[2] といったカメラ機能のついた様々なウェアラブルデバイスの普及に伴い，カメラの装着者から見た視界を記録した一人称視点映像が盛んに利用されている (図 1). 一人称視点映像の中では，映像の観測者の視界に別の人物が映り込む場合がある，このような映像を二人称視点映像と呼ぶ (図 2)．二人称視点映像中では，定点カメラなどの三人称視点映像と比べて対象が大きく映り込むことにより，対象の手元の動きなどの細かい情報を捉えることができる．

この二人称視点映像を利用することにより，人物同士のやり取りの検出や人間関係の推定する研究が行われてきた [1, 6]．検出された動作が頷きであるのか，首をかしげる動作であるのかといった動作種類の認識を行うことができれば，そのやり取りがポジティブな意思表示によるものなのかネガティブな意思表示によるものなのかといったより詳細な分析を得ることができる．

そこで，本研究では特に映像の記録者と他者が会話をしている状況をとりあげ，二人称視点映像からやり取りの中で生じる “呼びかけ” や “顔向け” といった動作を認識する問題に取り組む．文献 [22, 24] に見られるように，動作認識に関する従来手法の多くではまず画面全体から histograms of oriented optical flow(HOOF)[12] や histogram of oriented gradients(HoG)[3] といった局所特徴量を抽出し，それらを bag of features[20, 2] や Fisher vector[16] といったコーディング手法によって処理することで映像全体を表す高次特徴を生成するというアプローチを取る．

しかしながら，二人称視点映像を用いる本研究においては，カメラ装着者の頭部運動によって引き起こされる映像背景の動きが局所特徴群に含まれ，高次特徴の生成に影響を及ぼす可能性がある．また，今回取り扱うような二人称視点映像の記録者と他者との間のやり取りによる動作映像では，同映像中に複数人物が現れる場合にやり取りの相手である人物の動きのみを考慮した高次特徴を生成する方が好ましいと考えられる．本研究では，これらの問題を解決するために映像全体から生成された局所特徴群の中で動作認識において重要なものを推定しつつ高次特徴を生成するというアプローチを考える．

このような中，近年ユーザーの視線方向を計測する視線計測機器が従来より安価かつ手軽に利用できるようになった．特に Pupil Pro[3] などのウェアラブルカメラと一体になった視線計測機器を用いることにより，一人称視点映像中でのカメラ装着

[1] https://gopro.com
[2] https://www.google.com/glass
[3] https://pupil-labs.com

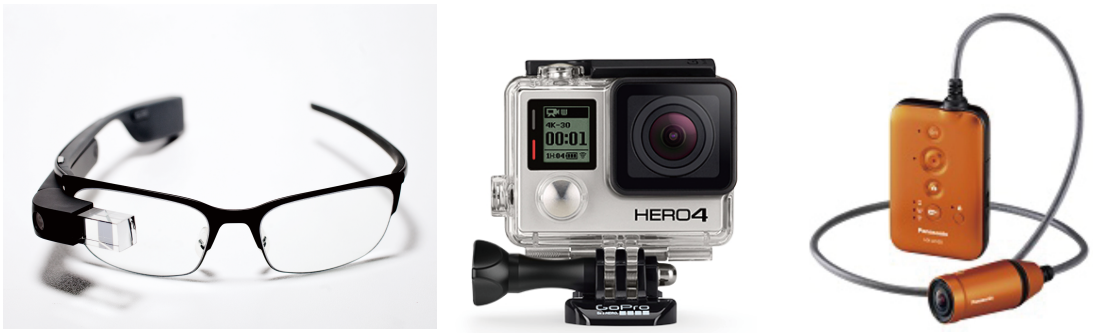

図 1: 一人称視点カメラの例 [4]

者の視線位置を計測することが可能である．このような情報は，映像の記録者がどのような場所に注意を向けているのかといった情報を知る手がかりになる．

そこで本研究では，このようなデバイスで計測された二人称視点映像中の視線位置を用いることで映像中のどの部位に現れる局所特徴が重要であるかを推定するアプローチを提案する．本研究で取り上げるような会話のやり取りの中では，やり取りの相手の人間の動きに映像の記録者の視線が向けられることが期待される．そのため，視線位置の周囲に現れる局所特徴を重要であるとみなすことで重要な局所特徴を推定することができ，上記の背景運動や複数人物の映り込みといった問題を解決することができる．提案手法では，視線位置からある一定の距離の範囲内値から生成された局所特徴を動作認識に有用な局所特徴とし，これらの局所特徴のみを用いて高次特徴を生成することで動作を認識する (図 3)．

本研究の貢献は以下の通りである．

- 二人称視点映像における映像記録者の視線を利用した動作認識手法を開発した．
- 提案手法の評価のために，映像記録者の視線情報を記録した二人称視点映像データセットを作成し，視線を利用した動作認識手法の評価を可能にした．収録したデータセットは 6 人の被験者による 12 種類の動作により構成され，全体で約 1300 サンプルの動作映像を含む．
- 構築したデータセットを用いて，視線を利用した動作認識手法と視線を利用しない動作認識手法を比較した．結果として，二人称視点映像中の対話動作の認識における視線利用の有用性を確認した．

[4]https://www.google.com/glass, https://gopro.com, http://news.panasonic.com/jp/topics/2015/38838.html

図 2: 二人称映像の例．このような映像が映像記録者の頭部に装着された一人称視点カメラにより収録される．本研究では，“呼びかけ”や “挙手” といった映像記録者と相手の間でのやり取りの中で発生する動作の映像を取り扱った．

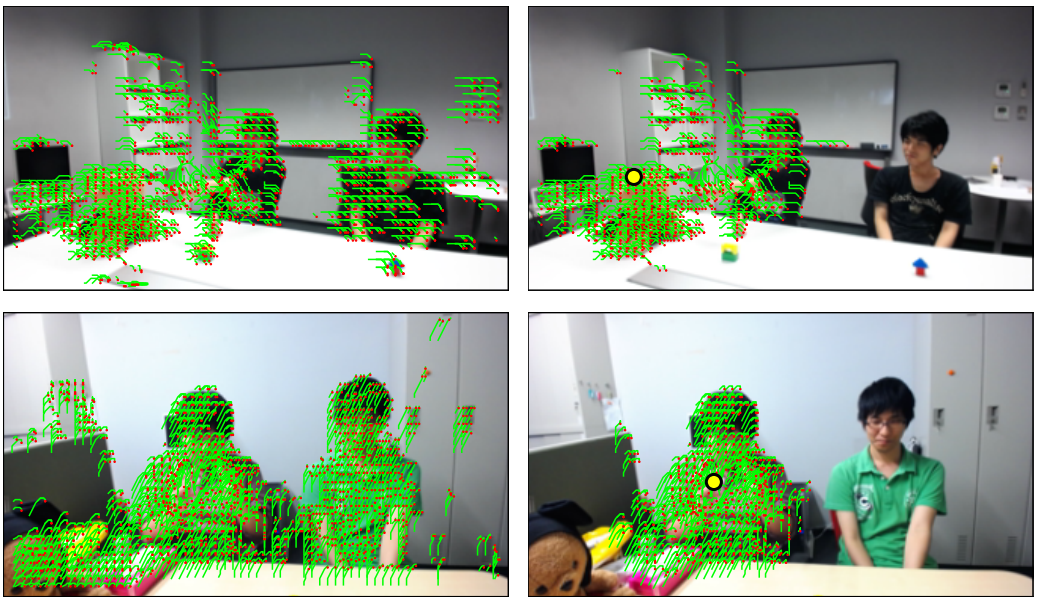

図 3: 視線情報を元にした局所特徴選択の例．緑色の線は dense trajectories[23] という，映像中の特徴点の軌跡を可視化したものである．左列は選択される前の全局所特徴に対応する dense trajectories を，右列は視線情報をもとに選択された dense trajectories を可視化している．右列では視線位置を黄色の丸で示した．

1.2 論文の構成

本論文は全 7 章で構成され，視線情報を利用した二人称視点動作認識の手法を提案しその有用性を検証するものである．

第 1 章では，序論として本研究の目的と背景について述べた．第 2 章では，本研究の関連研究として一人称視点映像や動作認識，視線情報を扱った研究を紹介する．第 3 章では，視線位置を加味した二人称視点映像動作認識手法を提案し，その詳細を説明する．第 4 章では，提案手法の実装の詳細として具体的なパラメタや細かい手法について説明する．第 5 章では，提案手法の評価を目的として収録した視線情報付きの二人称視点映像データセットについて説明する．第 6 章では，提案手法の評価のために行った実験の概要と結果について説明する．第 7 章では，本研究の結論と今後の展望について述べる．

2　関連研究

本研究の提案手法には二人称視点映像の利用，映像中での動作認識，視線情報の利用という要素がある．本章ではそれぞれの要素に関連する先行研究を紹介する．まず最初に 2.1 節で二人称視点映像の利用に関する先行研究研究について紹介する．次に，2.2 節で映像中での動作認識に関する先行研究について紹介する．最後に 2.3 視線情報の利用に関する先行研究について紹介する．

2.1　二人称視点映像の利用

本節では二人称視点映像を利用した先行研究について紹介する．

ある人物 A がウェアラブルカメラを用いて一人称視点映像を記録しつつ，別の人物 B とやり取りを行う状況を考える．このとき，人物 B が映り込んだ一人称視点映像を，B **に対する二人称視点映像**と定義する．

このような二人称視点映像では，従来利用されてきた固定カメラによる三人称視点映像と比較して人物の詳細な映像を得ることができる．こうした特性から，二人称視点映像は人物同士のやり取りや関連性を観測するための手段として活用されてきた．

二人称視点映像を利用した取り組みとして，Fathi らの研究 [6] と Alletlto らの研究 [1] が挙げられる．Fathi らは，被験者の集団がテーマパークを観光する様子を収録した 1 日分の二人称視点映像を分析することで，人物間の相互作用の検出を行った．被験者の集団は 30 人ほどの規模で，そのうち 8 人がヘッドマウントカメラを装着した．この研究では，映像中に映り込んだ人物の頭部位置と角度を推定することで人物間の顔向けや，複数の人物が共通の場所に注目する状態を検出した (図 4)．また，顔向けが発生した時点で人物間でやり取りが生じていると推定することで人物間のやり取りを検出した．加えて，カメラ装着者の視界への人物の映り込みを集計することでどの人物とどの人物が親しい関係にあるか，といった人物関係を推定した．

一方 Alletto らの研究 [1] では二人称視点映像中に現れる人物の頭部位置と向きから人物間の関係を推定し，クラスタリングの手法を用いてどの人物とどの人物が同じグループに属するかというグループ分けを推定した (図 9).

これらの研究では二人称視点映像に現れる人物の頭部の位置と向きを手掛かりに人物同士の関係性を推定することを可能にしている．一方で，やり取りの内容を含めたより詳細な分析のためには，映像中で起きている動作種類の推定まで行えることが望ましい．

二人称視点映像中での動作種類を識別した研究に [15] が挙げられる．本手法ではヘッドマウントカメラを人物に見立てた人形に固定し，“握手をする”や “手を振る”

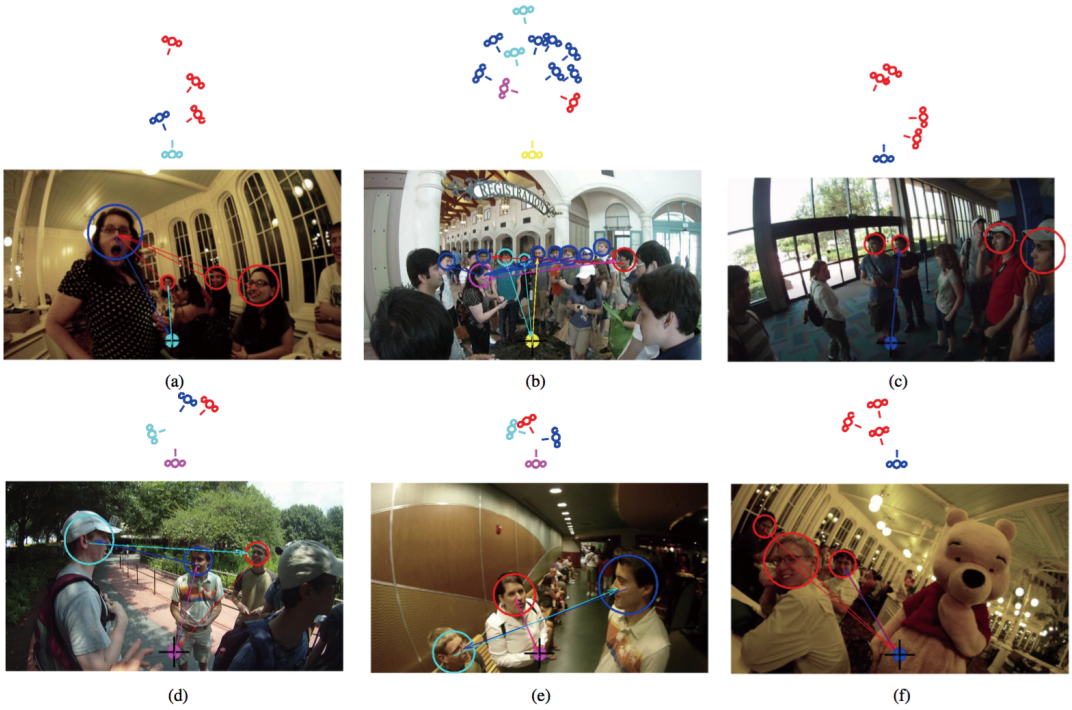

図 4: 二人称視点映像における人物の顔向け解析の例 ([6] より)．Fathi らは二人称視点映像中に現れる人物の頭部位置と向きを推定することで，人物間の相互作用の検出を行った．人物同士の顔向けは映像中の直線で示されている．この顔向けを検出することで人物間のやり取りの発生を推定した．また，映像中で共通の部分に注目している人物は同じ色で表されている．

といった動作を識別した．これらの動作はいずれも映像記録者と他者の間のやり取りを想定したものであった．

Ryoo らの研究では二人称視点映像は人形に固定されたカメラから収録されたものであり，自発的な頭部運動は含まれない．一方，実際の二人称視点からの映像は頻繁に頭部運動の影響を受け頭部運動から生まれる動作特徴が識別精度に影響をもたらす可能性がある．そこで本研究では実際に人物の頭部に取り付けられたカメラによる，自発的な頭部運動が含まれる二人称視点映像における動作認識ついて検討する．

2.2 映像中での動作認識

映像中での人物動作の特徴表現としては，映像中の身体の部位を認識した上でその位置と動きから特徴を抽出する部位ベースの手法 [9] と，画面全体から局所特徴を

図 5: 二人称視点映像における人物の顔向け解析の例 ([1] より)．Alletto らは二人称視点映像中での人物の頭部位置と向きを 3 次元空間上で推定し，その情報からクラスタリングにより人物のグループ分けを推定した．人物の頭部の色は，それぞれの人物がどのグループに所属するかを表している．また，赤色の丸はカメラを装着した人物を示している．

図 6: Ryoo らによる二人称視点映像動作認識のための実験装置 ([15] より)．Ryoo らの研究では，このような人形にウェアラブルカメラを装着することで二人称視点映像からの動作を収録した．

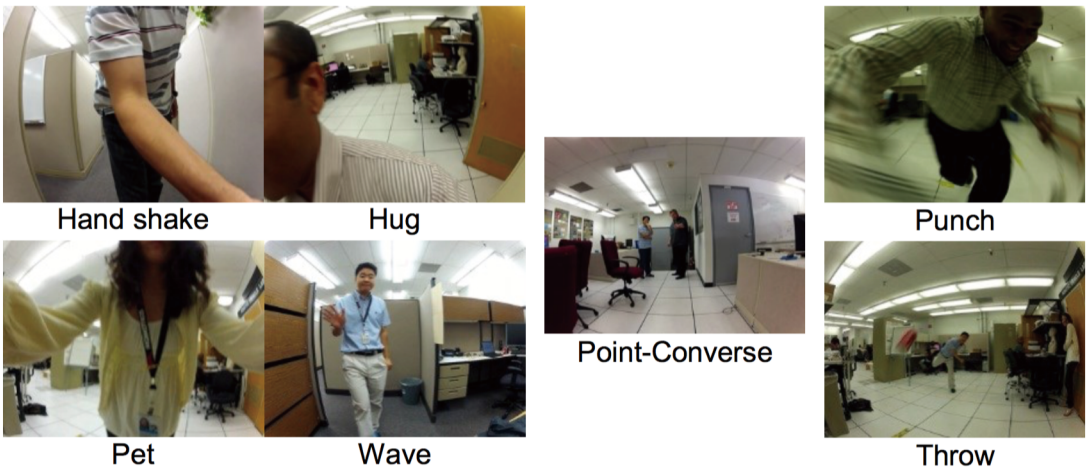

図 7: Ryoo らによる二人称視点映像動作認識における動作種類の例 ([15] より)．Ryoo らの研究では “握手”，“手を振る” などの 7 種類の動作を選定し，動作認識を行った．これらの動作はいずれも，カメラ装着者と相手の間のやり取りによって生じるものである．

抽出した上で Fisher Vector(FV)[16] などのコーディング手法を用いて高次特徴を生成する局所特徴ベースの手法 [22, 23, 24] が挙げられる．この中で部位ベースの手法は部位の認識に失敗した場合にそれ以降の処理が全て失敗してしまうため，観察対象の対象の激しい動きや画面外への離脱，視点移動などの影響を受けやすい．こうした性質から部位ベースの手法は，常に装着者の頭部運動の影響を受けるヘッドマウントカメラ等の映像を扱う際に必ずしも有効に働くとは限らない．そこで本研究では局所特徴ベースの手法を採用し動作認識に使用した．

特徴ベースの動作認識手法としてこれまでに on space-time interest points[11], dense trajectories(DT)[22], improved dense trajectories(IDT)[23] などが提案されてきた．この中でも IDT はカメラ運動がある場合にも利用可能な手法で，カメラ運動を推定しながら特徴抽出が行われる．IDT では optical flow[8] を用いて映像中の特徴点を追跡し，その周辺の histograms of oriented gradients(HoG)[3], histgrams of oriented optical flow(HOOF)[12], motion boundary histogram(MBH)[4] などの局所特徴量を抽出する．この中で HOOF や MBH といった optical flow から算出される特徴量を計算する際に，カメラ運動に起因する optical flow のバイアスを推定しキャンセルしてから特徴量を算出する．このことによりヘッドマウントカメラ等でも頑健に特徴中抽出を行うことができる．

IDT では，HoG, HOOF, MBH といった手法を用いて局所特徴の抽出を行った．一方，静止画像における一般物体認識ではこのような人手により設計された特徴表現に代替する手法として convolutional neural network(CNN)[10] が注目されている．

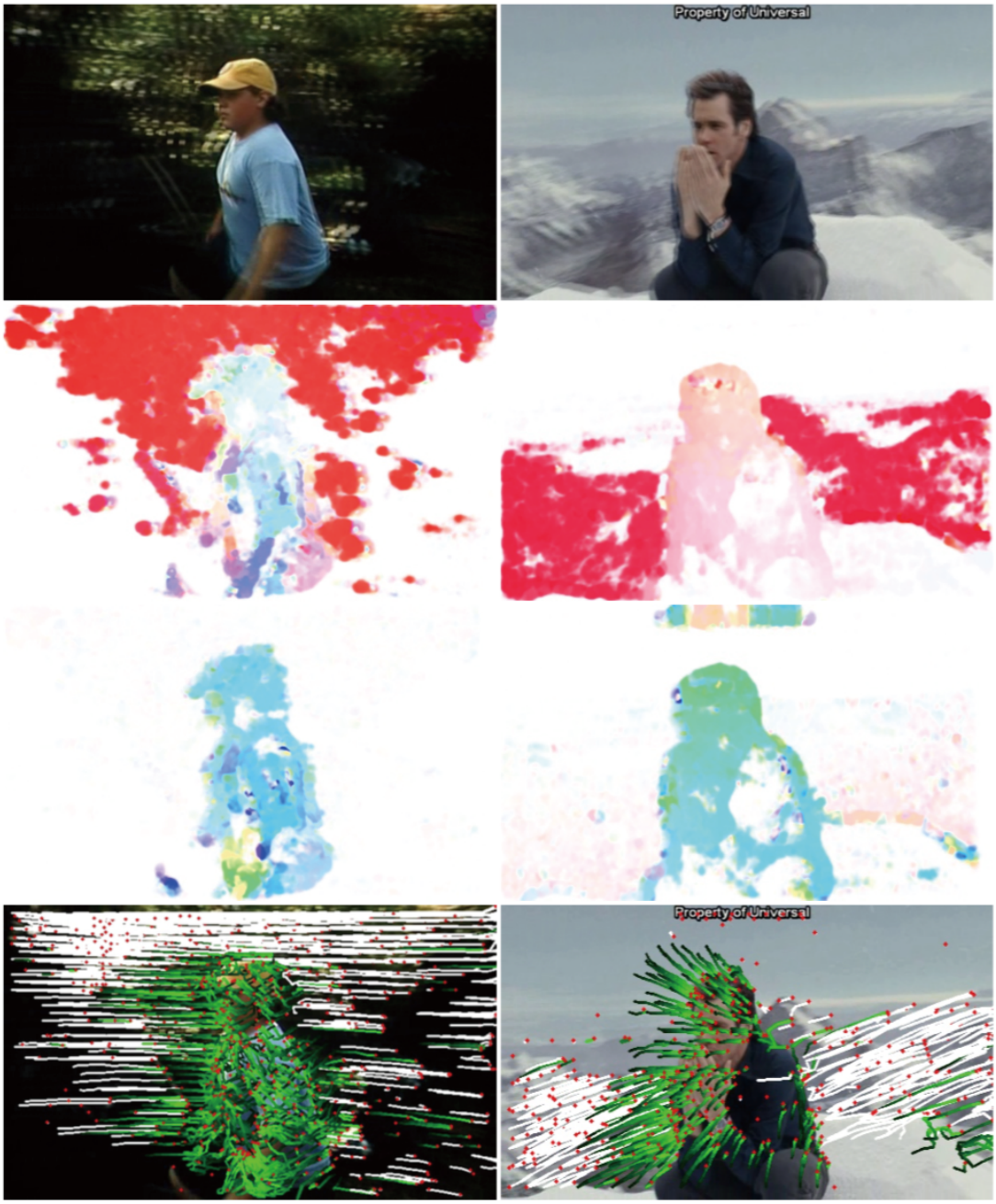

図 8: Improved dense trajectories(IDT) の概要 ([23] より)．IDT では，局所特徴抽出の際にカメラ運動の影響を差し引いた optical flow を用いることにより，カメラ運動がある映像でも頑健な特徴抽出を可能にした．図の 1 行目が入力映像，2 行目が入力映像から抽出された optical flow，3 行目がカメラ運動の影響を差し引いた optical flow，4 行目が最終的に生成された IDT の軌跡である．4 行目では IDT の特徴点の座標を赤点，特徴点の軌跡を緑線，カメラ運動の推定により取り除かれた特徴点の軌跡を白線で表している．カメラ運動による背景部分の IDT が取り除かれていることが確認できる．

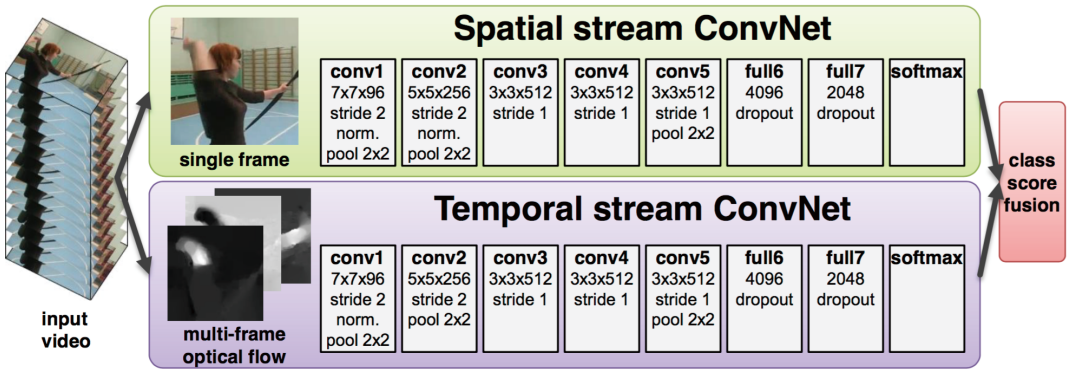

図 9: two-stream convolutional neural networks(TCNN) の概要 ([19] より)．TCNN では，spatial stream ConvNet と temporal stream ConvNet の 2 つの多層ニューラルネットにより，入力の映像とその optical flow それぞれ別に処理する．最終的に 2 つの多層ニューラ r ネットの結果を統合することにより，動作種類の識別等の結果を得る．

CNN では識別に用いる多層ニューラルネットのそれぞれの層が特徴抽出器の役割をしているため，特徴抽出の設計自体を学習することができる．

このような CNN の特性を動作認識に利用する手法として two-stream convolutional neural networks(TCNN)[19] が挙げられる．TCNN では，映像と映像から抽出された optical flow の 2 つを入力として用いる．この 2 つの入力を 2 つの CNN でそれぞれ独立して処理し，最後に結果を統合して用いることで CNN を用いて動作認識を行うことを可能にした．

TCNN の特徴表現を学習する能力と IDT のカメラ運動に対する頑健性を組み合わせた手法として trajectory-pooled deep-convolutional descriptors(TDD)[24] が提案された．TDD では IDT で使用する HoG, HOOF, MBH の代わりに TCNN で抽出された特徴マップを利用することで，従来の IDT と比較してより高精度の動作認識が可能となっている．本研究ではこの TDD を用いて局所特徴を生成する．

このようにして生成された局所特徴量群から高次特徴量を生成した上で識別に用いる．[22], [23] では映像全体の局所特徴から bag of features[20, 2] や Fisher vector[16] を用いて高次特徴量を生成した．一方，背景の影響や他の人物の映り込みの影響に対する頑健性のためには，映像全体の局所特徴を用いるのではなく有用な特徴量を選択しつつ利用する手法が望ましい．

[13] では局所特徴を座標と時間位置を元にクラスタリングで複数の部位に分け，生成されたそれぞれの部位について重要度を推定した．この重要度を元にそれぞれの局所特徴に重み付けを行い動作認識を行った．これに対して提案手法では映像記録者が会話の中でどこに注目を向けているかという視線情報を得ることができる．そ

こで，視線情報を用いて局所特徴の重要度を推定し動作認識の精度を向上できるか検討する．

2.3 視線情報の利用

近年，視線計測装置の小型化，低コスト化に伴い視線情報の研究利用が盛んに行われるようになった．これらの研究により，様々な識別タスクにおける視線情報の有用性が示されてきた．

Yun ら [25] は静止画像中での一般物体認識に視線の情報を用いる手法を提案した．Yun らは被験者が静止画像を見る際の視線の動きを計測した．また同時に，画像中に物体がどのように写り込んでいるかについて自然言語により説明された文章を被験者から集めた．このようにして集められた視線情報，言語情報と画像情報を統合することで静止画像に写り込んでいる物体の種類と位置を識別した．

視線情報を利用した動作認識に関する研究としては Fathi らの [7] や Shapovalova らの手法 [17] が挙げられる．Fathi らはヘッドマウントカメラによる一人称視点映像と，ヘッドマウントカメラに設置された視線計測機器による視線情報を利用することで “料理”や “歯磨き” といった手もとを見ながら行う日常動作を学習した．Fathi らは一人称視点映像中で視線が止まる注視点を検出し，その注視点周辺の画像特徴を抽出することで動作特徴を生成した．

一方，Shapovalova ら [17] は他人の動作を観察する三人称視点映像中での動作認識手法を提案した．ここで言う三人称視点映像とは，スポーツ競技映像など固定カメラから撮影された映像のことである．Shapovalova らは三人称視点映像中で動作認識の際に，映像を被験者に見てもらい，映像閲覧者の画面上での視線の位置を計測することで映像中の動作の位置情報を推測しつつ動作を識別した．

このように一人称視点，三人称視点からの視線情報を利用した認識手法が研究されてきた．これに対して本研究では，二人称視点での視線を利用した認識手法について検証する．

3　提案手法

本章では，提案手法の詳細について述べる．図 10 は，提案手法の概要を示したものである．提案手法では動作が起きるタイミングは既知のものであるとして，一つの動作が含まれた短い二人称視点映像についてその映像中でどのような動作が行われているかを識別する．入力映像の長さは動作の起点から動作の終了までの長さと大まかに等しくなるように編集されている．また，映像中には動作認識の対象となる人物が全フレームにわたって映り込んでいるように収録されている．映像の各フレームにはそのフレームが記録された時点での映像内での映像記録者の視線位置が与えられている．

提案手法ではまず映像全体から trajectory-pooled deep-convolutional descriptors(TDD)[24] を用いて局所特徴群を抽出する．この局所特徴群は，各映像につき多数生成される．次に，視線情報を用いてそれぞれの局所特徴の重要度を推定し，局所特徴に重み付けを行う．重み付けされた局所特徴を用いることにより，Fisher vector(FV)[16] を用いて高次特徴量を生成する．高次特徴は，各映像につき一つ作成される．最終的に生成された映像特徴に線形識別器を利用した多クラス分類手法を適用することで，それぞれの映像に含まれる動作があらかじめ定義された動作の種類のうちどの種類に該当するのかを識別する．

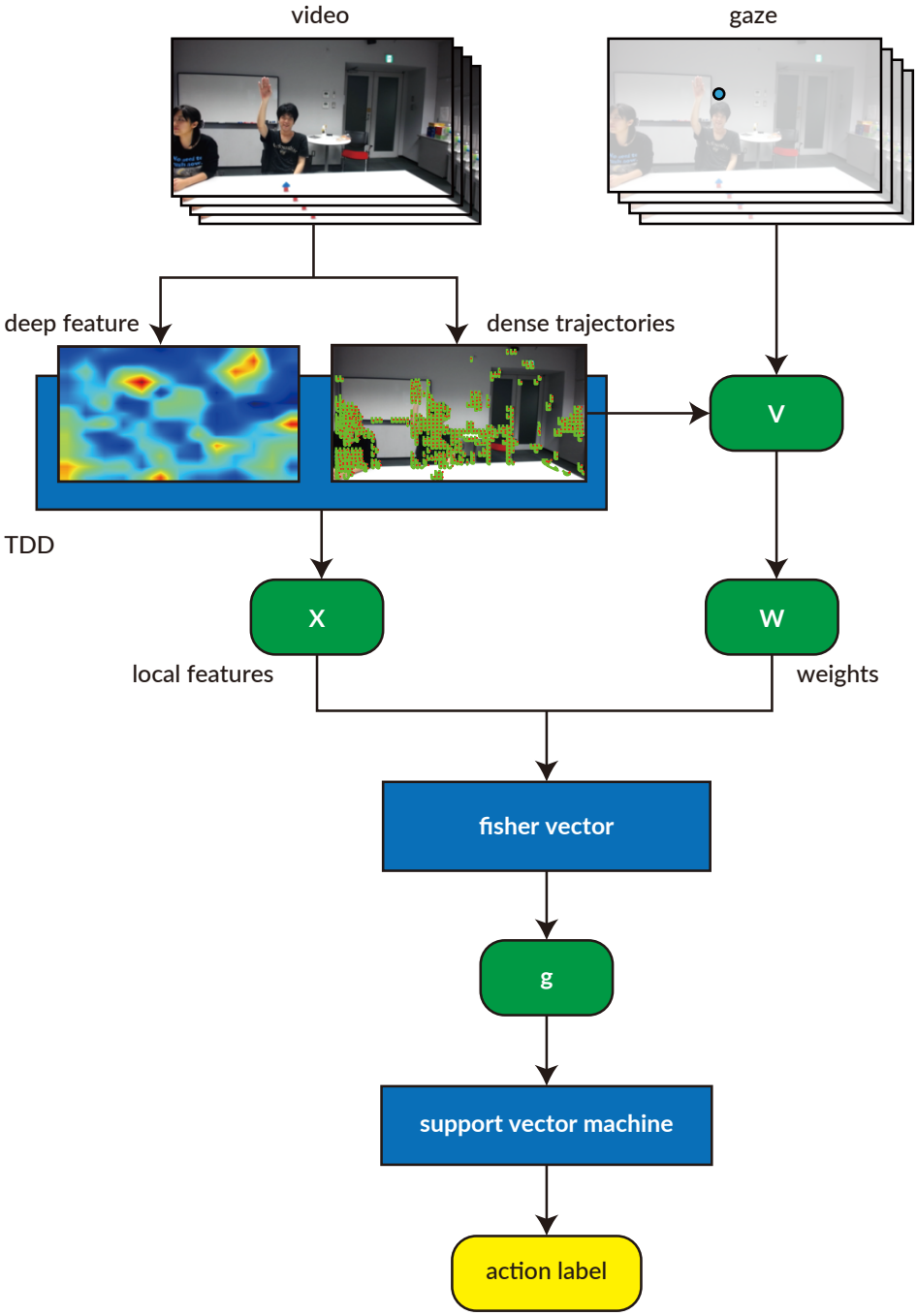

図 10: 提案手法の概要．まず入力映像から trajectory-pooled deep-convolutional descriptors[24] を用いて局所特徴群を生成する．この際に，convolutional neural networks[10] を用いて生成された特徴マップを dense trajectories[22, 23] の軌跡に沿って抽出する．次に，入力映像に付与された映像記録者の視線位置情報をもとに，生成された局所特徴に重み付けを行う．このようにして生成された重みつきの局所登頂群から Fisher vector[16] を用いて高次特徴を生成する．最後に，生成された高次特徴からその動作種類を識別する．

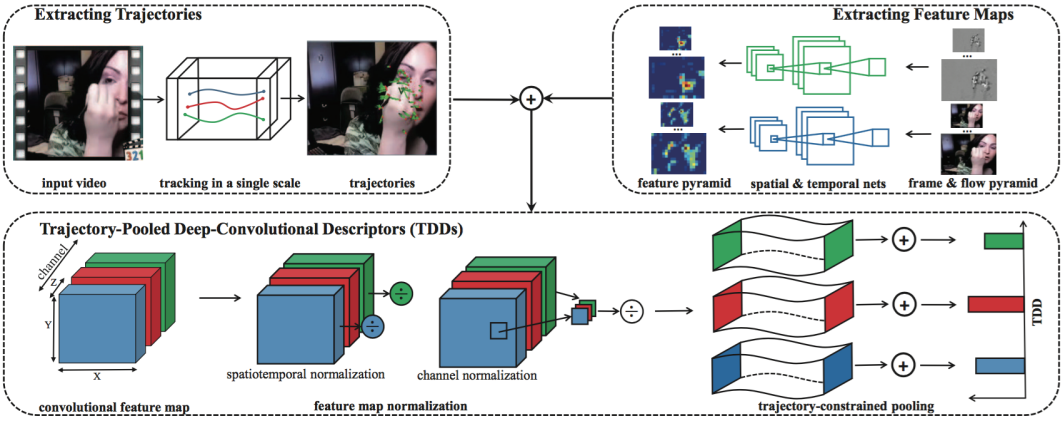

図 11: trajectory-pooled deep-convolutional descriptors(TDD) の概要 ([24] より)．TDD では，まず dense trajectories[22] を用いて特徴点を追跡した軌跡を生成する．一方，入力の画像とその optical flow から多層ニューラルネットを用いて特徴マップを生成する．この特徴マップを dense trajectories を用いて抽出することで各特徴点に対応する特徴量を生成する．

3.1 局所特徴の生成

本節では，提案手法における局所特徴の生成について述べる．本研究では，局所特徴群を生成するための手法として trajectory-pooled deep-convolutional descriptors (TDD)[24] を採用する．TDD は，dense trajectories[22, 23] と convolutional neural networks (CNN)[10] を組み合わせた手法である．Dense trajectories は特徴の抽出に身体部位の検出などを必要としないため，カメラ運動等に頑健である．Dense trajectories は局所特徴として histograms of oriented optical flow(HOOF)[12] や histogram of oriented gradients(HoG)[3] を用いていたが，TDD はこれらに変わり CNN により動作認識に効果的な特徴量を学習した上で使用することができる．TDD の概要を図 11 に示す．本節では，まず TDD のベースとなる手法として dense trajectories について説明する．次に，dense trajectories と CNN を組み合わせて TDD を生成する方法について述べる．

3.1.1 Dense trajectories

Dense trajectories は映像中での動作認識において広く使われてきた特徴表現であり，映像中の特徴点群の軌跡に沿って局所特徴を抽出する手法である．

Dense trajectories では，時間間隔 s_t で以下のような特徴点群 $\mathbb{P}$ を生成する．

$$\mathbb{P} = \{P_k \ | k = 1, 2, ..., K'\} P_k = (x_k, y_k) \in \mathbb{R}^2$$

ここで，K' は生成された全特徴量の数である．

特徴点群は，一定の間隔 s_p で映像中にグリッド上に配置される．抽出された特徴点群は，optical flow[8] のマップに従ってその位置を更新するが，模様のない壁などでは optical flow が計算できないため，初期位置がそのような場所に配置された特徴点群は追跡することができない．そこで，生成された特徴点群のうち追跡可能な領域にあるものを選定し，それ以外は追跡せずに破棄する．

特徴点群の初期位置が追跡可能な領域にあるかどうか判定するためには，Good Features to Track[18] で提案された手法を利用する．この手法ではまず，特徴点の周囲の正方形の近傍領域 $S(p)$ に対して勾配行列 M を以下のように計算する．

$$M = \begin{pmatrix} \sum_{S(p)} (dI/dx)^2 & \sum_{S(p)} (dI/dx dI/dy) \\ \sum_{S(p)} (dI/dx dI/dy) & \sum_{S(p)} (dI/dy)^2 \end{pmatrix}$$

この特徴点行列 M の最小固有値が一定の閾値以上であるかどうかで，与えられた特徴点が追跡可能な領域であるかどうか判定する．

生成された特徴点は，入力映像の動きを追跡するようにその位置を更新する．特徴点 $P_t = (x_t, y_t) \in \mathbb{R}^2$ の更新式は，入力映像の optical flow の場 $\omega = (u_t, v_t)$(図 12) を用いて以下のように表される．

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(\overline{x}_t, \overline{y}_t)} \tag{1}$$

ここで M はメジアンフィルタのカーネル，$(\overline{x}_t, \overline{y}_t)$ は整数に丸められた座標値である．この更新式により，optical flow のベクトルの示す方向に移動する形で特徴点の軌跡 (trajectories) が得られる．特徴点群 $\mathbb{P}$ は，p フレームの間更新され，長さ p の軌跡 T を生成する．

このようにして，与えられた映像から軌跡の集合である Dense trajectories

$$\mathbb{T} = \{T_1, T_2, .., T_K\}$$

を生成する．K は軌跡の数であり，k 番目の軌跡は

$$T_k = \{(x_1^k, y_1^k, t_1^k), (x_2^k, y_2^k, t_2^k), ..., (x_p^k, y_p^k, t_p^k)\}$$

と表される．$(x_p^k, y_p^k, t_p^k) \in \mathbb{R}^3$ は軌跡 T_k の p 番目の位置と時間である．映像から生

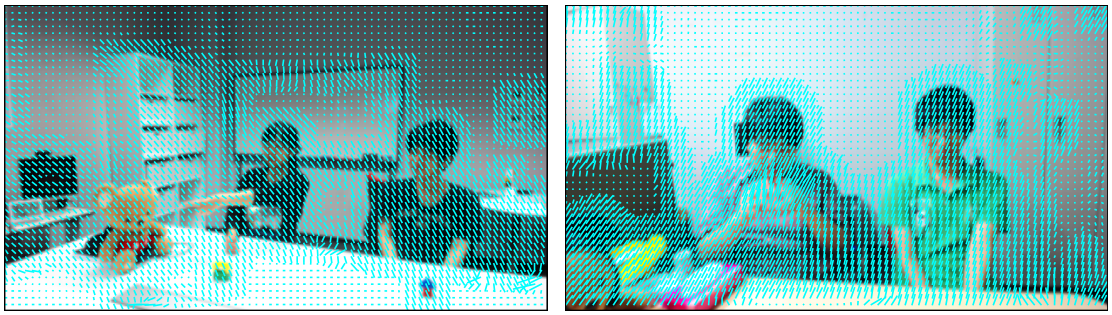

図 12: Optical flow 場 [8] の例．4 ピクセルごとに，その座標における optical flow のベクトルを青線で描画した．このベクトルに応じて特徴点の位置の更新式を得る．

成された Dense trajectories の例を図 13 に示す．

Dense trajectoires の周囲の局所特徴を抽出することで各軌跡に対して一つの特徴ベクトルが計算される．従来の dense trajectories では histograms of oriented gradients(HoG)[3], histgrams of oriented optical flow(HOOF)[12], motion boundary histogram(MBH)[4] 等の局所特徴を集めていた．それに対して TDD では，CNN の中間データとして得られる特徴マップを dense trajectories で抽出する．

3.1.2 Trajectory-Pooled Deep-Convolutional Descriptors

TDD では，CNN を用いて映像から特徴マップを生成する．CNN を用いた特徴マップの生成は，two-stream convolutional neural networks (TCNN)[19] で提案された手法を用いて行われる．

TCNN では，入力映像の各フレーム画像と各フレームの optical flow マップにそれぞれ CNN を適用することで，多層ニューラルネットワークの枠組みで見えの情報と動きの情報を両方加味した学習を行うことができる．この時に用いる二つの CNN を，それぞれ spatial net と temporal net と定義する．TDD では，この二つの CNN の中間層に現れる特徴マップを用いることで，CNN を特徴抽出器として用いる．映像 V を入力として生成される特徴マップは，以下のように表される．

$$\mathbb{C}(V) = \{C_1^s, C_2^s, ..., C_M^s, C_1^t, C_2^t, ..., C_M^t\}$$

ここで $C_m^s \in \mathbb{R}^{H_m \times W_m \times L \times N_m}$ は m 番目の spatial net に現れる特徴マップであり，H_m は高さ，W_m は幅，L は映像のフレーム数，N_m はチャネル数である．同様に，$C_m^t \in \mathbb{R}^{H_m \times W_m \times L \times N_m}$ は m 番目の temporal net に現れる特徴マップである．

抽出された特徴マップの例を図 14 に示す．このようにして得られる特徴マップの

図 13: 生成された dense trajectories の例．収録したデータセット (第 5 章参照) の “指差し”,“呼びかけ”,“考える”の 3 つの動作サンプルを，9 フレームごとに可視化した．赤が特徴点の座標，緑が特徴点の軌跡である．

一要素 C_m^a に対して，軌跡 T_k によって生成される特徴量は以下のように表される．

$$D(T_k, C_m^a) = \sum_{p=1}^{P} C_m^a(\overline{(r_m \times x_p^k), (r_m \times y_p^k), z_p^k})$$

(x_p^k, y_p^k, z_p^k) は軌跡 T_k の p 番目の位置であり，r_m は m 番目の特徴マップと入力映像のサイズ比である．また，$a \in \{s, t\}$ は spatial net か temporal net のいずれかを表す．

このようにして，それぞれの特徴点の軌跡 T_k に対して特徴

$$D(T_k) = \{D(T_k, C_1^s), D(T_k, C_2^s), ..., D(T_k, C_M^s), D(T_k, C_1^t), D(T_k, C_2^t), ..., D(T_k, C_M^t)\}$$

を得ることができる．TDD では，この $D(T_k)$ 局所特徴として識別タスクに利用する．

3.2 視線情報を用いた局所特徴の重み付け

提案手法では，二人称視点映像の記録者の視線位置を用いることで局所特徴に対して重み付けを行う．視線位置を用いた局所特徴選択の概要を図 15 に示す．

3.1 節で取り上げた手法により入力映像映像全体から生成された局所特徴の集合を以下のように定義する．

$$X = \{\boldsymbol{x}_n | n = 1, 2, ..., N\} \ (\boldsymbol{x}_n \in \mathbb{R}^d) \tag{2}$$

N は生成された全局所特徴の総数，d は局所特徴の次元数である．

それぞれの X の n 番目の局所特徴 x_n について，対応するの軌跡の t フレーム目での位置を $\boldsymbol{l}_n(t) \in \mathbb{R}^2$ とおく．一方で，そのフレームでの視線位置を $\boldsymbol{l}_g(t) \in \mathbb{R}^2$ とおく．それぞれの局所特徴の視線との位置関係

$$\boldsymbol{v}_n(t) = \boldsymbol{l}_n(t) - \boldsymbol{l}_g(t) \tag{3}$$

について，$v_n(t)$ の軌跡を以下のように表す．

$$V_n = (\boldsymbol{v}_n(1), \ldots, v_n(T)) \tag{4}$$

この V に対して，以下のような関数 f を定義する．

$$f(V) = \begin{cases} 1 & (g(V, r) \geq q) \\ 0 & (g(V, r) < q) \end{cases} \tag{5}$$

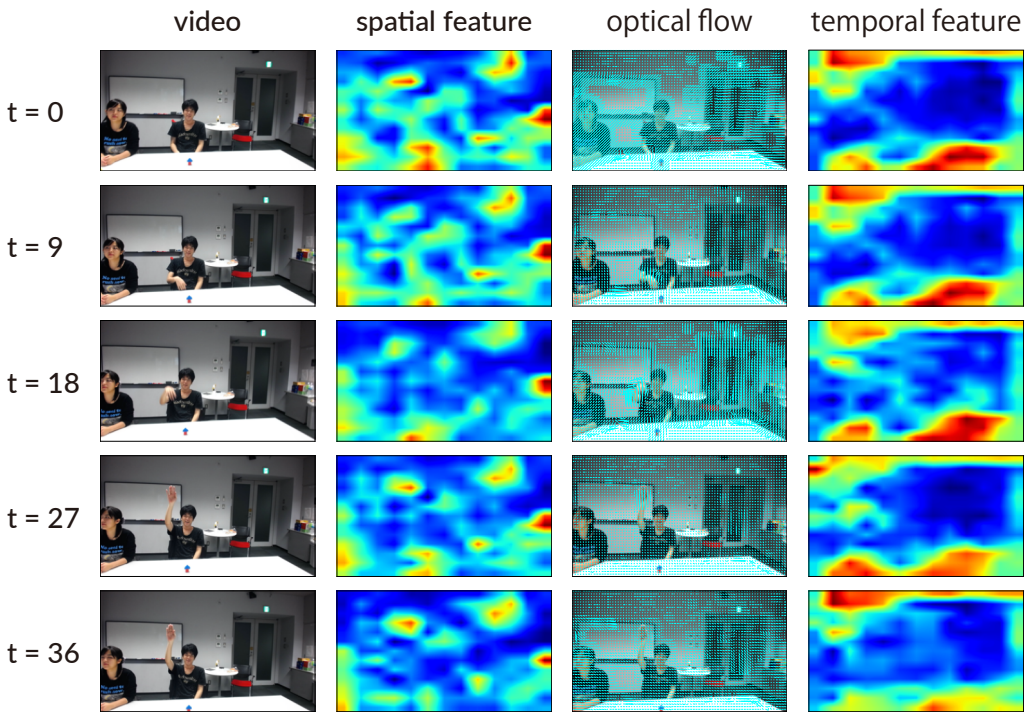

図 14: 多層ニューラルネットにより抽出された特徴マップの例．入力された映像と optical flow から convolutional neural networks(CNN)[10] により抽出された特徴の例を示す．左から数えて 1 列目は入力の映像を示す．2 列目は spatial net により入力映像から抽出された特徴マップの 512 チャネルのうち一つを可視化したものである．特徴マップには spatial net の 6 層目の出力を用いた．3 列目は入力の映像から生成された optical flow を示す．4 ピクセルごとに，optical flow のベクトルを水色の線で可視化した．4 列目は，temporal net により optical flow から抽出された特徴マップの 512 チャネルのうち一つを可視化したものである．特徴マップには temporal net の 5 層目の出力を用いた．

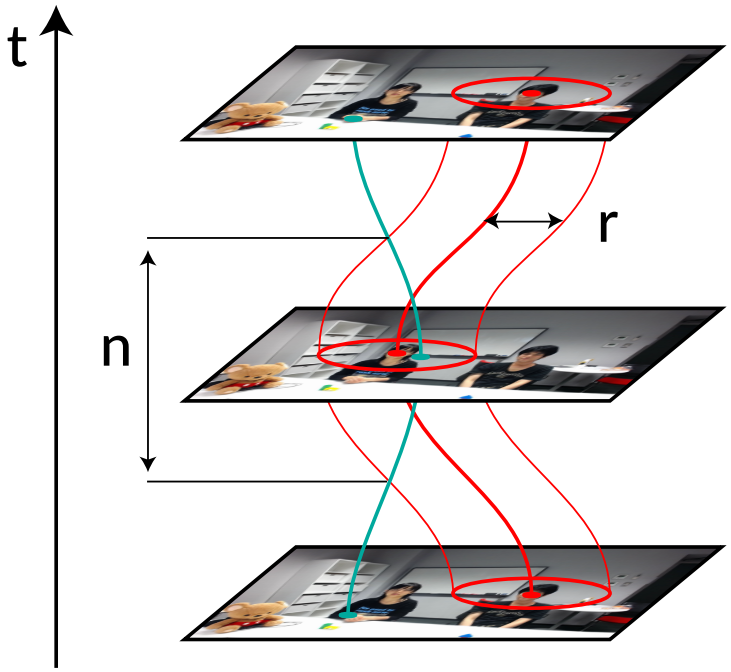

図 15: 視線情報を用いた局所特徴選択の概要．赤線は視線の軌跡を示す．緑線は重み付け対象の局所特徴の軌跡を示す．視線位置から半径 r の領域を**視線領域**と定義する．局所特徴が視線領域内にある時間 n を視線領域との**重複時間**と定義する．この重複時間 n が一定の閾値 q 以上であれば重み $w = 1$，そうでなければ重み $w = 0$ とすることで，視線の周りの局所特徴を選択する．

ここで，$g(V_n, r)$ は，軌跡 V の要素 $\boldsymbol{v}_n(t)$ の中で $|\boldsymbol{v}_n(t)| \leq r$ を満たすものの個数を表す．以上で定義した V_n, f を用いて，X の n 番目の局所特徴 $\boldsymbol{x}_n$ の重要度 w_n は以下のように表される．

$$w_n = f(V_n) \quad (w_n \in \mathbb{R}, w_n \geq 0) \tag{6}$$

提案手法では，視線位置から半径 r 以内の領域を “視線領域” と定義し，$g(V_n, r)$ を特徴点の軌跡と視線領域の “重複時間” と定義する．w_n は局所特徴 $\boldsymbol{x}_n$ の中で視線領域との重複時間が一定の閾値以上であるものを選択する働きを持つ．r は，視線領域をどの程度広くするかを決めるパラメタあり，$r = \infty$ の場合には全ての局所特徴から高次特徴を生成する従来手法に対応する．q は局所特徴 $\boldsymbol{x}_n$ と視線領域の重複時間が何フレーム以上であればその局所特徴を選択するかを決めるパラメタである．これは，どの程度注視された局所特徴を選択するかを調整する働きを持ち，適切に値を設定することで映像中で視線位置が大きく動いた場合に局所特徴が過剰に選択されることを防ぐことができる．

なお，視線計測機機器の誤差によりフレームから視線情報が欠落している場合は $|\boldsymbol{v}(t)| = \infty$, $w_n = 0$ であるものとして扱った．この w_n により選択された局所特徴の例を図 16 に示す．

以上で定義された各局所特徴の重み

$$W = \{w_n | n = 1, ..., N\} \tag{7}$$

を用いて，局所特徴を選択しつつ高次特徴を生成する．

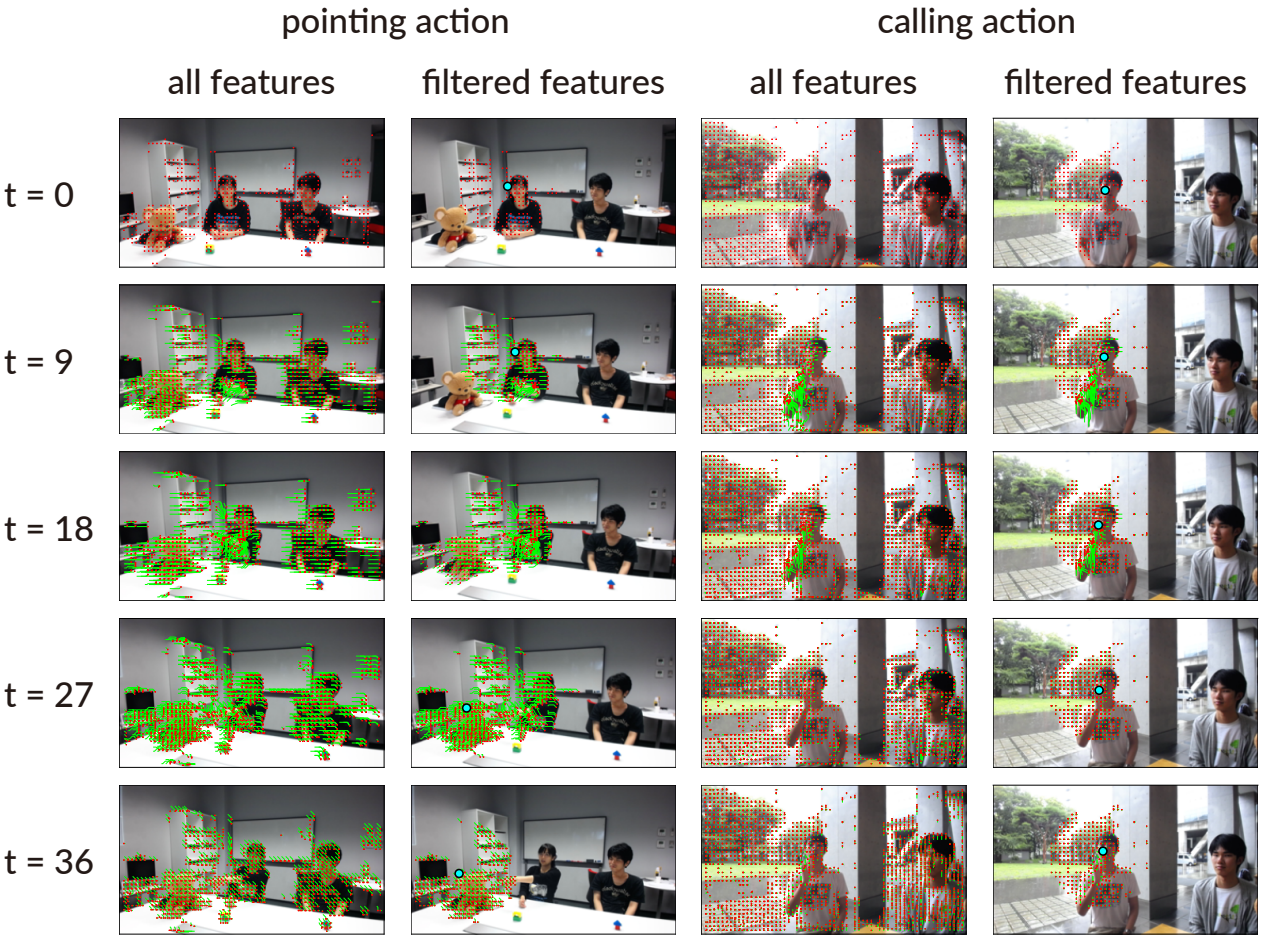

図 16: 視線による局所特徴の重み付けの結果．1, 2 列目では**指差し**の動作について，3, 4 列目では**呼びかけ**の動作について重み付けを可視化した．1 列目，3 列目は重み付けされる前のすべての局所特徴に対応する dense trajectories を描画している．それに対して 2 列目，4 列目は重み付けされた結果 $w = 1$ となった局所特徴に対応する dense trajectories のみを描画している．図中の赤丸は特徴点の位置を，緑線は特徴点の軌跡を，青丸は視線位置をそれぞれ表す．いずれの例も，背景部分や他の人物による局所特徴の影響が抑えられていることを確認できる．

3.3 高次特徴の生成

提案手法では，局所特徴群から高次特徴量を生成するためのコーディング手法として Fisher vector [16] を採用する．Fisher vector を用いることにより，線形識別器による効率的な学習が可能となる．本節では，局所特徴群からの Fisher vector の生成について説明する．

3.3.1 重みを考慮した Gaussian mixture model の学習

Fisher vector では，まず教師データの映像サンプルから局所特徴を生成し，その局所特徴の生成モデルを Gaussian mixture model(GMM) で学習する．GMM では局所特徴 $\boldsymbol{x} \in \mathbb{R}^d$ の生成確率を以下の式で表す．

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_k \pi_k \mathcal{N}(X_n|\boldsymbol{\mu}_k, \Sigma_k) \tag{8}$$

ここで $\mathcal{N}$ は正規分布の確率密度関数を表す．$\boldsymbol{\theta}$ は GMM のパラメタであり，以下の式により定義される．

$$\boldsymbol{\theta} = (\pi_1, ..., \pi_K, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K, \Sigma_1, ..., \Sigma_K) \tag{9}$$

GMM の k 番目のコンポーネントの正規分布に対して $\boldsymbol{\mu}_k \in \mathbb{R}^d$ は平均を，$\Sigma_k \in \mathbb{R}^{d\times 2}$ は共分散行列を表す．$\pi_k \in [0, 1]$ は k 番目のコンポーネントに割り当てられた重みを表し，$\sum_k \pi_k = 1$ となる．

教師データから得られる局所特徴サンプルの集合

$$X = \{x_n|n = 1, ..., N\} \tag{10}$$

に最もよく合致する GMM のパラメタ $\boldsymbol{\theta}$ は，以下の最大化問題を解くことにより求められる．

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \max_Y \prod_n \prod_k (\pi_k \mathcal{N}(X_n|\boldsymbol{\mu}_k, \Sigma_k))^{Y_{nk}} \tag{11}$$

$Y \in \{0, 1\}^{n\times k}$ は，成分 $\boldsymbol{y}_n$ が局所変数 $\boldsymbol{x}_n$ がどの正規分布から生成されたかを表す行列であり，$\boldsymbol{y}_n$ は一つの成分が 1 でそれ以外は全て 0 のベクトルである．この最適化問題は EM アルゴリズムにより近似的に解くことができる．

この最大化問題に対して，局所特徴の重み

$$W = \{w_n|n = 1, ..., N\} \tag{12}$$

を導入することで，GMM を局所変数の重みを考慮した分布 $\boldsymbol{\theta}_w$ に拡張する．

$$\boldsymbol{\theta}_w = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \max_{Y} \prod_n \prod_k (\pi_k \mathcal{N}(X_n|\boldsymbol{\mu}_k, \Sigma_k))^{Y_{nk} w_n} \tag{13}$$

このように定義することで，$w_n = 0$ の場合は局所特徴 $\boldsymbol{x}_n$ を無視した場合の GMM，$w_n \in \mathbb{N}$ の場合は局所特徴 $\boldsymbol{x}_n$ の個数を w_n 倍に増した場合の GMM に対応させることができる．この重みつき学習 GMM を利用して Fisher vector を計算する．

3.3.2 重みを考慮した Fisher vector の生成

前項で定義した重みつき学習 GMM を用いて Fisher vector を計算する．

事前に学習された GMM のパラメタ $\boldsymbol{\theta}$ に対して, サンプル X から生成される Fisher vector $\mathcal{G}_{\boldsymbol{\theta}}^X$ は以下のように求められる．

$$\mathcal{G}_{\boldsymbol{\theta}}^X = L_{\boldsymbol{\theta}} \boldsymbol{s}(X|\boldsymbol{\theta}) \tag{14}$$

ここで，$\boldsymbol{s}(X|\boldsymbol{\theta})$ はサンプル X の対数尤度勾配であり，以下のように定義される．

$$\boldsymbol{s}(X|\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log p(X|\boldsymbol{\theta}) \tag{15}$$

また，L_θ は Fisher 情報行列

$$F_{\boldsymbol{\theta}} = E_X[\boldsymbol{s}(X|\boldsymbol{\theta}) \boldsymbol{s}(X|\boldsymbol{\theta})^{\mathrm{T}}] \tag{16}$$

について，以下の式が成り立つような行列であり，GMM の分布 $\boldsymbol{\theta}$ に対応して一意に定まる．

$$L_{\boldsymbol{\theta}}^T L_{\boldsymbol{\theta}} = F_\theta \tag{17}$$

本研究ではこの対数尤度勾配 $\boldsymbol{s}(X|\boldsymbol{\theta})$ を局所特徴の重みを考慮できるように拡張することで，視線情報を考慮した高次特徴量を生成する．

サンプル X の尤度 $p(X|\boldsymbol{\theta})$ は以下のように X に含まれる各サンプルの生成確率 $p(\boldsymbol{x}_n|\boldsymbol{\theta})$ の積を用いて以下のように求められる．

$$p(X|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta}) \tag{18}$$

この尤度に局所特徴の重要度 W を加味して以下のように拡張する．

$$p_w(X, W, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta})^{w_n} \tag{19}$$

このように定義することで，$w_n = 0$ の場合は局所特徴 $\boldsymbol{x}_n$ を無視した尤度，$w_n \in \mathbb{N}$ の場合は局所特徴 $\boldsymbol{x}_n$ の個数を w_n 倍に増した場合の尤度に対応させることができる．この重み付け尤度の定義を用いることで，視線情報を加味した重み付け対数尤度勾配 $\boldsymbol{s}(X, W|\boldsymbol{\theta})$ は以下のように求められる．

$$\begin{aligned}
\boldsymbol{s}(X, W|\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \log p_w(X, W|\boldsymbol{\theta}) \\
&= \nabla_{\boldsymbol{\theta}} \log \prod_{n=1}^{N} p(\boldsymbol{x}|\boldsymbol{\theta})^{w_n} \\
&= \nabla_{\boldsymbol{\theta}} \sum_{n=1}^{N} w_n \log p(\boldsymbol{x}|\boldsymbol{\theta})
\end{aligned}$$

最終的に，視線情報を加味した重み付け Fisher vector $\mathcal{G}_{\boldsymbol{\theta}}^{X,W}$ は以下のように算出される．

$$\mathcal{G}_{\boldsymbol{\theta}}^{X,W} = L_{\boldsymbol{\theta}} \boldsymbol{s}(X, W|\boldsymbol{\theta}) \tag{20}$$

提案手法では，この重み付け Fisher vector を用いることにより局所特徴量から高次特徴を生成する．

4　実装

本章では，提案手法の評価のために作成したシステムの実装の詳細について説明する．

4.1　局所特徴の生成

本研究では，映像からの局所特徴の生成に第 2.2 節で説明した TDD[24] を採用した．TDD では dense trajectories により映像中に現れる特徴点を追跡し，その軌跡上の特徴マップを局所特徴として用いる．入力映像のサイズを 320px × 180px，フレームレートを 30fps とする．dense trajectories の生成の際には，映像の 5 フレームごとに画面全体から特徴点を選択し，各特徴点から軌跡を生成した．特徴点の生成の際には 8px 間隔のグリッド上に特徴点の初期位置を配置し，その中で追跡可能なものを選択した上で optical flow[8] により追跡した．

生成された特徴点が追跡可能であるかどうかの判定には Good Features to Track で提案された手法 [18] を採用した．この手法では特徴点の生成の際に初期位置の勾配行列の最小固有値を計算することにより，その特徴点が追跡可能であるかどうか評価する．この最小固有値が一定の値に満たない場合，特徴点は模様のない壁などの追跡することができない領域にあると判定する．提案手法では，この固有値の評価の際に，勾配行列の固有値の閾値を 10^{-4} に設定し特徴点の選定を行った．また，optical flow の生成には Farneback のアルゴリズム [5] を採用し，Dense trajectories の長さは 15 フレームとした．

特徴マップの生成には Wang ら [24] により開発された CNN モデル [5] を使用した．Wang らによる CNN モデルの構成を表 1 に示す．Wang らは，spatial net と temporal net の両方で同じ構成の CNN を用いた．CNN の学習のための教師データには UCF101[21] データセットが用いられている．TDD では spatial net と temporal net で各フレームの画像と optical flow 入力をそれぞれ入力とし，その中間データを特徴マップとして用いる．中間データの利用に際しては，文献 [24] を参考に画像の特徴マップには **conv4** 層の出力を，optical flow の特徴マップには **conv3** 層の出力を採用した．

[5] https://github.com/wanglimin/TDD

表 1: TDD で使用した CNN のモデル構成 (Wang et al. [24] より)

Layer	size	stride	channel	map size ratio	receptive field
conv1	7 × 7	2	96	1/2	7 × 7
pool1	3 × 3	2	96	1/4	11 × 11
conv2	5 × 5	2	256	1/8	27 × 27
pool2	3 × 3	2	256	1/16	43 × 43
conv3	3 × 3	1	512	1/16	75 × 75
conv4	3 × 3	1	512	1/16	107 × 107
conv5	3 × 3	1	512	1/16	139 × 139
pool5	3 × 3	2	512	1/32	171 × 171
full6	-	-	4096	-	-
full7	-	-	2048	-	-

4.2 視線情報の利用

提案手法では局所特徴の重み付けに視線位置の情報を用いる．視線情報の収集には，収録には Pupil Labs 社の Pupil Pro[6] を使用した．

提案手法には視線からどの程度の距離を視線周辺領域にするかという半径 r と，視線領域との重複時間がどの程度の局所特徴を選択するかという閾値 q の二つのパラメタが存在する．そこでこの二つのパラメタを様々に変えた上で，認識精度が最も高くなるような変数の組み合わせを提案手法のパラーメータとして決定した．これらのパラメタは交差検定で決定した．交差検定の結果と採用した値に関しては第 6 章で説明する．

4.3 Fisher vector による識別

提案手法では，線形識別器で効率的に学習を行うために Fisher vector により高次特徴を生成する．

4.1 の項で特徴マップの生成に使用した CNN の conv4 層，conv3 層はそれぞれ 512 のチャネルから構成される．この特徴から dense trajectories により集められる局所特徴は spatial net の conv4 層と temporal net の conv 3 層のチャネル数を合わせた 1024 次元となる．

Fisher vector では Gaussian mixture model の共分散成分を考慮していないため，特徴次元間に相関がないことを仮定している．そのため，Wang らの研究 [24] では，

[6]http://pupil-labs.com

局所特徴をそのまま Fisher vector に適用するよりも主成分分析により次元削減をし，特徴次元間の相関を無くした上で Fisher vector を生成した．提案手法においても，主成分分析により局所特徴の次元を 1024 次元から 64 次元に削減した上で Fisher vector を作成した．

Gaussian mixture model の学習の際には，Wang らの手法に基づきコンポーネント数 $K = 256$ として学習を行った．最終的に，生成された Fisher vector は $64 \times 2 \times 256 = 32768$ 次元となった．

生成された Fisher vector は，Florent らの手法 [14] を用いて正規化した上で使用した．この手法では，Fisher vector $\boldsymbol{x} \in \mathbb{R}^d$ に対して L2 正規化

$$f(x_i) = \frac{x_i}{\sqrt{\sum_k^d x_k^2}}$$

と power 正規化

$$g(x_i) = \mathrm{sign}(x_i)|x_i|^\alpha$$

を適用する．このようにすることで Fisher vector による学習の精度が向上することが示されている [14]．本研究では $\alpha = 0.5$ とした．

識別の際にはサポートベクタマシンを用いた．また，最終的には 12 種類の多クラス分類を行う必要があったため，one vs one classifier を用いることで多クラス識別を行った．

5　データセット

5.1　概要

本研究では，二人称視点映像における映像の記録者の視線位置を利用することにより動作認識の精度を上げる手法を提案する．映像の記録者の視線位置を含む二人称視点映像は存在しないため，今回は提案手法の評価のために新たにデータセットを構築した．データセット収録の際には，被験者が視線計測機器と一体になったウェアブルカメラを装着し，他の被験者とやり取りをするという設定でデータ収集を行った．

本データセットでは人物同士のやり取りで発生すると考えられる 12 の動作種類について，その二人称視点からの動作映像の収録を行った．収録の際には 6 人の被験者が動作認識の対象として参加した．また，2 人の被験者が二人称視点映像の記録者として実験に参加した．最終的に各動作種類で 108 サンプル，全体で 1296 サンプルの動作映像を収録した．収集したデータセットの各動作種類の映像例を図 17 に示す．

5.2　動作クラスの選定

本研究で作成したデータセットでは，人物間のやり取りで発生する動作を 12 種類選定し収録した．人物間のやり取りで発生する動作には，**指差し** や **挙手** などの手振りによる動作と，**うなずき** や **顔向け** といった頭の動きによる動作が考えられる．そこで，収録する動作種類には，手による動作と頭部運動による動作が含まれるように選定を行った．本データセットで収録した動作クラスの一覧を表 2 に示す．

5.3　データセットの収集

本データセットでは，6 人の被験者の動作映像を屋内，屋外を含む 3ヶ所で収録した．二人称視点映像と視線情報は 2 人の被験者から収集し，収録には Pupil Labs 社の Pupil Pro[7] を使用した．データセットの収録の様子を図 19 に示す．

データセットの収録の際には，映像の記録者 (observer) と被験者 2 人が向かい合い座った状態でデータセットの収録を行った．このようにすることで映像の記録者の視界に同時に 2 人の映像が映るようになる．これは視線を利用することで複数人物が同時に別の動作をする，といった場合に認識精度を向上することが可能であるか評価することを目的としている．

収録された動作群は映像の記録者と動作認識の対象者との間のコミュニケーションとして行われるようにしており，動作認識の対象者同士のコミュニケーションと

[7] http://pupil-labs.com

図 17: データセットの映像例 (1/2)．黄丸は視線位置を示す．

図 18: データセットの映像例 (2/2)．黄丸は視線位置を示す．

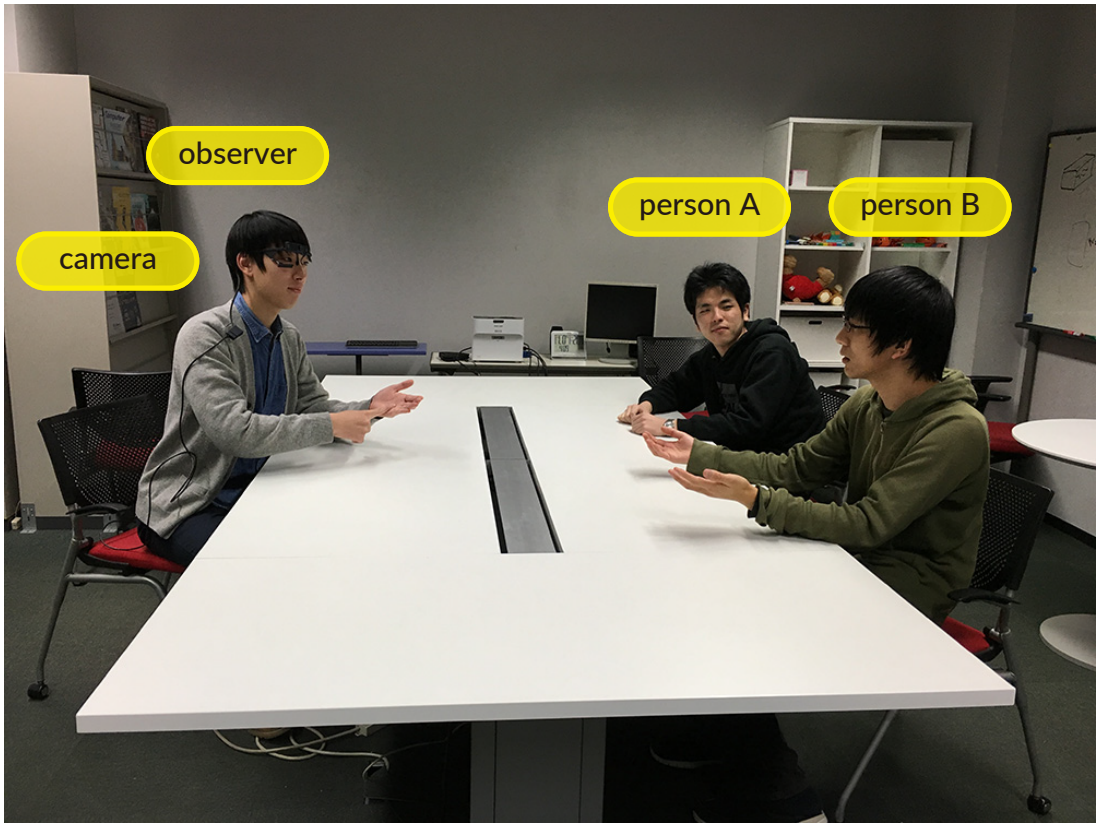

図 19: データセットの収録風景の一例．1 人の二人称視点映像記録者と 2 人の動作認識対象の被験者が互いに向かい合い座った状態でデータセットの収録を行った．映像の記録者の視界には常に 2 人の解析対象者が映り込むようにして収録を行った．

表 2: 本研究で収録した動作映像の種類

身体部位	動作種類
手による動作	指差す
	呼びかける
	主張する
	手を挙げる
	腕組み
	手を頭に乗せる
	物を拾う
手と頭による動作	考える
頭による動作	頷く
	首を傾げる
	こちらを向く
	顔を上げる

した行われた動作は含まれないようにしている．よって映像の記録者の視線は動作認識の対象者との間のコミュニケーションの一部としての振る舞いをする．

データセットの収録の際にはあらかじめ作成した台本を読み上げ，それに従って被験者が動作することで一連の動作群を収録した長い動画を収録した．収録された長い動画の中から，動作が含まれる部分を切り出すことで動作サンプル群を作成した．

5.4 動作サンプルのフォーマット

本データセットにおける二人称視点映像は 30fps のフレームレート，解像度 320px × 180px のフォーマットで収録した．解像度は，一度収録した高解像度の映像の解像度を落とすことにより 320px × 180px に変更した．これは，dense trajectories の生成や CNN による特徴マップの生成を実用的な計算資源で実現するためである．

それぞれの映像サンプルの中には一つの動作が収録されており，映像の長さは動作の起点から動作の終了までの長さと大まかに等しくなるように編集されている．ここでは，例えば “挙手” の動作であれば手を上げる動きが始まる瞬間を動作の起点と定義する．今回収録した動作はほとんどが 1.5 秒で完結する動作であったため，映像サンプルの長さを 1.5 秒に統一した．

5.5 動作サンプルに対するアノテーション

動作認識の教師データとするため，それぞれの動作サンプルに対してその動作種類を付与した．また，教師データとテストデータの間で認識対象の人物を分離する交差検定を可能にするため，各動作サンプルに対してその動作の主体が誰であるかといった人物ラベルを付与した．本研究の実験では比較手法として，人物領域の局所特徴のみを使用する手法を評価する．そのために，動作をしている人物を含む矩形領域を人手で付与した．矩形領域の際には，動作の起点の時点での人物領域を矩形で付与し，その後の人物位置は矩形領域内で生成した dense trajectories の重心を矩形領域で追跡することで各フレームでの人物領域の情報を付与した．付与された人物領域の例を図 20 に示す．

図 20: 付与された人物領域の例．赤色の矩形が付与された矩形領域である．まず $t = 0$ の時点での人物領域を人手で付与し，それ以降の人物領域は dense trajectories の重心を追跡することで判定している．

6 実験

6.1 実験概要

本研究では提案手法の評価のために第 5 章で作成したデータセットを用いて認識精度の評価を行った.

本研究で提案する手法は，カメラ装着者の視線情報をもとに動作映像全体に現れる局所特徴の中から動作認識において重要なものを選択し高次特徴を生成した上で認識を行う．そのため，カメラ装着者の自発的頭部運動により現れる映像背景の動きや，複数人物の映り込みによる認識対象でない人物の動きの影響を抑えて動作認識を行うことができると考えられる．提案手法のこのような性質を確かめるために以下の三つの手法を比較した.

- GAZE: 3.2 節で論じた手法により二人称視点映像の観測者の視線周辺の局所特徴を用いて高次特徴を生成する手法 (提案手法)
- ALL: 映像中に現れる全ての局所特徴から高次特徴を生成する手法 (ベースライン)
- BOX: データセットに長方形であらかじめ付与した人物領域から生成された局所特徴から高次特徴を生成する手法 (比較手法)

ALL は映像中に現れる全ての局所特徴から高次特徴を生成する従来手法に相当する．それに対し提案手法である GAZE では，視線周辺の局所特徴を選択した上で用いるため ALL よりも認識精度が高くなることが期待される．BOX は人物領域に由来する局所特徴を選択的に用いているため，理想的な局所特徴選択に近い手法である.

認識精度の評価では，6 人の被験者のうち 1 人のサンプルをテストデータ，残りの 5 人のサンプルを教師データとして交差検定を行った.

6.2 視線による局所特徴選択のための変数の決定

第 3 章で論じた視線による局所特徴のための変数 (r, q) を決定するために，(r, q) を様々な値に変えて精度比較を行った．ここで最も良い精度を与えた (r, q) を用いて提案手法の評価を行った.

認識精度の比較の際には，12 種類の動作識別の精度の平均値を評価した．また，精度比較の際には 6 人の被験者のうち 5 人によるサンプルを教師データとして Gaussian mixture model，サポートベクタマシンの学習を行い，残りの 1 人の被験者によるサ

ンプルをテストデータとして識別対象にすることで，(r, q) のパラメタの算出の際に教師データとテストデータが分離するようにした．このようなテストデータ，教師データの取り方で 6 人分の交差検定を実施し，その平均値を (r, q) のスコアとして比較を行った．

精度比較の結果を表 3 に示す．結果として，$r = 60\text{px}, q = 1\text{pixel}$ で最も精度の良い結果が得られたため，提案手法の評価に際してはこの値を採用した．

表 3: 各 r, q に対応する GAZE の認識精度

$r \backslash q$	1 frames	5 frames	10 frames	15 frames
30 px	31.3 %	28.2 %	24.7 %	22.4 %
60 px	37.3 %	36.2 %	32.3 %	29.4 %
90 px	34.5 %	34.9 %	34.3 %	32.5 %
120 px	31.0 %	31.1 %	30.0 %	26.9 %

6.3 実験結果

GAZE，ALL，BOX の三つの手法の間での精度比較を表 4 に示す．本実験では 12 種類の動作識別の他に，手による動作種類内での識別，頭部による動作内での識別を行い，その平均スコアを比較した．いずれの場合も GAZE が 従来手法による ALL を上回ることが確認された．これは，視線情報を用いて重要な局所特徴を推定しつつ動作の学習及び識別を行ったため，背景や別の人物の動きの影響を軽減できたためと考えられる．

一方，認識精度には動作種類により差が出た．各動作種類における識別結果の f-score を図 21 に示す．

識別結果の大まかな傾向として，頭部による動作の識別精度が手による動作の識別精度よりも低くなることが確認された．これは，頭部動作に含まれる動きが微細である傾向にあるため，動きの特徴を捉えきれず識別に失敗したものと思われる．

表 4: 各手法での認識精度比較

	All actions	Hand actions	Head actions
GAZE	37.3 %	51.2 %	32.4 %
ALL	23.6 %	38.1 %	28.0 %
BOX	41.6 %	57.1 %	35.5 %

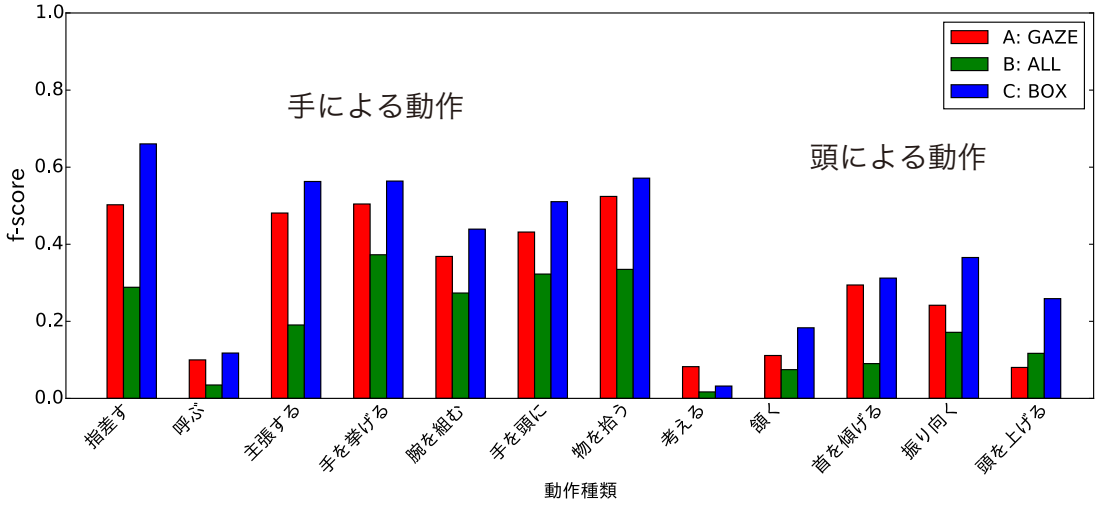

図 21: 各動作での認識結果の f-score

手による動作の中でも “呼びかけ” の識別精度が他と比べて低いが，これも同様の理由によるものと考えられる．

今回の実験で識別精度が高かった動作種類の例を図 22 に，識別精度の低かった動作種類の例を図 23 に示す．

6.4 考察

本実験では GAZE, ALL, BOX の三つの手法を比較し，提案手法の評価を行った．平均スコアでは提案手法による GAZE が従来手法の ALL を上回り，視線の利用により二人称動作認識の精度が向上することが確認された．しかしながら，動作種類ごとに認識精度に差があることが確認された．

図 21 に示した動作種類ごとの F スコアの比較では，“指差す” や “主張する” といった動作では GAZE による認識精度が ALL による認識精度を大きく上回ることが確認された．これらの動作ではカメラ装着者の視線が動作認識の対象者の手元を追いかけるように動くことが確認されている．このことから，手元の局所特徴を効果的に選択されたことにより GAZE による認識精度が ALL による認識精度よりも高くなったと考えられる．また，“指差す” ではカメラ装着者が指をさした方向に顔向けを行うことで映像に自発的カメラ運動が生まれ，背景動作特徴が盛んに生成されていることが確認された．この背景動作特徴が ALL による認識の際に結果に影響を与えたことも GAZE と ALL の認識精度に差が出た要因の一つであると考えられる．“指差す” の動作映像の例を図 24 に示す．

図 22: 識別精度の高かった動作種類の例．提案手法により選択された局所特徴に対応する dense trajectories と視線位置を可視化している．“手を挙げる”など，大きい動きの含まれる動作種類の識別精度が高くなる傾向にあることが確認された．

図 23: 識別精度の低かった動作種類の例．提案手法により選択された局所特徴に対応する dense trajectories と視線位置を可視化している．“頷く”など，大きい動きの含まれる動作種類の識別精度が低くなる傾向にあることが確認された．

図 24: 指差し動作の映像例．視線位置を黄丸，選択された局所特徴に対応する dense trajectories を緑線で可視化した．カメラ装着者の手の先を視線が追いかけており，その周囲の特徴が選択されていることを確認できる．

また，“指差す”の動作映像では GAZE と BOX の間でも認識精度に大きな差が確認され，BOX による認識精度が GAZE による認識精度を上回った．GAZE による“指差す”の動作認識では視線が手を伸ばした先を追いかけた際に ALL よりは少量ながらも，手先の周りの領域の背景の動きが局所特徴に含まれる．それに対し，BOX では選択された局所特徴のほとんどが人物領域に由来するものであるため，背景の動きの影響を受けずに GAZE よりも精度が高くなったと考えられる．

図 23 に示されるような“考える”，“頷く”，“呼ぶ”といった動作では GAZE, ALL, BOX のいずれも良い認識精度を得ることができなかった．これらは比較的小さな動きにより構成される動作であることが多く，局所特徴の抽出の段階で動きの特徴を捉えきれなかったことが原因と考えられる．

今回の実験ではほとんどの動作種類において，GAZE による認識精度は BOX による動作特徴に達しなかった．今回作成したデータセットでは，被験者が相手の顔の位置を注視する傾向にあることが確認されている．それに対して，動作時に特徴的な動きが現れるのは顔から手にかけての領域である．このことから，動作認識に重要な局所特徴は視線の下側の領域に分布することが推定される．このように，視線により選択される局所特徴の位置と実際に重要な動作が現れる位置に相違が生じていることが，GAZE による認識精度が BOX による認識精度よりも低くなった要因であると考えれる．このような事実を考慮して局所特徴量の重み付けを求めるためには，視線との距離だけではなく上下左右の位置関係も考慮できるように $f(V)$ を構成するという拡張が考えられる．

7　結論

7.1　結論

本論文では二人称視点動作認識における問題点を示し，その解決策として二人称視点映像の記録者の視線位置の情報を用いた動作認識手法を提案した．提案手法では映像中から生成された局所特徴に対して視線位置との位置関係に応じた重みを割り当てることにより，視線の周辺の局所特徴のみから高次特徴を生成した．

また，観測者の視線情報が付与された二人称視点映像のデータセットを作成することにより，視線を利用した動作認識手法の評価を可能にした．データセット作成の際には，動作認識対象と映像の記録者との間のやり取りの動作を収録するようにすることで，二人称視点動作に特有のコミュニケーションの一部としての振る舞いをする視線データを収録した．また，映像の記録者の視界に複数人の人物が映り込むように設定を行うことで従来の手法では動作認識が困難な場合を再現した．

最終的に作成したデータセットを用いて，視線を利用した動作認識手法と視線を利用しない動作認識手法を比較した．結果として，ほとんどすべての動作種類において視線利用により二人称視点映像中での動作認識の精度が改善することを示した．

7.2　提案手法の限界と課題

7.2.1　特定の動作種類における認識精度の低下

提案手法では，二人称視点映像の記録者の視線情報を用いて映像全体から抽出された局所特徴に対して重みを割り当てることで，動作認識の精度を向上した．しかしながら，その認識精度は動作種類ごとに差があり，特に頭による動作は手による動作と比べて認識精度が下がる傾向にあることが確認された．認識精度の低い動作種類には微細な動きにより構成される動作種類が多いことから，局所特徴の抽出の段階で微細な動きを捉えきれていない可能性が考えられる．

今後の改善としてはまず，フレームレートや解像度などの映像フォーマットを見直し最適な値を見つけることが挙げられる．また，頭による動作に関しては，動作認識の対象者もヘッドマウントカメラをつけている場合にその頭部運動がカメラ映像の背景運動として現れるため，これを利用して動作認識を行うという手法が考えられる．

7.2.2　視線計測機器による視線推定の失敗

提案手法では視線情報を利用して局所特徴の選択を行ったが，視線計測機器が視線の推定に失敗したフレームにおいては局所特徴の選択を行うことができないという問題がある．視線計測機器は必ずしも全てのフレームで視線位置の推定に成功するとは限らないため，頑健な動作認識のためには視線位置の計測に失敗した場合の対応をする必要がある．

今後の課題として，視線の推定に失敗したフレームにおいては直近のフレームの視線情報をもとに視線位置の補完を行うといった改善が必要である．

7.2.3　正面以外からの動作映像の考慮

今回収録したデータセットでは正面からの動作映像の身を取り扱っているため，側面や斜め方向からの動作認識が困難になる可能性がある．実際の二人称視点映像では必ずしも正面からの人物映像が得られるとは限らないため，正面以外からの動作映像を含めたデータセットでも動作解析が可能であるか今後検証する必要がある．

7.3　今後の展望

最後に，今後の展望について述べる．

今回は動作の起きるタイミングは既知のものとしてその種類の識別のみを行っていた．しかし，今回の視線を用いた動作認識の手法を動作をしている人物映像と何もしていない状態の人物映像の識別タスクに応用することで，動作検出も可能である．そこで，次のステップとして視線を利用した二人称視点動作検出が考えられる．

動作検出と動作認識の両方が可能となった段階で，二人称視点映像に現れる動作に基づいた映像要約や，人物相関の推定といったより実用的な応用を目指す．二人称視点に現れる動作に基づいた映像要約では，会議や協調作業などを記録した二人称視点映像の一連のやりとりの中で, “手を挙げる”や “指差しをする”といった重要なキーアクションを検出する．これらの動作が起きているタイミングは映像中重要な箇所であると推定することができる． このようにして推定された映像の重要な箇所をつなぎ合わせることで二人称視点映像の要約が可能であると考えられる．

人物相関の推定では，二人称視点映像中に現れる人物と映像の記録者の間のやりとりを検出することで，映像の記録者がどの人物と頻繁にやりとりをするかを集計することができる．この情報を用いれば，例えば映像の記録者と頻繁にやりとりをする人物は映像の記録者と親しい間柄にある，といったような推定をすることができると考えられる．また，他者どうしのやりとりにおいて “呼びかけ”と “振り向く”

などの動作とそれに対する反応のペアを検出することができれば，他者同士の人物関係も推定することが可能となる．

二人称視点動作認識に関するこのような実用的な応用を通して システムによるヒューマンインタラクションのより深い理解を目指す．

参考文献

[1] Stefano Alletto, Giuseppe Serra, Simone Calderara, Francesco Solera, and Rita Cucchiara. From ego to nos-vision: Detecting social relationships in first-person views. In *Proc. of 3rd Workshop on Egocentric (First-person) Vision*, Columbus, Ohio, June 2014.

[2] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cdric Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–22, 2004.

[3] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *International Conference on Computer Vision & Pattern Recognition*, Vol. 2, pp. 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005.

[4] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human Detection Using Oriented Histograms of Flow and Appearance. In Ales Leonardis, Horst Bischof, and Axel Pinz, editors, *European Conference on Computer Vision (ECCV '06)*, Vol. 3952 of *Lecture Notes in Computer Science (LNCS)*, pp. 428–441, Graz, Austria, May 2006. Springer-Verlag.

[5] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, SCIA'03, pp. 363–370, Berlin, Heidelberg, 2003. Springer-Verlag.

[6] Alireza Fathi, Jessica K. Hodgins, and James M. Rehg. Social interactions: A first-person perspective. In *CVPR*, pp. 1226–1233. IEEE, 2012.

[7] Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part I*, ECCV'12, pp. 314–327, Berlin, Heidelberg, 2012. Springer-Verlag.

[8] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *ARTIFICAL INTELLIGENCE*, Vol. 17, pp. 185–203, 1981.

[9] N. Jammalamadaka, A. Zisserman, and C. V. Jawahar. Human pose search using deep poselets. In *International Conference on Automatic Face and Gesture Recognition*, 2015.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

[11] Ivan Laptev. On space-time interest points. *Int. J. Comput. Vision*, Vol. 64, No. 2-3, pp. 107–123, September 2005.

[12] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*, jun 2008.

[13] Bingbing Ni, Pierre Moulin, Xiaokang Yang, and Shuicheng Yan. Motion part regularization: Improving action recognition via trajectory selection. June 2015.

[14] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pp. 143–156, Berlin, Heidelberg, 2010. Springer-Verlag.

[15] Michael S. Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, pp. 2730–2737. IEEE, 2013.

[16] Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. 2013.

[17] Nataliya Shapovalova, Michalis Raptis, Leonid Sigal, and Greg Mori. action is in the eye of the beholder: eye-gaze driven model for spatio-temporal action localization. In c.j.c. burges, l. bottou, m. welling, z. ghahramani, and k.q. weinberger, editors, *advances in neural information processing systems 26*, pp. 2409–2417. curran associates, inc., 2013.

[18] Jianbo Shi and Carlo Tomasi. Good features to track. In *1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, pp. 593 – 600, 1994.

[19] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. Vol. abs/1406.2199, 2014.

[20] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, Vol. 2, pp. 1470–1477, October 2003.

[21] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, Vol. abs/1212.0402, , 2012.

[22] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pp. 3169–3176, Colorado Springs, United States, June 2011.

[23] Heng Wang and Cordelia Schmid. Action Recognition with Improved Trajectories. In *ICCV 2013 - IEEE International Conference on Computer Vision*, pp. 3551–3558, Sydney, Australia, December 2013. IEEE.

[24] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. *CoRR*, Vol. abs/1505.04868, , 2015.

[25] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg. Studying relationships between human gaze, description, and computer vision. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Computer Society Conference on*. IEEE, 2013.

発表文献

村上晋太郎，米谷竜，佐藤洋一．“視線を利用した二人称視点動作認識”，コンピュータビジョンとイメージメディア研究会 (CVIM200), 2016.

表彰

第 200 回コンピュータビジョンとイメージメディア研究会 CVIM 研究会奨励賞受賞

謝辞

本論文は，東京大学大学院 情報理工学系研究科 電子情報学専攻コースの修士研究として行われた 2 年間の研究の成果をまとめたものです．研究の進行及び論文の執筆に際しては，多くの方々にご助力いただきました．ここに感謝の意を述べさせていただきます．

本研究は，東京大学生産技術研究所の佐藤洋一教授のもとで行われました．佐藤先生には，指導教官として研究の道筋を示していただいた他，研究室内での活発な知識交換の場を設けていただきました．この研究をやり遂げることができたのも，先生のご指導があってこそでした．

東京大学生産技術研究所 助教の米谷竜さんには，研究の成立に不可欠なコンピュータビジョンの要素技術や機械学習全般の知識について指導をしていただきました．私は学部時代から専門分野を変えたために修士課程からこれらの技術に触れてきましたが，新しい分野で研究をできたのは米谷さんのご指導のおかげでした．また，米谷さんには研究の要素技術の知識のみならず “伝える” というのはどういうことなのか，相手に応じてどのようなことに留意するべきかなど，これからの人生全般にためになるような知識を伝授していただきました．

研究室のメンバーには，日々の生活の中で心の支えになっていただきました．特に，データセット収集の際には貴重な時間を割いて実験に参加していただき，大変感謝しております．

ここに書ききれない多くの友人たちの協力, 多くの先生方のご指導によって本論文を完成させることがきました. この場を借りて, 深く御礼申し上げます.

平成 28 年 2 月 4 日

村上 晋太郎