エゴモーションを利用した自己動作カテゴリの教師無し学習

木谷 クリス 真実† 岡部 孝弘†† 佐藤 洋一†† 杉本 晃宏†††

† 電気通信大学 大学院情報システム学研究科 182–8585 東京都調布市調布ヶ丘 1–5–1
†† 東京大学 生産技術研究所 153–8505 東京都目黒区駒場 4–6–1
††† 国立情報学研究所 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †kitani@is.uec.ac.jp, ††{takahiro,ysato}@iis.u-tokyo.ac.jp, †††sugimoto@nii.ac.jp

あらまし 本論文では，一人称視点動画による自己動作，つまり映像を撮影している本人の動作を対象としたカテゴリ学習手法を提案する．従来研究では，カメラの動き，すなわちエゴモーションを外乱として扱い，カメラの「ブレ」を補正した後，人物動作のカテゴリ分類及び認識に取り組んできた．これに対して本研究では，このエゴモーションが，動作学習のために有効な手がかりとなることを示す．本提案手法では、ディリシュレ過程を用いた混合モデルを利用し，（1）特徴の高速クラスタリング，および（2）自己動作カテゴリの教師無し学習を実現する．また実験では，動的な野外動作の映像を用いて，エゴモーションが一人称視点による動作カテゴリの発見に有効であることを示す．
キーワード エゴモーション，一人称視点ビジョン，ディリシュレ過程，教師無し学習

Using Ego-motion to Learn First-Person Action Categories

Kris M. KITANI†, Takahiro OKABE††, Yoichi SATO††, and Akihiro SUGIMOTO†††

† University of Electro-Communications, Graduate School of Information Systems,
1–5–1 Chofugaoka, Chofu, Tokyo 182–8585 JAPAN
†† University of Tokyo, Institute of Industrial Science, 4–6–1 Komaba, Meguro, Tokyo 153–8505 JAPAN
††† National Institute of Informatics, 2–1–2 Hitotsubashi, Chiyoda, Tokyo 101–8430 JAPAN
E-mail: †kitani@is.uec.ac.jp, ††{takahiro,ysato}@iis.u-tokyo.ac.jp, †††sugimoto@nii.ac.jp

Abstract We propose a new framework for learning first-person action categories without supervision from a first-person point-of-view video sequence. In the past, first-person motion, namely ego-motion, has been treated as unwanted noise and is usually removed to improve action recognition performance. To the contrary, we show that ego-motion, far from being unwanted, is actually a vital source of information for understanding first-person actions. We implement a Dirichlet process multinomial mixture model for the two tasks of (1) learning low-level feature clusters online and (2) learning first-person action categories. In our experiments, we use a video of dynamic outdoor actions to show that the use of ego-motion is both effective and necessary for discovering first-person action categories.
Key words Ego-motion, First-person vision, Dirichlet process, Unsupervised Learning

1. はじめに

本研究は一人称視点の動き，すなわちエゴモーションを利用した行動解析に着目する．具体的には，エゴモーションそのものを観測特徴として活用し，教師無し学習の枠組みで自己動作，つまり映像を撮影している本人の動作カテゴリを発見する．一人称視点の自己動作解析技術は，最近，コンピュータビジョンの分野で注目されていて，幅広いアプリケーションへの応用が期待されている．視覚障害者用支援技術(注1)や患者のモニタリング [1] やモバイルアシスタント [2] への応用は既に成果が得られているものの一例である．

従来の動作認識に関する研究は三人称視点を仮定するものが多く，初期の研究ではカメラの運動を考慮せず，固定カメラを前提としたアプローチが提案されてきた．最近ではカメラのブレを考慮するアプローチも提案されている [3]〜[5]．これらの手法はカメラ運動の補正や特徴点の選定方法を工夫し，エゴモーションの影響を排除している．すなわち，カメラのエゴモーションは基本的に外乱として扱われている．

(注1)：GROZI:A Grocery Shopping Assistant for the Visually Impaired. grozi.calit2.net

図 2　障害物コース映像．三人称視点画像（上行），一人称視点画像（下行）

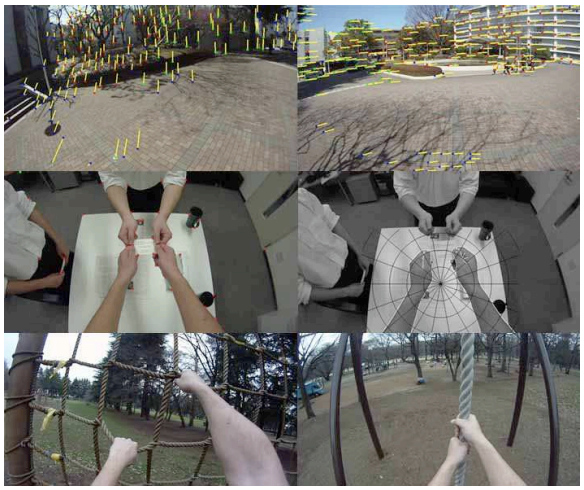

図 1　一人称視点映像データセット: 広場映像（上），訪問者映像（中），障害物コース映像（下）．

しかしながら，エゴモーションは必ずしも外乱ではない．エゴモーションには，映像を撮っている本人とその環境に関する情報が含まれている．それゆえ，エゴモーションは自己動作を理解するために重要な手がかりである．エゴモーションから，人が立ち止まって撮影しているのか，あるいは対象を追いかけているかといった自己動作が推定できる．さらに，ウエアラブルカメラシステムの場合は，身体のグローバルな動きとともに，頭や肩に装着したカメラから人の頭部方向の情報も得ることができる [6].

本研究の目的は，図 1 のような一人称視点映像から人物の自己動作カテゴリ（例．走る，渡す，登る）を認識することである．本提案手法では人物のエゴモーションを利用し，教師無し学習の枠組みで自己動作カテゴリを発見する．具体的には，ディリシュレ過程を事前分布を持つ混合モデル（トピックモデル）を用いて，低次の特徴点クラスタリングおよび自己動作カテゴリの学習に利用する．実験では，エゴモーション特徴と手の輪郭特徴を併用することにより，屋内映像だけでなくエゴモーションの変動の大きい野外映像（図 2）からも自己動作カテゴリを学習し，本提案手法の有効性と頑健性を示す．

2. 関 連 研 究

2.1 一人称視点の映像解析

一人称視点映像の最も扱い難い点はカメラの自由な運動である。そのため，多くの従来研究ではエゴモーションの影響を受けにくい特徴を利用し，エゴモーションの直接的な扱いを避けてきた．

画像処理にもとづく初期のウエアラブル（一人称視点）システムは，使用範囲を制限し，エゴモーションのないことを前提条件としていた．例えば，一人称視点における手話認識 [7] やジェスチャによる入力インターフェース [8], [9] は，認識の際には身体が動いていないことを前提条件としている．これらのような使用目的が限定されている場合，エゴモーションを考慮する必要はない．しかし，より一般的な動作認識においては，エゴモーションは非常に重要な手がかりとなり得る．これに対し，一人称視点映像による動作認識・行動解析の研究では，エゴモーションを利用した例はほとんど存在しない．一般的には手の動作に着目した研究が多く，手の輪郭による動作認識 [10]，モデルによる手の動作の認識 [11]，または手の Motion History Image (MHI) を利用した動作認識 [12] が主である．エゴモーションを利用した研究としては，カメラと加速度センサを利用した手法 [13] がある．ただし，[13] のカメラはシーン全体のテキスチャを表すために使用されていて，人物の動作を記述するためには加速度センサが利用されていた．

ビジョンを利用した動作認識研究ではエゴモーションの扱いが少ない一方，ウエアラブルシステムの研究分野では人物の動きを動作認識に利用することは一般的である．最新の研究では，モーションセンサを利用した手法が有効であることが示されている [14]．体に装着した二つの加速度センサ情報から Latent Dirichlet Allocation (LDA) を用いて高次の人物行動カテゴリを教師無しで学

習している．本研究は，動きの情報を利用している点，また自己動作カテゴリを学習している点においては[14]に似ているが，注目されたい相違点は，本提案手法ではカメラのみを利用していることである．また, [14] では動作ではなく，長時間における行動カテゴリ学習を対象とし，30 分間隔の加速度データからカテゴリ発見していることに対して，本提案手法では短い動作を対象とし，2 秒区切りでカテゴリの発見を行っている．

モーションセンサはビジョンの運動推定と比べ精度は優れているが，ビジョンベースのシステムにもいくつかの利点がある．加速度センサでは混在した運動（例えば車や電車の中での人の動き）は扱い難い一方，ビジョンベースの運動推定は比較的頑健である．また，人の動作を理解するという課題においては，一つのカメラで複数のコンテキストを解析することができる．例えば，一つのカメラで，手と環境と物体のコンテキスト認識とカメラ運動による動作認識を同時に解析ができるという利点がある．

2.2 トピックモデルと三人称視点映像解析

長期間録画された映像から人物の動作に人手でカテゴリのラベル付けをすることは多大な労力を要するため，教師無しでカテゴリを発見する手法が注目されている．実際，言語処理と一般物体認識分野に次いで動作認識研究においても教師無しで学習できるトピックモデル（混合モデル）が注目され，多くの研究に使われている．混合モデルを用いて動きの局所特徴量を利用した研究 [15]，構造的な空間情報を用いた手法[16], [17]，さらに物体と動作の関係を考慮したアプローチ [18], [19] が提案されている．しかし，これらの手法は有限次元の混合モデルを利用しているため，教師無しとはいえ，混合数を予め決定しておく必要がある．

この混合数をも推定する枠組みとして，無限次元分布を持つディリシュレ過程（Dirichlet Process - DP）を利用したノンパラメトリック混合モデルがある．本提案手法ではこのディリシュレ過程を利用し学習を行う．カテゴリ数を推定する点，特徴量クラスタリングを自動で行う点，また入力映像を予め分割する必要がない点においては，提案手法は [20] の枠組みに類似している．しかし, [20] で使用したサンプリング推定ではなく，より高速なオンライン推定と変分ベイズ推定の組み合わせを提案している．そして，最大の相違点は，本研究はエゴモーションを扱っている点である．

3. 特徴の記述

カメラの環境に対する相対的な運動を計算するために，まず安定した空間的特徴点 [21] を抽出し，画像フレーム間の特徴点のオプティカルフローを求める [22]．次に RANSAC を用いて，特徴点集合の移動を平面ホモグラフィで近似し，射影誤差が小さい特徴点を残すことにより，環境の局所的な動き（人，車，自転車など）や視差の大きい物（自分の手や電信柱などの近傍の物体）の影響を避けることができる．

各々のフレーム（正確にはフレーム間）のエゴモーションヒストグラムを以下のような表現で蓄積する．エゴモーションヒストグラムは図 3 のように合計 8 個のビンで形成されている．まず一つのフレームの特徴点移動の平均強度 μ と分散 σ を計算，平均 μ が一定の閾値（実験では $\mu_m = 2$）より小さい場合は考慮しないこととする．分散が一定の閾値より小さい場合（実験では $\sigma_m = 9$）は下段の 4 ビンに各々の特徴点の移動方向の度数を蓄積し，分散が大きい場合は上段の 4 ビンに同様に蓄積する．この記述方法を利用することによりエゴモーションの方向と移動の強度情報が特徴として残されることとなる．本手法ではシンプルなエゴモーションの記述を利用しているが，より正確な姿勢推定 [23] を用いてエゴモーションを記述することも可能である．

この 8 次元エゴモーションヒストグラムはトピックモデルの逐次的推定アルゴリズムによりオンラインでクラスタリングされ，一つの映像セグメント（n フレーム区間）についても一つの bag-of-features（BOF）がオンラインで蓄積される．実際激しい運動（例えば，走る）を行う際の特徴追跡は困難であり，ホモグラフィの計算に誤りが生じることがある．しかし，BOF といった統計量を利用しているため，ホモグラフィ計算の誤りには頑健である．一般的に使用されている局所運動を表す時空間ボリューム特徴記述 [24] [25] はフレーム間の厳密なカメラ移動の補正が必要なため，本研究では時空間ボリューム特徴記述を利用せず，フレーム単位のグローバル記述を利用している．

エゴモーション特徴はグローバルな運動を記述する一方，手のローカルな動きは検出できないため，先行研究に倣い本提案手法では手の動き特徴も利用する．手の輪郭の記述は Histogram of Oriented Gradients (HOG) [26] を利用している．まず HSV 色空間内で閾値処理を行い手の領域を求める．領域内の勾配の強い輪郭線を抽出し，輪郭線上の全ピクセルから勾配ヒストグラムを蓄積する．手の HOG には 16 個の空間ビン（手の重心を中心に放射線ビン 4 つと 90° 刻みに方向ビンを 4 つ）があり，各々の空間ビンに対して, [26] と同様に 9 個の勾配方向ビンが設定されている（図 3 参照）．計 144 次元の HOG はエゴモーションヒストグラムと同様にオンライン処理によってクラスタリングされ，各々の映像セグメント（n フレーム区間）に対して BOF が生成される．

4. ディリシュレ過程トピックモデル

自己動作カテゴリを学習するためには，本提案手法では三つのディリシュレ過程多項分布混合モデル（Dirichlet Process Multinomial Mixture Model - DPMMM）を二

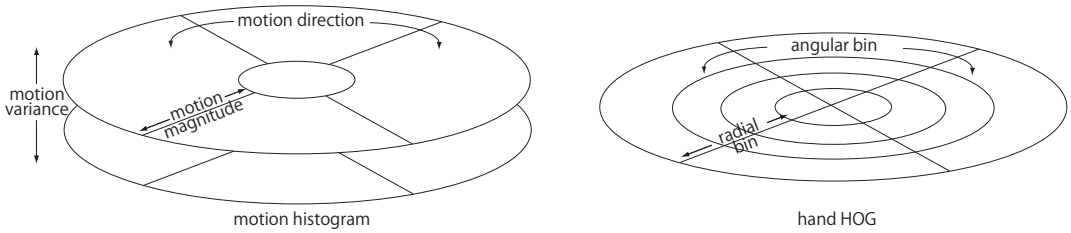

図 3　エゴモーションの 8 次元（4×2）ヒストグラムと 144 次元（4×4×9）の手の輪郭ヒストグラム．

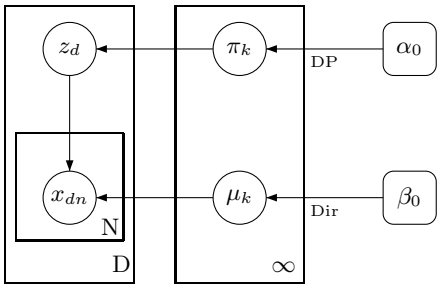

図 4　DPMMM のグラフィカルモデル．四角プレートは含まれたグラフ構造の複製，丸は確率変数，丸い角の四角は分布・過程のハイパーパラメタ．

つの段階に分けて用いる．第一段階では，一つの DPMMM でエゴモーションヒストグラムをクラスタリングし，同時に手の HOG も二つ目の DPMMM でクラスタリングする．クラスタリング処理を高速に行うために Sequential Update Greedy Search (SUGS) アルゴリズムを利用して，中華料理店過程（Chinese Restaurant Process - CRP）により高速なオンラインクラスタリングを実現する．第二段階では，一つの DPMMM を用いて，生成された D 個映像セグメントの BOFs から自己動作カテゴリを変分ベイズ推定で学習する．オンラインクラスタリングのための SUGS については第 4.1 節で紹介し，また変分ベイズ推定については 4.2 節で簡単に復習する（詳細は付録に記載）．DPMMM のグラフィカルモデルを図 4 に示す．

4.1　オンライン特徴点クラスタリング

映像データベースが大きいほど，抽出される特徴ベクトルの数が膨大になるため，実用性という観点からはメモリ使用の少ない高速なクラスタリング手法が望ましい．これに対して本研究では DPMMM を利用し，高速なオンライン・クラスタリング処理の可能性を示す．

本提案手法では低次特徴ヒストグラムのクラスタリングと BOFs の蓄積を，SUGS [27] アルゴリズムを利用して実現する．本手法の利点としては，処理が高速であると同時にクラスタの数もディリシュレ過程の枠組みで自動的に発見されることである（計算コストはクラスタ数に応じて線形である）．

新しい d 番目の観測 $\boldsymbol{x}_d$（低次特徴ヒストグラム）に対して最適なクラスタ $\hat{k}$ を選択し，観測の所属 z_{dk} を決める．最適なクラスタは以下の尤度関数（式 1）を最大にすることで求まる．

$$\hat{k} = \arg\max_k \left\{ p(z_{dk} | \boldsymbol{x}_d, \mathbf{X}^{-d}, \mathbf{Z}^{-d}; \alpha_0, \boldsymbol{\beta}_0) \right\} \qquad (1)$$

この尤度関数は，現在の観測 $\boldsymbol{x}_d$，過去の観測 $\mathbf{X}^{-d}$（$-d$ は d を含まないの意味），過去の所属 $\mathbf{Z}^{-d}$ 及びディリシュレ分布のハイパーパラメタ $\boldsymbol{\beta}_0$ とディリシュレ過程のハイパーパラメタ α_0 を条件とする確率である．

この尤度関数の利点は，クラスタの事前分布として CRP を利用している点である．CRP はディリシュレ過程から生成される無限次元離散分布の生成方法の一つであり [28]，CRP を利用することにより，今まで観測されていないクラスタの確率（式 A·6）が常に確保され，クラスタ数を必要に応じて増やすことができる．

4.2　変分ベイズ推定による自己動作カテゴリ学習

各々の映像セグメントのカテゴリの割り当ては，変分ベイズ推定により，観測の周辺尤度を最大化する適切なカテゴリを割り振ることにより実現する．以下の説明では，理解を促すために前節と同じ変数名を用いているが，ここでは別のものをさしている．各々の観測ベクトル $\mathbf{x}_d \in \mathbf{X}^D$ はエゴモーションの BOF と手の特徴の BOF を正規化して結合したものである．

最大化対象の周辺尤度の潜在変数は所属分布 z_d，ディリシュレ過程の混合確率 $\boldsymbol{\pi}$，ディリシュレ分布のパラメタ $\boldsymbol{\mu}$ である．与えられているハイパーパラメタ α_0 と $\boldsymbol{\beta}_0$ は，それぞれディリシュレ過程のハイパーパラメタとディリシュレ分布のハイパーパラメタである．ここで使用する添字 d は映像セグメントのインデクスを意味する．

$$\begin{aligned} p(\mathbf{X}; \alpha_0, \boldsymbol{\beta_0}) = & \int_{\boldsymbol{\pi},\boldsymbol{\mu}} p(\boldsymbol{\pi}; \alpha_0) p(\boldsymbol{\mu}; \beta_0) \\ & \times \prod_d \sum_{z_d} p(\mathbf{x}_d | z_d, \boldsymbol{\mu}_{z_d}) p(z_d | \boldsymbol{\pi}) d\boldsymbol{\mu} d\boldsymbol{\pi} \end{aligned} \qquad (2)$$

この周辺尤度は厳密に計算することができないため，変分ベイズ推定で事後分布を近似する分布 $\boldsymbol{q}$ を仮定し，ジェンセンの不等式により尤度の下限を反復的に最大化する．本研究ではディリシュレ過程を近似する事後分

布 $\boldsymbol{q}(\boldsymbol{\pi})$ に有限対称ディリシュレ分布（Finite Symmetric Dirichlet-FSD）を利用する．事後分布は以下の通りである．

$$\boldsymbol{q}(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{Z}; \alpha_0, \beta_0) = \prod_k q(\pi_k; \alpha_k) \prod_m q(\mu_{km}; \beta_{km}) \times \prod_d q(z_{dk}; \phi_{dk}) \quad (3)$$

ただし，α_k は k 番目の FSD 混合のハイパーパラメタ，μ_{km} は k 番目のディリシュレ分布の m 次元目のハイパーパラメタ，ϕ_{dk} は $z_d = k$ の確率である．多項分布混合モデルの事後分布 $\boldsymbol{q}$ のハイパーパラメタ α_k と μ_{km} 及び所属確率 ϕ_{dk} の更新式は付録に記載されており，尤度下限の一般的な導出は [29] にある．更新式の計算を繰り返すことによりギブスサンプリング等に比べて高速なカテゴリ学習とカテゴリ数の推定が行われる．

5. 実　　験

一人称視点の映像解析研究の多くは室内の安定性の高い動作を前提としている [12], [13] が，本研究では新たな動的な動きを含む野外映像と複数人物のインタラクション映像データセットを提案する．映像データセットの図 1 に示す．

5.1 実験環境

全ての実験映像は帽子に装着された GoProHD カメラで録画し，解像度は 840 × 480，フレームレートは 60 fps（実験では間引いて 30 fps），カメラの視野角は 170° である．

DPMMM のハイパーパラメタ α_0 と β_0 は全て 1 に設定し，DPMMM-SUGS の場合のみ, [27] のように α_0 を推定の枠組みに含んで周辺化している．変分ベイズ推定は 50 回乱数で初期化された所属変数 z_{dk} から計算し，その中から尤度最大のものを採用した．DP の事後確率を $K = 40$ 次元の FSD を利用して，真の DP を近似した．エゴモーションの BOF は運動の平均強度で正規化し，手の特徴の BOF はエッジ面積で正規化した．

真値と発見された動作カテゴリの関連付けは未知であるため，以下の流れで決定した．まず真のカテゴリと発見されたカテゴリの中からもっとも値の高い F-measure を持つ組を関連付けて，今後の候補として無効にする．次は同様に，残されたカテゴリの中から最適な組を関連付けて，候補がなくなるまで繰り返す．

5.2 広場映像データセット

広場映像データセットを用いて，エゴモーション特徴のみを利用した自己動作カテゴリ学習を行い，エゴモーションそのものの有効性を確認する．このデータセットには，*walk* (立ったまま歩く)，*walk-forward*, *jump*, *start-jump*, *run*, *walk-run*, *stand*, *right*, *left*, *left-right*, *up-down*, *right-left-up-down*，といった 12 種類の動作を含む．映像の内容は一連の動作パターンを野外で 3 回繰り返したものである．真値と学習結果を並べたものを図 5 に示す．広場の映像は 2 秒区間（60 フレーム）で区切り，124 個映像セグメントから形成されている．

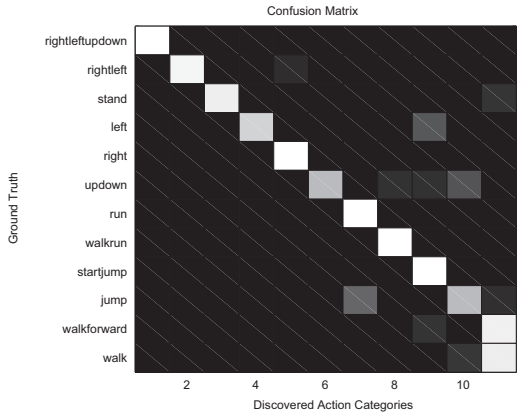

図 6　広場データセットの混同行列．

12 個の動作クラスの内 11 個のカテゴリが発見され，発見されなかった *walk* は *walk-forward* 非常に似ているため一つのカテゴリに合併されている．混同行列の対角要素の加重平均である平均分類精度は 83.4% であり，混同行列は図 6 に示す．

混同行列から，*up-down*，*jump*，*run* のような似た動作は時には混同されることが分かる．似た動作カテゴリ以外は問題なく学習され，エゴモーションのみでも動的な動作を学習することが可能であることを示した．

5.3 訪問者映像データセット

次の実験では，エゴモーションと手の形を特徴として併用した場合の学習性能を評価する．訪問者映像データセットでは，二人の営業マンが客のオフィスを訪ね，インタラクションを行う設定になっている．このデータセットはお辞儀，名刺交換，相槌，握手といった動作を含む 10 種類の動作から形成されている．

このデータセットにおける時系列の分類結果は図 7 に示す．混同行列を図 8 に示す．エゴモーションと手の情報を併用した場合の平均分類精度は 61.6% であり，12 個の動作カテゴリが学習された．手の情報のみを利用した場合の平均分類精度は 41.7% で 6 種類の動作が学習された．エゴモーションのみを利用した場合の平均分類精度は 33.6% であり，8 種類の自己動作が学習された．エゴモーションと手の情報を併用した方が精度の高い学習結果が得られることが分かる．具体的には，名刺交換を学習するためには手の形状（手の情報）とお辞儀（エゴモーション）を認識する必要がある．

5.4 障害物コース映像データセット

前述の実験ではエゴモーションの有効性，そして動作カテゴリを正しく学習するためにエゴモーションと手の情報を併用することの重要性を示した．最後の実験で

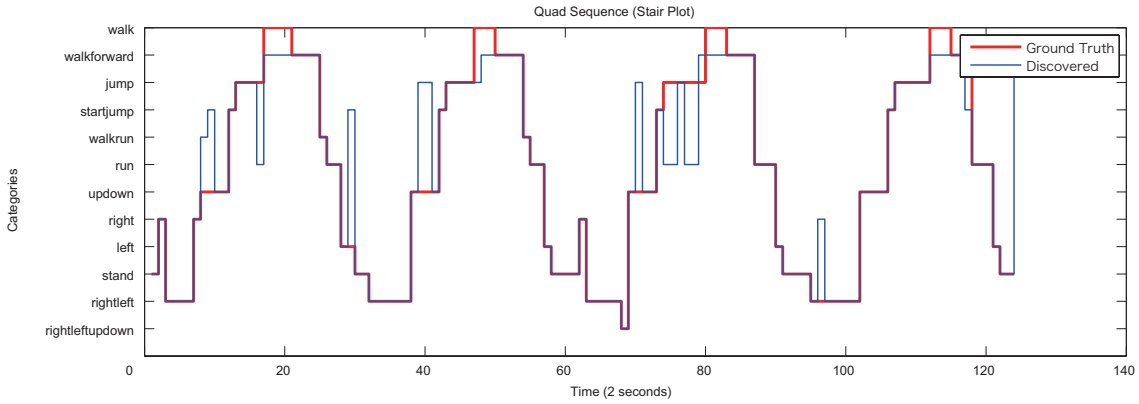

図 5　広場映像データセットの自己動作カテゴリ学習の結果と真値.

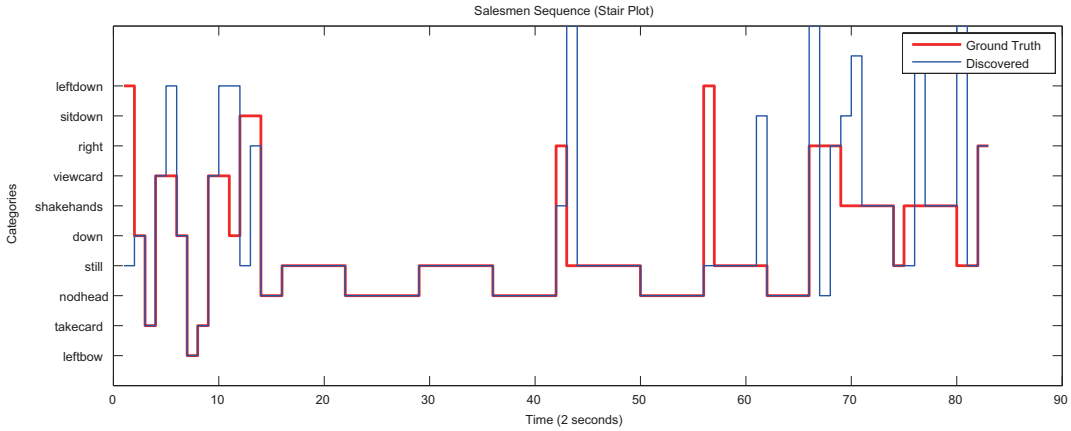

図 7　訪問者映像データセットの動作カテゴリの学習結果と真値.

は，さらに長く，より動的な動作を含んだ，障害物コース映像を準備し，本提案手法の頑健性を示す．このデータセットは，這う（*crawl*），走る（*run*），縄からぶら下がる（*slide-rope*），ネットを登る（*climb-net*）といったチャレンジングな動作を含む 19 種類の自己動作から形成されている．内容としてこれらの一連の動作パターンを 3 回繰り返し，カットなしの 25 分間の映像である．データセットは合計 766 個の映像セグメントから形成され，前述の実験と同様に 60 フレーム（2 秒）ごとに映像を区切っている．

エゴモーションと手の情報を併用した場合の平均分類精度は 37.81% であり，37 個の動作カテゴリが学習された．手の情報のみを利用した場合の平均分類精度は 32.17% で 34 種類の動作が学習された．エゴモーションのみを利用した場合の平均分類精度は 26.91% であり，15 種類の自己動作が学習された．情報を併用した場合の混同行列は図 9 に示す．混合行列から，真の動作カテゴリが複数カテゴリに分割さていることが分かる．複数のモードを持つカテゴリの教師無し学習は今後の課題である．

6. おわりに

本論文では，一人称視点における自己動作カテゴリの発見のために，エゴモーションの特徴と手の情報との併用が自己動作カテゴリ学習に重要であることを示した．さらに，ディリシュレ過程多項分布混合モデルが低次の特徴クラスタリングと高次の動作学習に有力であることを示した．実験では，エゴモーションの有用性及びエゴモーションと手の情報との併用の必要性を示し，さらに長時間における動的な動きを含む映像データセットから動作を学習し，本提案手法の有効性を示した．

文　献

[1] R. Megret, D. Szolgay, J. Benois Pineau, P. Joly, J. Pinquier, J.-F. Dartigues and C. Helmer: “Wearable video monitoring of people with age dementia: Video indexing at the service of healthcare”, Proceedings of the International Workshop on Content-Based Multimedia Indexing, pp. 101–108 (2008).

[2] T. E. Starner: “Wearable agents”, IEEE Pervasive Computing, **1**, pp. 90–92 (2002).

[3] K. Mikolajczyk and H. Uemura: “Action recognition with motion-appearance vocabulary forest.”, Proceedings of the IEEE Conference on Computer Vision and Patter Recognition (2008).

[4] I. Laptev, M. Marszałek, C. Schmid and B. Rozenfeld: “Learning realistic human actions from movies”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008).

[5] J. Liu, J. Luo and M. Shah: “Recognizing realistic actions from videos in the wild”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1996–2003 (2009).

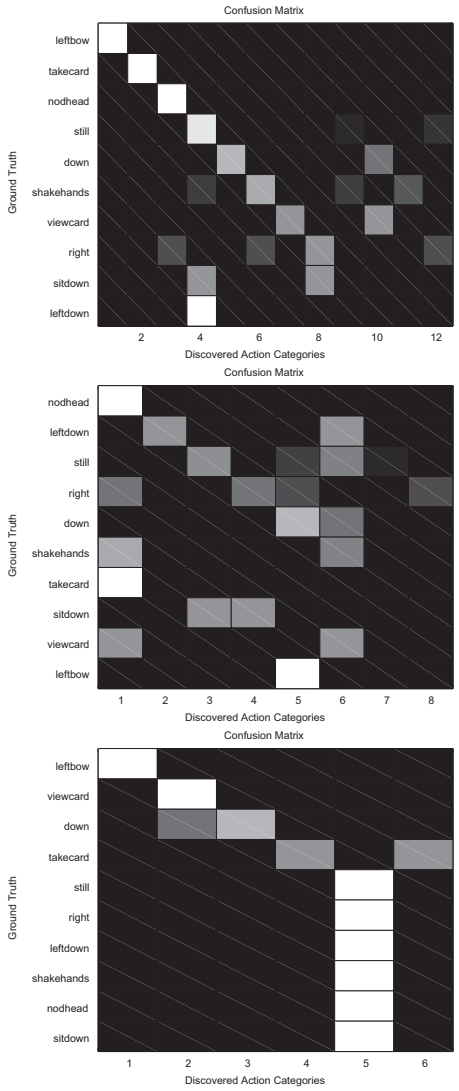

図 8　訪問者映像データセットの混同行列．エゴモーションと手の特徴の併用（上），エゴモーションのみ（中），手の情報のみ（下）．

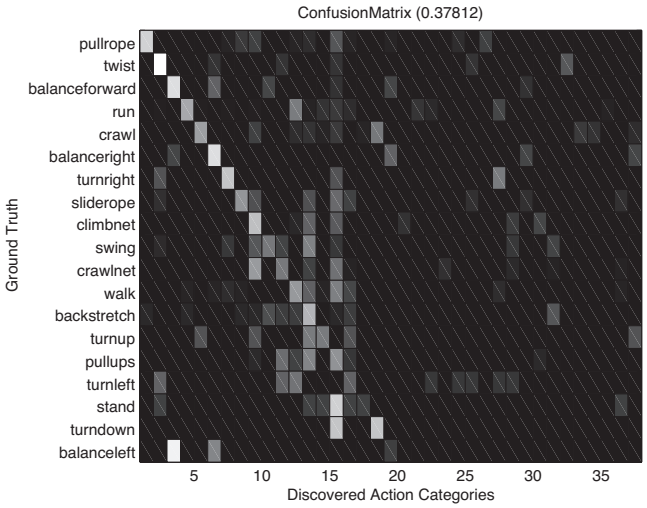

図 9　障害物コース映像の混同行列.

[6] B. Tordoff, W. Mayol, T. de Campos and D. Murray: “Head pose estimation for wearable robot control”, Proceedings of the British Machine Vision Conference, pp. 807–816 (2002).

[7] T. Starner, A. Pentland and J. Weaver: “Real-time American sign language recognition using desk and wearable computer based video”, IEEE Transactions on Pattern Analysis and Machine Intelligence, **20**, 12, pp. 1371–1375 (1998).

[8] T. Kurata, T. Okuma, M. Kourogi and K. Sakaue: “The hand mouse: GMM hand-color classication and mean shift tracking”, Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, p. 119 (2001).

[9] M. Kolsch, M. Turk, T. Hollerer and J. Chainey: “Vision-based interfaces for mobility”, Proceedings of the International Conference on Mobile and Ubiquitous Systems (2004).

[10] W. W. Mayol and D. W. Murray: “Wearable hand activity recognition for event summarization”, Proceedings of the Ninth IEEE International Symposium on Wearable Computers, pp. 122–129 (2005).

[11] L. Sun, U. Klank and M. Beetz: “EYEWATCHME: 3D hand and object tracking for inside out activity analysis”, Proceedings of the IEEE Workshop on Egocentric Vision, pp. 9–16 (2009).

[12] S. Sundaram and W. Cuevas: “High level activity recognition using low resolution wearable vision”, Proceedings of the IEEE Workshop on Egocentric Vision, pp. 25–32 (2009).

[13] E. H. Spriggs, F. D. la Torre Frade and M. Hebert: “Temporal segmentation and activity classification from first-person sensing”, Proceedings of the IEEE Workshop on Egocentric Vision (2009).

[14] T. Huynh, M. Fritz and B. Schiele: “Discovery of activity patterns using topic models”, Proceedings of the International Conference on Ubiquitous Computing, pp. 10–19 (2008).

[15] J. C. Niebles, H. Wang and L. Fei-Fei: “Unsupervised learning of human action categories using spatial-temporal words”, Proceedings of the British Machine Vision Conference, pp. III:1249–1258 (2006).

[16] S. Wong, T. Kim and R. Cipolla: “Learning motion categories using both semantic and structural information”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6 (2007).

[17] J. C. Niebles and L. Fei-Fei: “A hierarchical model of shape and appearance for human action classification”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007).

[18] K. M. Kitani, T. Okabe, Y. Sato and A. Sugimoto: “Discovering primitive action categories by leveraging relevant visual context”, Proceedings of the IEEE International Workshop on Visual Surveillance, pp. 1–8 (2008).

[19] R. Filipovych and E. Ribeiro: “Recognizing primitive interactions by exploring actor-object states”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008).

[20] X. Wang, X. Ma and E. Grimson: “Unsupervised activity perception by hierarchical Bayesian models”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007).

[21] J. Shi and C. Tomasi: “Good features to track”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (1994).

[22] B. D. Lucas and T. Kanade: “An iterative image reg-

istration technique with an application to stereo vision", Proceedings of the International Joint Conference on Artificial Intelligence, pp. 674–679 (1981).
[23] R. Hartley and A. Zisserman: "Multiple view geometry", Cambridge university press Cambridge, UK (2000).
[24] P. Dollár, V. Rabaud, G. Cottrell and S. Belongie: "Behavior recognition via sparse spatio-temporal features", Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005).
[25] I. Laptev: "On space-time interest points", International Journal on Computer Vision, **64**, 2, pp. 107–123 (2005).
[26] N. Dalal and B. Triggs: "Histograms of oriented gradients for human detection", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005).
[27] L. Wang and D. Dunson: "Fast Bayesian inference in Dirichlet process mixture models", Technical Report, Department of Statistical Science, Duke University (2008).
[28] D. J. Aldous: "Exchangeability and related topics", Ecole dEte de Probabilites de Saint-Flour XIII–1983 (1985).
[29] D. M. Blei and M. I. Jordan: "Variational inference for Dirichlet process mixtures", Journal of Bayesian Analysis, **1**, 1, pp. 121–144 (2006).
[30] S. Yu, K. Yu, V. Tresp and H.-P. Kriegel: "Variational Bayesian Dirichlet-multinomial allocation for exponential family mixtures", Proceedings of the European Conference on Machine Learning, pp. 841–848 (2006).

付　　録

SUGS オンラインクラスタリング

ここでは，4.1 節で紹介した DPMMM によるオンラインクラスタリング手法 SUGS を説明する．観測ベクトル $\mathbf{x}_d$ は d 番目の観測であり，観測 $\boldsymbol{x}_d$ の次元は V である．SUGS は観測 $\mathbf{x}_d$ の最適な所属クラスタ k をオンラインに推定し，$z_d = k$ は d 番目の観測がクラスタ k に所属することを意味する．

（1） 初めての観測 x_1 の所属 z_1 を $k = 1$ に初期化する．

$$z_1 = 1 \tag{A·1}$$

（2） 観測 $\boldsymbol{x}_1$ を利用して, クラスタ $k = 1$ の観測度数 $n(k_1, \mu_v)$ を更新し，$k = 1$ のディリシュレ分布のパラメタ $\boldsymbol{\mu}_{k_1} = \{\mu_1, \ldots, \mu_v, \ldots, \mu_V\}$ の確率の期待値を再計算する．

$$p(\mu_{k_1 v}|\boldsymbol{x}_1, z_1) = \frac{n(k_1, \mu_v) + \beta_0/V}{n(k_1) + \beta_0} \tag{A·2}$$

（3） $d > 1$ の場合

(a) 観測 $\boldsymbol{x}_d$ に対して，最適な所属 z_d を以下の尤度関数から推定する．

$$\hat{k} = \arg\max_k \left\{ p(z_d = k|\boldsymbol{x}_d, \mathbf{X}^{-d}, \mathbf{Z}^{-d}; \alpha_0, \beta_0) \right\} \tag{A·3}$$

ただし， $\boldsymbol{x}_d$と $\mathbf{X}^{-d}$ はこれまでの全ての観測を意味し，$\mathbf{Z}^{-d}$ は過去の全ての所属（$-d$ は d を含まないことを意味する）を表す．

(b) $\hat{k}$ のクラスタの観測度数 $n(\hat{k}, \mu_v)$ を更新し，$\boldsymbol{\mu}_{\hat{k}}$ の確率の期待値を再計算する．

本研究では，ディリシュレ過程のハイパーパタメタ α_0 を推定の枠組みの中で周辺化し，式 A·4 のように複数の α_t を利用する．

$$p(\alpha_0) = \sum_t p(\alpha_t)\delta_{\alpha_t}(\alpha) \tag{A·4}$$

実験では 0.1 ∼ 160 の範囲の 8 個の離散値を持つ一様分布を利用したが，広い範囲であれば結果への影響は少ない．これ以降の α の推定は式 A·5 で逐次的に求まる．

$$p(\alpha_t|\mathbf{X}^d, \mathbf{Z}^d) \propto p(\alpha_t|\mathbf{X}^{-d}, \mathbf{Z}^{-d}) p(z_{dk}|\alpha_t, \mathbf{Z}^{-d}) \tag{A·5}$$

観測の所属 z_{dk} の条件付確率は式（A·6）の CRP を利用して計算する．CRP の更新式は，以下の通りである．

$$p_{crp}(z_{dk}|\alpha_t, \mathbf{Z}^{-d}) = \begin{cases} \frac{N_k^{-d}}{\alpha_t + (d-1)} \\ \frac{a_t}{\alpha_t + (d-1)} \end{cases} \tag{A·6}$$

クラスタ所属 z_{dk} の事後確率は以下のように求まる．

$$\begin{aligned} p(z_{dk}|\mathbf{X}^d, \mathbf{Z}^{-d}; \beta_0) &= p(z_{dk}|\mathbf{X}^{-d}, \mathbf{Z}^{-d}) \\ &\quad \times p(\boldsymbol{x}_d|\mathbf{X}^{-d}, \mathbf{Z}^{-d}; \beta_0) \qquad \text{(A·7)} \\ &\propto p(z_{dk}|\mathbf{X}^{-d}, \mathbf{Z}^{-d}) \\ &\quad \times \prod_{n_d} \frac{n(k, v)^{-d} + \beta_0/V}{n(k)^{-d} + \beta_0} \qquad \text{(A·8)} \end{aligned}$$

ただし，

$$\begin{aligned} p(z_{dk}|\mathbf{X}^{-d}, \mathbf{Z}^{-d}) &= \sum_t p(z_{dk}|\alpha_t, \mathbf{Z}^{-d}) \\ &\quad \times \ p(\alpha_t|\mathbf{X}^{-d}, \mathbf{Z}^{-d}) \qquad \text{(A·9)} \end{aligned}$$

ガウシアン混合モデルを利用した例と SUGS に関する詳細は [27] にある．

DPMMM 変分ベイズ推定の更新式

周辺尤度の下限を最大化するための更新式は以下の通りである．

$$\alpha_k = \frac{\alpha_0}{K} + \sum_d \phi_{dk} \tag{A·10}$$

$$\beta_{kv} = \frac{\beta_0}{V} + \sum_d x_{dv}\phi_{dk} \tag{A·11}$$

$$\begin{aligned} \phi_{dk} \propto \exp\Bigg(& \sum_v x_{dv} \left\{ \Psi(\beta_{kv}) - \Psi(\sum_{v'} \beta_{kv'}) \right\} \\ & + \Psi(\alpha_k) \Bigg) \end{aligned} \tag{A·12}$$

ただし，x_{dv} はデータ d における特徴 v の度数，$\Psi()$ はディガンマ関数である．また，FSD [30] を利用しているため所属変数 ϕ_{dk} の更新式は，truncated stick-breaking 過程 [29] を利用した更新式と異なる．