Vol.2016-CVIM-200 No.32
2016/1/22

視線を利用した二人称視点動作認識

村上 晋太郎[1,a)]　米谷 竜[1,b)]　佐藤 洋一[1,c)]

概要：ウェアラブルカメラを用いた撮影では映像の記録者の視界に別の人物が映り込むことにより，映り込んだ人物にとっての二人称視点映像が生じる．二人称視点映像では，定点カメラなどによる三人称視点映像と比べて，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．本研究ではこの二人称視点映像中での人物の動作認識について検討する．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．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装置を用い，映像の記録者が注視している位置との関係から画面中のそれぞれの部位の重要度を考慮しつつ特徴量を生成することで，二人称視点映像で頑健な動作認識を目指す．

Second person action recognition using gaze

Shintaro Murakami[1,a)]　Ryo Yonetani[1,b)]　Yoichi Sato[1,c)]

1. 序論

近年，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，一人称視点からの映像が盛んに利用されている．ウェアラブルカメラからの撮影では，映像の観測者の視界に別の人物が映り込むことにより，映り込んだ人物の二人称視点からの映像が記録される．こうした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．

本研究では，特に複数人物が会話をしている状況をとりあげ，人物の二人称視点映像から”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．二人称視点映像中で動作を認識することで，映像の記録者と他者のやり取り内容を認識し，共同作業の記録等の応用が可能となる．

文献 [17], [19] に見られるように従来手法の多くではまず histograms of oriented optical flow(HOOF)[9] や histogram of oriented gradients(HoG)[3] といった局所特徴を画面全体から抽出し，bag of features[2], [16] や Fisher vector[13] といったコーディング手法によって高次特徴を生成するというアプローチをとる．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全体の動きが高次特徴に影響を及ぼす可能性がある．また，同映像中に複数人物が現れる場合，認識対象となる人物の動きのみを考慮した高次特徴を生成する必要がある．

このような課題を解決するための手段として，映像中に現れた局所特徴の中で動作認識に有用であるものを見つけ出した上で，背景運動や第三者の動きによる不要な局所特徴の影響を抑制するという方法が考えられる．そこで本研究では，二人称視点映像を記録する人物（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計測し，動作認識に利用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，重要な人間の振る舞いには視線が向けられることが期待される．そのため，視線の位置から画面中のどの位置に現れる動きが重要であるか推定することができる．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

[1] 東京大学　生産技術研究所, 〒 153-8505 東京都目黒区駒場 4-6-1
[a)] shin-m@ut-vision.org
[b)] yonetani@iis.u-tokyo.ac.jp
[c)] ysato@iis.u-tokyo.ac.jp

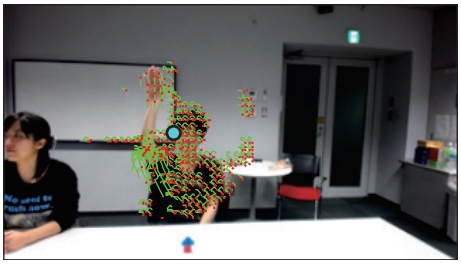


図 1: 視線情報を元に選択された局所特徴の例 (水色の丸は視線位置)

を用いて動作を認識する．

本研究の貢献は以下の通りである．

- 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．
- 二人称視点映像における，相手とのやりとりの中で生じ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．
- 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，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示した．

2. 関連研究

2.1 二人称視点映像の利用

二人称視点映像の利用に関するする研究として，文献 [1], [5] が挙げられる．これらの研究では一人称視点映像に現れる人物の頭部の位置と向きを推定し，それを元に登場する人物の位置関係と人物同士の顔向けの発生を観測することで，誰と誰がどのくらいの頻度でやり取りをしているかという関係を推測した．一方，やり取りの内容を含めたより詳細な分析のためには，どのような種類の動作が起きているのかという動作種類の認識も可能であることが望ましい．

二人称視点映像から動作種類を識別した研究に [12] 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」「こちらを指差す」といった映像記録者と他者のやり取りを想定した動作を識別した．[12] で使用されたヘッドマウントカメラは固定されたものであったが，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，二人称視点からの動作認識のためには頭部運動に由来するカメラ運動にうまく対処できるような動作特徴表現を用いることが望ましい．

2.2 映像の特徴表現

映像中での人物動作の特徴表現として Improved Dense Trajectories(IDT)[17], [18] が挙げられる．IDT は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)[9], motion boundary histogram(MBH)[4] などの見えと動きに由来する局所特徴量を集めるものである．IDT は身体部位の認識などの複雑な工程を経ずに特徴を抽出するためカメラ運動の影響に頑健である．

一方，静止画像における一般物体認識では局所特徴量ベースの手法に変わる手法として convolutional neural network(CNN)[8] の利用が注目されている．人手により局所表現を定義する必要のあった従来の手法とは異なり，CNN では特徴表現の方法を学習することができる．このようにして学習された特徴表現を deep feature と呼ぶ．two-stream CNN(TCNN)[15] では，二つの CNN で独立して画像と optical flow[7] を処理することで，動きの情報と見えの情報の両方から deep feature を生成し動作を認識することを可能にした．

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)[19] 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 deep feature 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．[17], [18] では映像全体の局所特徴から bag of features[2], [16] や Fisher vector[13] 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．

[10] 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した．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 [20] 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，画像中のどの箇所が注視され

るかという情報を集めた．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 [6] や Shapovalova らの手法 [14] 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装置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．ここで言う一人称視点映像とは，映像記録者自身の動作が記録された映像のことを指す．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収集することで動作特徴を生成した．

一方，Shapovalova ら [14] 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，被験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

3.1 概要

図 2 は，提案手法の概要を示したものである．提案手法では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．それぞれの映像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．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映像記録者の視線位置情報が与えられている．

提案手法ではまず解析対象の映像全体から局所特徴を抽出し，Fisher vector(FV)[13] を用いて高次特徴量を生成する．高次特徴を生成する際に，それぞれの局所特徴がどれだけ重要であるかを視線を用いて推定する．

最終的に生成された映像特徴に線形識別器を利用した他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．

3.2 視線を考慮した高次特徴の生成

3.2.1 視線情報を用いた局所特徴の重み付け

映像全体から抽出された局所特徴の集合を以下のように定義する．

$$X = \{\boldsymbol{x}_n | n = 1, 2, ..., N\} \tag{1}$$

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

$$\boldsymbol{v}_n(t) = \boldsymbol{l}_n(t) - \boldsymbol{l}_g(t) \tag{2}$$

について，$v_n(t)$ の軌跡を以下のように表す．

$$V_n = (\boldsymbol{v}_n(1), \ldots, v_n(T)) \tag{3}$$

この V に対して，以下のような関数 f を定義する．

$$f(V) = \begin{cases} 1 & (g(V, r) \geq q) \\ 0 & (g(V, r) < q) \end{cases} \tag{4}$$

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．

$$w_n = f(V_n) \quad (w_n \in \mathbb{R}, w_n \geq 0) \tag{5}$$

w_n は局所特徴 $\boldsymbol{x}_n$ の中で対応する軌跡が視線に近いものを選択する働きを持つ．r は，視線にどの程度近い局所特徴を選択するか決めるパラメータである．$r = \infty$ 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．視線から r 以内の領域を視線周辺領域と定義すると，q は軌跡が何フレームの間視線領域内に存在した場合にその局所特徴を選択するかを決定する．r を適切に設定することにより，視線位置が激しく動くような場合に選択される局所特徴が多くなりすぎることを防ぐことができる．

なお，視線計測装置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 3 に示す．

以上で定義された各局所特徴の重み

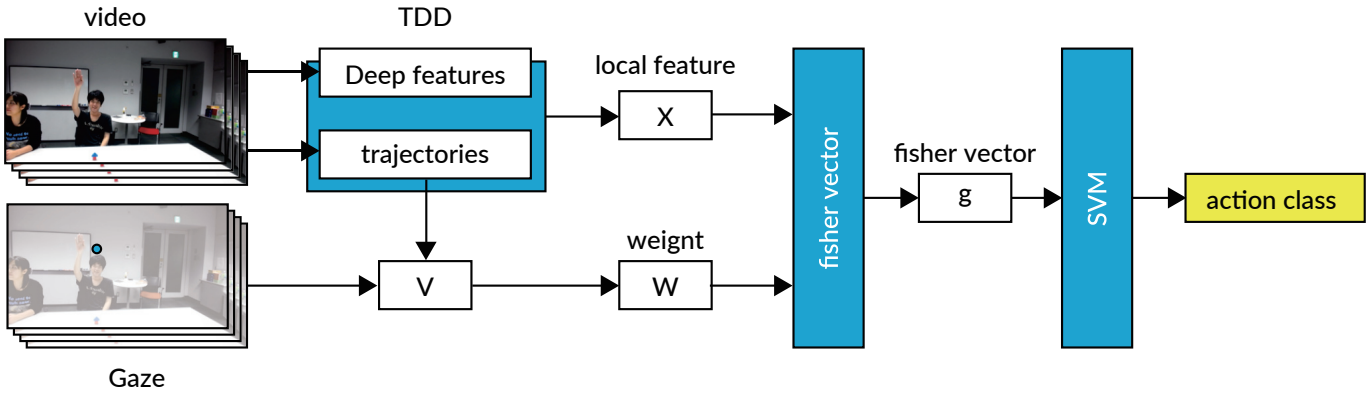
$$W = \{w_n | n = 1, ..., N\} \tag{6}$$

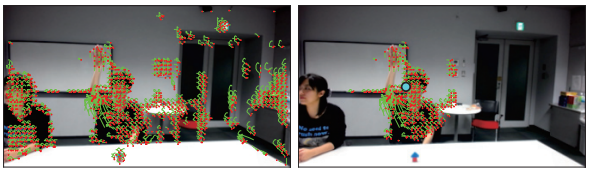
を用いて，高次特徴を生成する．

3.2.2 視線情報を用いた Gaussian mixture model の学習

本研究では，視線を用いて局所特徴量から高次特徴量

図 2: 提案手法の概要





(a) 全ての局所特徴　**(b)** 視線により選択された局所特徴

図 3: 視線による局所特徴の選択

を生成するために Fisher vector を拡張する．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．

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_k \pi_k \mathcal{N}(X_n|\boldsymbol{\mu}_k, \Sigma_k) \tag{7}$$

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

$$\boldsymbol{\theta} = (\pi_1, ..., \pi_K, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K, \Sigma_1, ..., \Sigma_K) \tag{8}$$

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

$$X = \{x_n|n = 1, ..., N\} \tag{9}$$

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

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \max_Y \prod_n \prod_k (\pi_k \mathcal{N}(X_n|\boldsymbol{\mu}_k, \Sigma_k))^{Y_{nk}} \tag{10}$$

$Y \in \{0, 1\}^{n\times k}$ 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．

この最大化問題に対して，局所特徴の重み

$$W = \{w_n|n = 1, ..., N\} \tag{11}$$

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．

$$\boldsymbol{\theta}_w = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \max_Y \prod_n \prod_k (\pi_k \mathcal{N}(X_n|\boldsymbol{\mu}_k, \Sigma_k))^{Y_{nk} w_n} \tag{12}$$

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．

3.2.3 視線情報を考慮した 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．ある動画サンプルから得られた局所特徴群 X に対して，その対数尤度 $\boldsymbol{s}(X|\boldsymbol{\theta})$ を以下のように定義する．

$$\boldsymbol{s}(X|\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log p(X|\boldsymbol{\theta}) \tag{13}$$

サンプル X から生成される Fisher vector $\mathcal{G}^X_{\boldsymbol{\theta}}$ は $\boldsymbol{s}(X|\boldsymbol{\theta})$ に対して一意に定まり，以下のように求められる．

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

提案手法ではこの対数尤度勾配 $\boldsymbol{s}(X|\boldsymbol{\theta})$ に以下のように局所特徴の重み W を導入することで，重みを考慮した対数尤度勾配 $s(X, W|\boldsymbol{\theta})$ に拡張する．

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

このように定義することで，$w_n = 0$ の場合は局所特徴 $\boldsymbol{x}_n$ を無視した場合の対数尤度勾配，$w_n \in \mathbb{N}$ の場合は局所特徴 $\boldsymbol{x}_n$ の個数を w_n 倍に増した場合の対数尤度勾配に対応させることができる．

上記で定義された重みつき対数尤度勾配を用いて，視線情報を加味した重み付け Fisher vector $\mathcal{G}_{\boldsymbol{\theta}}^{X,W}$ は以下のように求められる．

$$\mathcal{G}_{\boldsymbol{\theta}}^{X,W} = G(\boldsymbol{s}(X, W|\boldsymbol{\theta})) \tag{15}$$

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．

3.3 実装

本節では，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．

3.3.1 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 2.2 節で説明した TDD[19] を採用した．TDD では映像中に現れる特徴点を追跡し，その軌跡上の deep feature を局所特徴として用いる．Deep feature の生成には Wang ら [19] により開発された TCNN モデル [*1] を使用した．独立した二つの CNN 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 deep feature として用いた．文献 [19] を参考に，画像の deep feature には **conv4** 層の出力を，optical flow の deep feature には **conv3** 層の出力を採用した．

IDT の生成の際には，30fps の映像の 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．映像のサイズは 320px × 180px とし，特徴点の選択の際には隣り合う特徴店同士は 8px 以上の間隔が開くようにした．また，IDT の軌跡の長さは 15 フレームに統一した．

3.3.2 視線情報の利用

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変数 r と，局所特徴に対応する軌跡が何フレームの間視線周辺領域内に存在した場合に局所特徴を選択するかという変数 q の二つのパラメータが存在する．そこでこの二つのパラメー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する．

3.3.3 Fisher vector による識別

3.3.1 項の実装により生成される局所特徴は，各フレームの画像から 512 次元，optical flow から 512 次元の合わせて 1024 次元である．また，生成された Fisher vector は，Florent らの手法 [11] を用いて正規化した上で使用した．

識別には support vector machine を用いた．また，多クラス分類には one vs one classifier を用いた．

[*1] https://github.com/wanglimin/TDD

表 1: 本研究で収録した動作映像のクラス

part of body	action class
hand	pointing
	calling
	insisting
	raising hand
	arm wrestling
	scratching head
	picking up
hand and head	thinking
head	nodding
	cocking neck
	tuning head
	raising face

図 4: 各動作の映像例

4. データセット

本研究では，提案手法の有効性を評価するため視線データを含む二人称視点映像のデータセットを作成した．本節では，作成したデータセットの詳細について説明する．

4.1 動作クラス

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．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．本データセットで収録した動作クラスのリストを表 1 に示す．また，各動作の映像例を図 4 に示す．

4.2 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[*2] を使用した．二人称視点映像は 30fps のフレームレート，320px × 180px のサイズで収録した．

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

[*2] http://pupil-labs.com

は，例えば”raising hand”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は 1.5 秒に統一した．

それぞれの動作クラスに対して 1 人あたり 18 サンプル，合計で 108 サンプルを収録した．全ての動作クラスで計 1296 サンプルの動作映像を収録した．

それぞれのサンプルに対して，収録されている動作のクラスと動作をしている人物のラベルを付与した．また，動作をしている人物の位置が分かっている状態での動作認識と視線情報を利用した提案手法を比較するために，動作をしている人物の領域を長方形で付与した．

なお，視線情報により動作認識対象以外の人物の映り込みの影響を抑制できるかどうかを評価するため，映像サンプルでは映像記録者を含む 3 人の被験者が向かい合って互いに動作をし，記録者の視界に他の 2 人が入るようにした．

5. 実験

5.1 実験概要

本研究では提案手法の評価のために以下の三つ手法を比較した．

- 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
- 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)
- 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

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，従来手法に対応する．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．

まず，GAZE のパラメータ r と q を決定するために，これらのパラメータを様々に変えて提案手法を評価した．その中で認識精度が最も高くなるような変数の組み合わせを GAZE のパラメータとして採用した．

その上で，ALL, GAZE, BOX の識別精度を比較評価した．腕による動作クラスと頭による動作クラスの間で識別精度が大きく異なるため，それぞれのクラス内での認識精度も評価した．

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．

表 2: 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	28.0 %	25.7 %	21.6 %	21.6 %
60 px	35.7 %	36.7 %	27.9 %	27.9 %
90 px	32.6 %	32.6 %	30.3 %	30.3 %
120 px	29.1 %	27.8 %	26.8 %	27.8 %

表 3: 各手法での認識精度比較

	All actions	Hand actions	Head actions
ALL	23.6 %	38.1 %	28.0 %
GAZE	36.7 %	51.2 %	32.4 %
BOX	41.6 %	57.1 %	35.5 %

5.2 実験結果

GAZE で r と q を様々に変えた際の認識精度の変化を表 2 に示す．$r = 60, q = 5$ の値で最も認識精度が高くなったため，この値を GAZE のパラメータとして採用した．

ALL, BOX, GAZE の三つの手法の間での精度比較を表 3 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った．提案手法の 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．

また，全ての手法を通して手による動作の識別精度が頭部による動作の識別精度を大きく上回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．手による動作の中でも”calling”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．

各動作クラスにおける識別結果の f-score を図 5 に示す．多くの動作種類で GAZE の f-score が ALL の識別精度を上回った．

5.3 考察

本実験では ALL, BOX, GAZE の三つの手法を比較し，ベースラインの ALL の識別精度を提案手法の GAZE の識別制度が上回ることを確認した．このことにより，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うことで，動作識別の精度が向上することが確認された．しかしながら，提案手法の識別精度は多くの動作クラスで BOX の識別精度には及ばなかった．このことから，提案手法の局所特徴量選択方式には改良の余地があると言える．

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

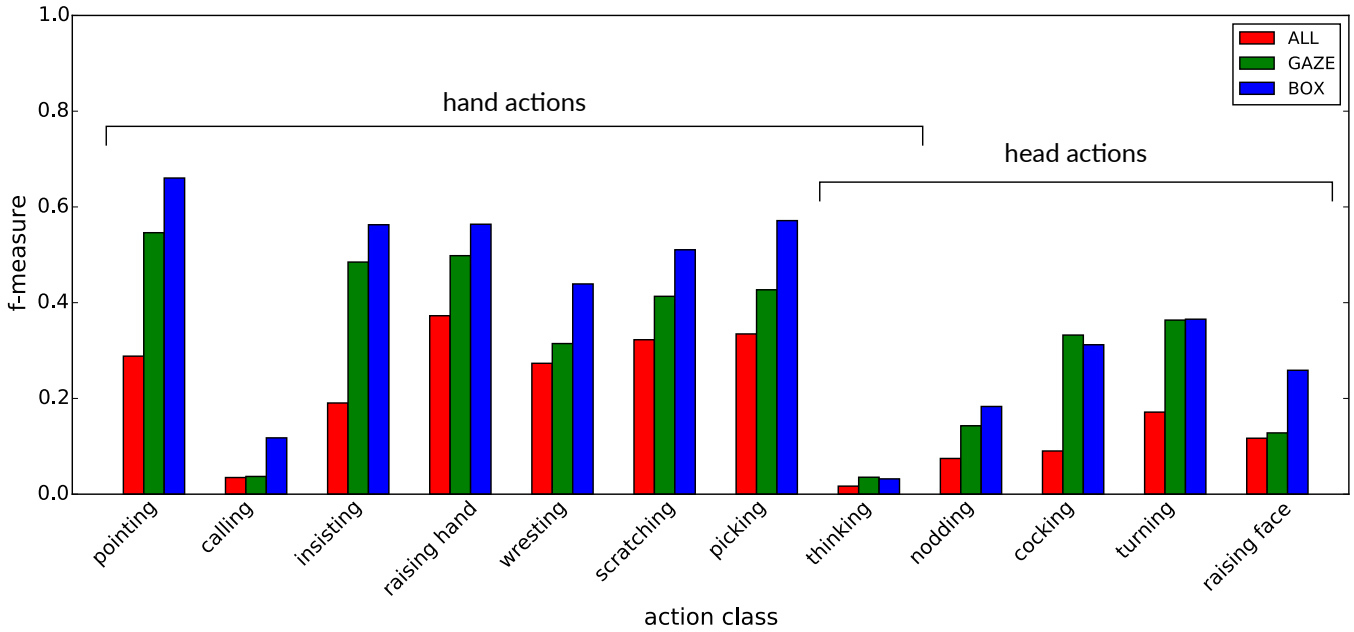


図 5: 各動作での認識結果の f-score

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 $f(V)$ を構成する必要がある．この点は今後の課題とする．

6. 結論

本研究では視線情報を用いて動作認識に重要な局所特徴を推定する手法を提案した．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．それらを用いて，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することにより，動作認識における視線利用の有用性を示した．

謝辞

本研究の一部は JST CREST および栢森情報科学振興財団の助成により行った．

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

[6] 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.

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

[8] 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.

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

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

[11] 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.

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

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

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.