視覚的文脈を考慮した人物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††

† 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,takahiro,ysato}@iis.u-tokyo.ac.jp, ††sugimoto@nii.ac.jp

あらまし 近年, Bag-of-words アプローチは文書解析に次いで, 一般物体認識と行動認識に適用され, その有用性が示されている. 但し, 映像にもとづく動作カテゴリの学習手法は動き特徴のみを用いており, 動作に関連している物体や背景のアピアランス特徴を用いていなかった. ここで本研究では, 人物の動きのみならず, 視覚的文脈をも考慮し, 人物映像のデータベースから動作カテゴリを教師無しで学習する手法を提案する. 具体的には, 動作カテゴリを学習するための, (1) 動き特徴と視覚的文脈の二つを考慮した生成モデルと (2) 大量のデータを処理するためのクラスタリング手法を提案する. 実験では, 視覚的文脈を用いた際の改善を示し, 物体を用いる動作を中心としたデータベースから動作カテゴリが自動的に得られることを示す. 更に, 複雑な背景を持つシーンから動作カテゴリを学習することにより本手法の有用性を示す.

キーワード 動作分類, 視覚的文脈, 教師無し学習, bag of features, 潜在変数モデル

Unsupervised Action Category Discovery Using Visual Context

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

† The 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,takahiro,ysato}@iis.u-tokyo.ac.jp, ††sugimoto@nii.ac.jp

Abstract Under the bag of words framework we aim to learn primitive action categories from video without supervision by leveraging relevant visual context. We implement a bi-modal latent variable model that utilizes both motion features as well as relevant visual context, and a two stage clustering technique using nearest representative point clustering and non-negative matrix factorization to deal with the large number of features produced by video. Our experiments show that the combination of relevant visual context and motion features improve the performance of action discovery and that our method is able to leverage relevant visual features for action discovery despite the presence of irrelevant background objects.

Key words Action classification, visual context, unsupervised learning, bag of features, latent variable model

1. はじめに

本稿では, 大量の人物映像データベースから動作カテゴリを抽出するという課題に着目し, 自動的に人物のプリミティブ動作を学習する枠組を提案する. プリミティブ動作とは, 短い時間で行われる動作を意味し, コップを手に取る, 本のページを捲るなど, 数秒で認識できる動作である. 人物の高次的な行動はプリミティブ動作 (以降, 動作) から構成されている ([1]〜[3]) ということから, 動作の学習は映像における人物行動を理解するための重要な課題である.

近年, 確率的潜在変数モデルを用いて, 教師無しで映像データベースから人物の動作を学習するという研究が盛んに行われている ([4]〜[6]). **確率的潜在変数モデル**(混合モデル [7], PLSA [8], LDA [9], HDP [10]) は, bag-of-words のアプローチに基づき, 一つの文書を単語の集合として表し, 文書中のトピックを学習するモデルでもある.

潜在変数モデルを人物の行動学習に適用した例としては, Niebles らの [4] 研究がある. 彼らは PLSA [8] をビデオに応用し, 文書の代わりにビデオ, そして単語の代わりに時空間 (ST) 特徴を用いて, フィギュアスケート映像からスピンの種類 (カテゴリ) を教師無しで学習した. そして, 彼らの研究は, 人物映像も文書と同様に動作を特徴の集合として表すことができることを示した. しかし, ここで用いられた PLSA モデルでは時空間特徴という単一のモードしか扱っていないため, 後述するよう

図 1　視覚的文脈を用いた動作学習：動作と関係のある視覚的特徴 (緑) と動作と関係のない特徴 (紫) を区別し動作カテゴリを学習する.

に視覚的文脈を考慮していない.

人物の動作は, 動き (時空間的特徴) と見え (空間的特徴) から構成されていることが経験から分かる. そして, 脳科学の分野においても同じ結論に至り, 人は動作を動きとその動作に関係する物体の見えによって認識を行うと言われている [11]. 例えば, ピアノを弾く時の手の動きとキーボードを使う時の手の動きはよく似ているが, ピアノやキーボードの存在 (見え) によって, 異なる動作だと簡単に区別することができる. このような動作に関連している物体または背景の見え (空間特徴) を**視覚的文脈**と呼び, 本研究では人物動作を動きと視覚的文脈の組み合わせとして考える.

しかし, これまで提案されてきた動作学習の研究は視覚的文脈を考慮していない. 例えば, Wong らは時空間特徴と特徴の位置情報を扱うモデルを提案し, 体や顔や手の動作を学習した [5]. しかし, 彼らの枠組みでは, 視覚的文脈を用いて似ている動作を区別することはできない. 同じく, Wang ら [6] が提案した手法はフレーム間の画素値の変化を用いて車や歩行者の動きを記述しているが, 認識の手掛かりとなる移動物体の見えを無視している.

視覚情報を用いた研究としては, Fanti ら [12] と Niebles ら [13] の研究があるが, 人物の形状に関する情報をモデルに加え, 体の部品の見えを表しているものである. そのため, 人の体という対象に特化されており, 他の対象へ適用することは容易ではないという問題が存在する. また, 本稿で提案する手法との大きな相違点としては, 動作に関係する物体や認識の手掛かりとなる背景の特徴を考慮していないという点がある.

本研究では, 対象物体の事前情報 (形状情報) に依存しない二つのモード (動きと視覚的文脈) を考慮した動作学習手法を提案する. 実験では, 視覚的文脈を用いることによる学習結果の改善を示し, 物体を扱う動作を中心としたデータベースから動作カテゴリが自動的に得られることを示す. 更に, 複雑な背景を持つシーン (図 1) でも, 背景の視覚的ノイズに影響されず, 動作カテゴリを正しく学習することにより本手法の有用性を示す.

2. 提 案 手 法

本研究の目的は映像データベースから人物動作を学習することである. この目標を達成するため, まず映像から視覚的文脈 (空間的特徴) と動作の動き (時空間的特徴) を抽出する (2.1 節). 次に, 2 段階クラスタリング手法を用いて, 各ビデオセグメントを特徴のヒストグラムで表す (2.2 節). そして, 最後に二つのモードを持つ潜在変数モデルを用いて, 各々のビデオセグメントのヒストグラムから, ビデオに含まれる動作カテゴリを教師無しで学習する (2.3 節).

Algorithm 1 – Nearest representative point clustering

```
for every video segment d in corpus d do
    Initialize histogram v_d = 0
    for every extracted feature x_di do
        Find the nearest representative point c_j to x_di
        if L_2(x_di, c_j) > θ then
            Create new representative point c_k ← x_di
            Set count of cluster v_dk = 1
        else
            Increment count v_dj of nearest cluster c_j
        end if
    end for
end for
```

2.1 空間的特徴と時空間的特徴の抽出

ここでは空間的特徴と時空間的特徴を抽出方法を説明する. まず, 空間的特徴は各々の映像のフレームから SIFT 特徴点 [14] を抽出し, 正規化された 128 次元ベクトルを特徴量として得る. なお, ここでは SIFT 特徴と用いているが, 他の特徴点や特徴量も使うことができる.

時空間特徴は [15] と同様に, サイズ $7 \times 7 \times 4$ (7×7 の画像平面を 4 フレーム分) の時間勾配ボリュームを各々の画素の近辺から抽出する. ボリュームの各々の要素は画素値の時間微分である. そして, 特徴量はボリュームの要素を並べた 196 次元ベクトルである. なお, 他の特徴点, 時空間キューボイド [16] や時空間点 [17] も使うことができる.

2.2 特徴の二段階クラスタリング

2.2.1 オンライン・クラスタリング

各ビデオセグメントから抽出された特徴をクラスタリングし (コードブックを作成), 各ビデオセグメントを特徴ヒストグラムとして表したい. しかし, 映像データから抽出される大量の特徴を処理するために, 膨大な計算コストがかかる K 平均法のようなオフライン手法は避けたい. ここで, 本手法はクラスタリングの第一段階とし, ビデオセグメントから得られる特徴のコードブック作成とビデオセグメントのヒストグラム作成を同時に行う高速なオンラインクラスタリング法を利用する. 具体的には Algorithm 1 のようなクラスタリング法を定義し, それを最近代表点クラスタリングと呼ぶ.

まず, 映像データベース $\mathbf{d}$ からビデオセグメント d が与えられているとする. 最近代表点クラスタリングでは, ビデオセグメント d から得られる各々の特徴 $\mathbf{x}_{di}$（セグメント d の i 番目の特徴ベクトル）に対して, 新たなクラスタ $\mathbf{c}_j$ を作るか, 既存の最近代表点 $\mathbf{c}_k$ の度数 v_{dk}（セグメント d のヒストグラムの k 番目の度数）を一つ増やすかを, 閾値 θ で決定する. 具体的には, 特徴 $\mathbf{x}_{di}$ から最近代表点 $\mathbf{c}_k$ の L_2 距離が θ より大きい場合, 新しいクラスタの代表点 $\mathbf{c}_j = \mathbf{x}_{di}$ を作成する. また, 距離が θ より小さい場合は最近代表点の度数 v_{dk} に一つ足す. すべての特徴についてクラスタリングを行い, 結果的には n 個の

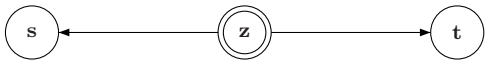

図 2　2 モード潜在変数モデル：モデルは動作カテゴリ $\mathbf{z}$ と空間特徴 $\mathbf{s}$ と時空間特徴 $\mathbf{t}$ から構成される．

クラスタの代表点集合 $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ と各クラスタの度数を持つヒストグラム $\mathbf{v}_d = (v_{d1}, \cdots, v_{dn})^T$ が得られる．そして，m 個のビデオセグメントを処理することによりヒストグラム行列 $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_m)$ が得られる．クラスタの数 n はビデオセグメントを処理するたびに増える可能性があるので，次元を統一するために，以前に処理したヒストグラムの後ろを 0 で埋める処理を行う．なお，空間特徴と時空間特徴のそれぞれについてこのクラスタリング処理を行う．

2.2.2　非負行列因子分解

第一段階では高速なオンライン処理を行った反面，映像データベース全体の特徴を考慮しなかった．第二段階の役割は，各ビデオセグメント間の関係を考慮した上でヒストグラムの次元削減を行うことである．ここで，データ $\mathbf{V}$ を非負部分空間 $\mathbf{H}$ に射影する非負行列因子分解 (NMF) [18] を用いる．人物動作は必ず正に特徴を生成する (負の特徴は存在しない) ことから，NMF は主成分分析のように正と負を混合した基底を用いる次元削減手法より相応しいと言える．

NMF では $n \times m$ の非負データ行列 $\mathbf{V}$(列がビデオセグメントのヒストグラム) を $n \times r$ の基底行列 $\mathbf{W}$ と $r \times m$ の符号 (係数) 行列 $\mathbf{H}$ に分解する．

$$\mathbf{V} \approx \mathbf{WH}. \tag{1}$$

結果として得られる符号行列 $\mathbf{H}$ は，データ $\mathbf{V}$ を r 次元空間に射影したものである．

二つのモードに対して NMF を独立に行い，空間と時空間ヒストグラム行列 $\mathbf{V}^s$ と $\mathbf{V}^t$ をそれぞれの符号空間 $\mathbf{H}^s$ と $\mathbf{H}^t$ に射影する．$\mathbf{H}^s$ と $\mathbf{H}^t$ は文書解析で使用される単語対文書行列の様なものであり，行列の要素 $n(w,d)$ は単語 w が文書 d の中で発生した度数である．なお，NMF と PLSA は multinomial PCA の一例であることが [19] で示されている．本手法では，潜在変数モデルのカテゴリ $\mathbf{z}$ の次元 q と NMF の部分空間の次元 r を同じ値に設定しているので，第二段階の NMF を各モードにもとづく動作カテゴリの学習としても解釈できる．

2.3　動作モデルによるカテゴリ学習とモードの統合

ベイジアンネットワークの枠組みでは，各変数の条件付き独立性を仮定することにより，確率変数の結合確率をより簡潔に表現できる．ここで，[7] で提案された単一モードの混合モデルを拡張し，動作カテゴリ変数 $\mathbf{z}$ により空間特徴変数 $\mathbf{s}$ と時空間特徴変数 $\mathbf{t}$ を独立した観測として扱う 2 モード混合モデルを提案する (図 2)．一つのビデオセグメント $d \in \mathbf{d}$ の確率は以下のように表す．

$$p(d) = \sum_z p(d|z)p(z) \tag{2}$$

$$p(d|z) \propto \prod_{s \in d} p(s|z) \prod_{t \in d} p(t|z) \tag{3}$$

$$= \prod_s p(s|z)^{n(s,d)} \prod_t p(t|z)^{n(t,d)} \tag{4}$$

動作カテゴリ z が与えられた時の空間特徴 s と時空間特徴 t の独立性により，ビデオセグメントの確率を特徴の条件付き確率の積として表すことができる (式 4)．条件付き確率の指数 $n(s,d)$ と $n(t,d)$ はビデオセグメント d に含まれている空間特徴 s と時空間特徴 t の正規化された度数を表している．

2.3.1　パラメータ学習

動作モデルのパラメータを学習するために，ビデオデータベース $\mathbf{d}$ の尤度を最大にするパラメータ $p(s|z)$ と $p(t|z)$ と $p(z)$ を求める．

$$\log p(\mathbf{d}) = \sum_d \log \sum_z p(d|z)p(z) \tag{5}$$

最適な局所解を求めるために EM アルゴリズムを用いる．事後確率 $p(z|d)$ における完全データ尤度 $E[\mathcal{L}^c]$ を最大にすることによりデータ尤度の下限を最大にすることができることから，以下の関数を最大にするパラメータを求める．

$$E[\mathcal{L}^c] = \sum_{d,z} p(z|d) \log p(d|z)p(z) \tag{6}$$

最初の E ステップでは潜在変数の事後確率をベイズ定理で求める．

$$p(z|d) = \frac{p(d|z)p(z)}{\sum_{z'} p(d|z')p(z')} \tag{7}$$

パラメータは乱数で初期化する場合が多いが，本手法では NMF の次元 r と動作カテゴリの次元 q が等しいため，一つのモードの正規化された符号行列 $\mathbf{H}$ を $\hat{p}(d|z)$ の初期値として使用する．

次に M ステップでは，完全データ尤度とパラメータの条件から形成されるラグランジュ関数の極値を求める．データ尤度を最大にするパラメータの再推定方程式は以下のように導出することができる．

$$\hat{p}(s|z) \propto \sum_d n(s,d)p(z|d) \tag{8}$$

$$\hat{p}(t|z) \propto \sum_d n(t,d)p(z|d) \tag{9}$$

$$\hat{p}(z) \propto \sum_d p(z|d) \tag{10}$$

以後 E ステップと M ステップを繰り返し，対数尤度が最大値収束するまで計算を続けることによって最適なパラメータが得られる．

2.3.2　認識と推定

提案手法の範囲外であるが，前述のように動作モデルのパラメータを学習することにより，得られたベイジアンネットワークで認識を行うことも可能である．例えば，学習で得られた空間特徴のクラスタ集合（コードブック）$\mathbf{C^s}$ を用いて，入力ビデオセグメント d の空間ヒストグラム $\mathbf{v}_d^\mathbf{s}$ を作成し，NMF で学習された空間基底行列 $\mathbf{W^s}$ で係数ベクトル $\mathbf{h}_d^\mathbf{s}$ を [20] と同様に求める．時空間特徴の係数ベクトル $\mathbf{h}_d^\mathbf{t}$ も同様に計算できる．$\mathbf{h}_d^\mathbf{s}$ と $\mathbf{h}_d^\mathbf{t}$ を正規化することにより $n(t,d)$ と $n(s,d)$ が得られ，信念伝搬 [21] を利用してカテゴリ $\mathbf{z}$ の分布が求まる．

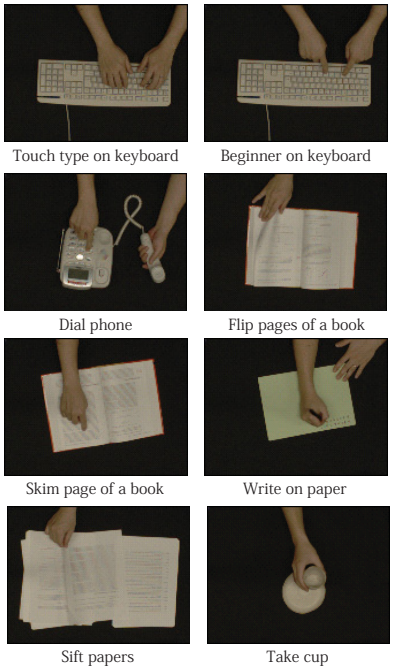

図 3　動作と物体のデータベース C_{OBJ}：8 種類の動作と物体の組が含まれている．

3. 実　　験

公開されている人物動作のデータセットは単純な背景を利用するのみで, 動作に関連している物体や背景が含まれていない ([5], [16], [22]). ここで動作と動作に関係のある物体と背景を含む新たなデータベースを提案する. これらのデータベースを用いて, 本提案手法は動作の動きと共に動作の視覚的文脈を利用し, より正確な動作カテゴリの学習ができることを示す.

3.1　動作データベース

3.1.1　動作・物体コーパス

動作と物体のデータベース C_{OBJ} は 8 種類の物体を扱う動作で構成され (図 3), 視覚的文脈を用いて異なる動作を学習する実験で使用される. 動作の内容は以下の通りである.

(1) キーボードをブラインドタッチで打つ (touch keyboard)
(2) 初心者がキーボードを打つ (beginner on keyboard)
(3) 電話をかける (dial phone)
(4) 本のページを捲る (flip page)
(5) 指で本を走り読みする (skim page)
(6) ペンで紙に書く (write paper)
(7) 紙をより分ける (sift paper)
(8) コップを手に取る (take cup)

各動作の映像を 3 秒間隔で区切り, 一つの動作に対して 5 つのセグメントを利用し, データベースを合計 40 個のビデオセグメントで構成している. 各セグメントは 90 フレームであり, 解像度は 160×120 である. ここではビデオセグメントの長さは同じであるが, 長さを統一する必要はない.

3.1.2　動作・背景コーパス

動作と背景のデータベース C_{BG} は 3 種類の動きと 3 種類の背景を含み, 合計 9 種類の動作で構成され (図 4), 視覚的文脈を用いて似ている動作を学習する実験で使用される. 動きと背景の種類は以下の通りである.

(1) 手で取る (Take)　　(1) ゲーム背景 (Game)
(2) 手で拭く (Wipe)　　(2) 工具背景 (Tools)
(3) 手を開く (Open)　　(3) 料理背景 (Cooking)

一つの動作に対し 5 つのセグメントがあり, データベースは合計で 45 個のビデオセグメントで形成されている. 各セグメントの長さは 90 フレームであり, 解像度は 160×120 である.

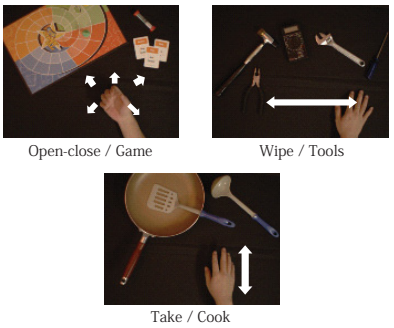

図 4　動作と背景データベース C_{BG}：3 種類の動きと 3 種類の背景で 9 種類の動作を含む. 白い矢印は動きを示す.

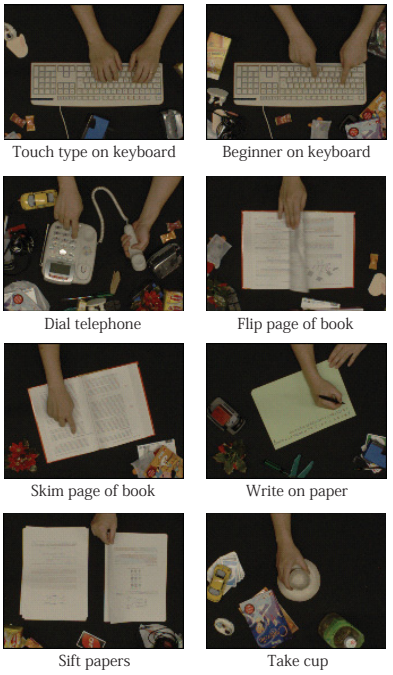

図 5　データベース C_{BGOB}：8 種類の動作と物体が含まれ, 各ビデオセグメントの背景 (周辺物体) が異なる.

3.1.3　動作・物体・背景コーパス

動きと物体と背景を含むデータベース C_{BGOB} は最初に紹介したデータベース C_{OBJ} と同じ動作を含むが, 各セグメントの背景は異なる (図 5). 背景には様々な動作と関係のない物体が置かれ, セグメント毎に内容が異なる. C_{OBJ} と同様に合計 40 個のセグメントで構成され, 解像度も同等である. このデータベースは, 動作と関係ない空間的特徴を含む背景から異なる動作を学習する実験で使用される.

表 1　従来手法で学習した動作カテゴリの平均確率行列.

	Discovered Actions							
	2	4	5	8	6	1	3	7
Touch-key	**0.97**	0.01	0.02	0.00	0.00	0.00	0.00	0.00
Begin-key	0.02	**0.82**	0.05	0.09	0.01	0.01	0.00	0.00
Dial-phone	0.02	0.00	**0.97**	0.00	0.00	0.00	0.00	0.00
Flip-page	0.00	0.00	0.00	**0.53**	0.03	0.40	0.00	0.03
Skim-page	0.00	0.00	0.02	0.00	**0.96**	0.00	0.00	0.01
Write-paper	0.01	0.48	0.27	0.09	0.03	**0.01**	0.00	0.11
Sift-paper	0.01	0.01	0.01	0.02	0.00	0.01	**0.94**	0.00
Take-cup	0.00	0.00	0.03	0.40	0.00	0.00	0.00	**0.56**

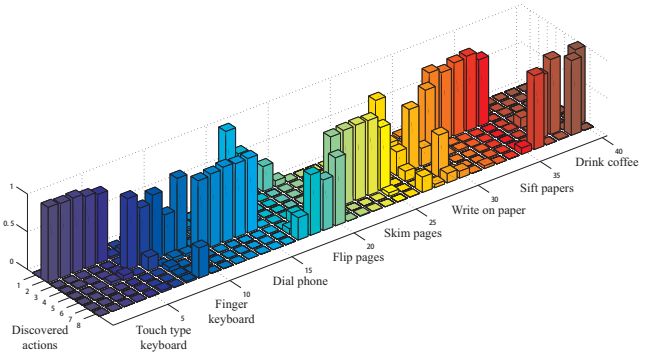

図 6　従来手法の解析結果：従来手法 [4] で動作と物体データベース C_{OBJ} を解析した結果 (時空間特徴のみ).

3.2　実験結果

最初にベースライン実験として従来手法を用いて, 動き特徴で動作カテゴリの学習を行う. 次に, 3 つのデータベースに対して 3 つの実験を行い, 提案手法による学習結果の改善を示す.

最近点クラスタリングの距離の閾値は $\theta_t = 0.02$ と $\theta_s = 0.01$ に設定した. ヒストグラム行列 $\mathbf{V}$ の 8 つの主成分 (PCA を利用) でデータを K 平均法でクラスタリングし, NMF の係数行列 $\mathbf{H}$ の初期化を行った. 各々の実験について, 係数行列の次元 r と動作モデルのカテゴリ数 q は既知としているが, [23] のようにモデル選択基準を使用することも考えられる.

3.2.1　ベースライン実験：従来手法の結果

ベースライン実験として, [4] と同様に PLSA を用いて, 観測として時空間ボリューム (時間勾配のみ) を使用し動作カテゴリを学習した結果を示す. データベース C_{OBJ} の動作を正しく分類する確率 (Probability of correct categorization-PCC) は 72% である. PCC は平均確率行列 (表 1) の対角要素の平均である. そして, 平均確率行列の列は各カテゴリに属するセグメントの確率 $p(z|d)$(図 6) の平均値であり, PCC を最大にするように列の順番を決める. 平均確率行列から, 紙にペンで書くという動作の学習精度が低いことが分かる. この実験を通して, 視覚的文脈を考慮せず, 動きのみで動作カテゴリを正しく学習することは困難であることが分かる.

3.2.2　物体の見えを考慮した学習結果

提案手法を用いて, 動作と物体のデータベース C_{OBJ} から動作カテゴリを行い, PCC は 99.6% であった (図 7). すべての動作カテゴリが学習され, 従来の時空間特徴を利用したモデルに対して, 視覚的情報と時空間的情報を併用することにより, 学習精度が向上されていることが分かる.

3.2.3　背景の視覚的情報を用いた動作カテゴリの学習結果

ある動きと一緒に発生する視覚的特徴はその動作と強い関係

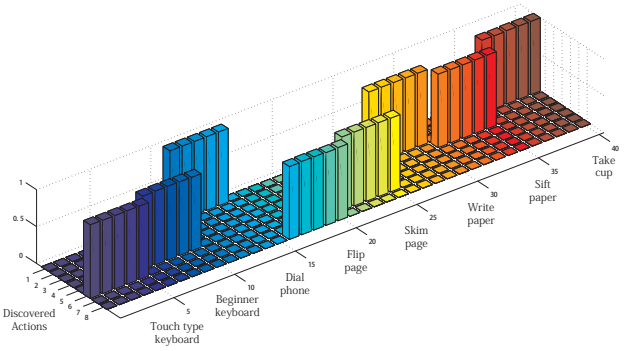

図 7　動作と物体データベース C_{OBJ} の解析結果：提案手法で 8 つの動作カテゴリが正しく分類されている.

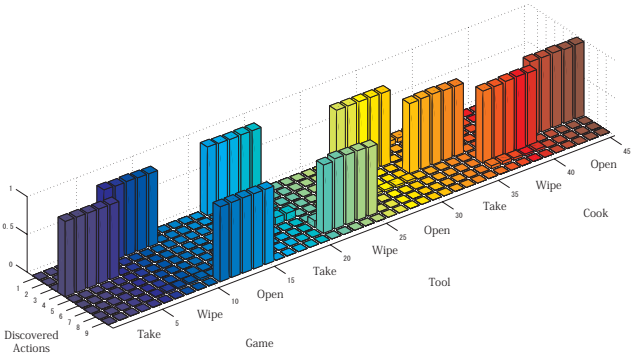

図 8　動作と背景データベース C_{BG} の解析結果：C_{BG} に含まれている 9 種類の動作カテゴリが学習されている.

があるといえる. ここでは, 動作と背景のデータベース C_{BG} を用いて, 本手法が似ている動作を視覚的情報を用いて区別できるか検証する. つまり, 3 種類の動きと 3 種類の背景から 9 種類の動作が学習されるか検証する. 結果は, 図 8 で示す通り, 9 種類の動きと背景の組合せが学習されていることが分かる. PCC は 95.7% である. このデータベースの場合, 動きと背景から発生した視覚的情報により, 各動作の分類が可能となった.

前述の通り, ある動きと一緒に発生する視覚的特徴はその動作と強い関係があるといえるが, 動きの種類に比べ, 背景の種類が少ない場合には問題が発生する. つまり, ある動作が同じ関係のない物体の前で何度も観測されると, その物体の特徴が関係のある視覚的特徴として学習される. しかし, 次の実験のように, 実際机上で手の動作を観測する際には, 背景が常に変化する傾向があり, 動作と関係のない視覚的特徴が学習されることは少ない.

3.2.4　視覚的文脈を用いた動作カテゴリの学習結果

実世界では, 動作は様々な環境の中で観測されるため, 動作と関係のある視覚的情報だけを区別する必要がある. ここでは, 動作と物体と背景データベース C_{BGOB} を用いて, 本手法が動作と関係のある視覚的情報のみを利用し, 背景に含まれる視覚的ノイズ (動作と関係のない物体) に影響されず, 動作カテゴリを正しく学習出来ることを示す.

本手法のデータベース C_{BGOB} における PCC は 98.2% であり, 結果は図 9 で示す. 背景による視覚的ノイズに頑健である理由の一つとしては, 動作から発生する特徴が一定であるのに対して, 動作に関係のない背景から発生する特徴はセグメント

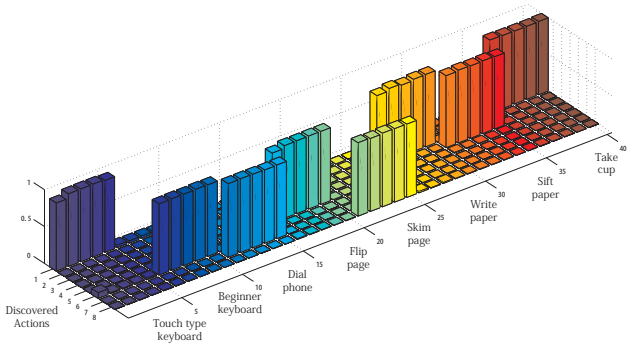

図 9　動作と物体と背景データベース C_{BGOB} の解析結果：散かっている机上環境でも, 8 種類の動作カテゴリが学習されている.

毎に違うので, NMF の次元削減の段階で関係のある視覚的特徴が残されているからである. 二つのモードから得られた情報を動作モデルで統合し, 8 種類のプリミティブ動作カテゴリを正しく学習している.

4. おわりに

人の動きだけでなく, 動作の視覚的文脈を使うことにより, プリミティブ動作をより正確に教師無しで学習する手法を提案した. 提案手法では, 二つの段階でクラスタリングを行った. 第一段階では最近代表点クラスタリングを行い, 高速なオンライン処理で同時にコードブック生成とヒストグラム生成を実現した. 第二段階では非負行列因子分解を用いて, ヒストグラムの次元削減を行い, 各モードから動作の種類を求めた. クラスタリングの結果を, 二つのモードを持つ確率的潜在変数モデルを用いて, 動作の動きと動作の視覚的文脈の両方を考慮し, 動作カテゴリを学習した. 実験を通して, 視覚情報を用いることにより動きのみを考慮した手法では学習できなかった動作を, 学習できることを示した. 更に, 本手法は視覚情報から発生するノイズに対しても頑健であり, 動作に関係のある視覚的文脈を用いて, 正しくプリミティブ動作カテゴリを教師無しで学習できることを示した.

文　献

[1] D. J. Moore and I. A. Essa: “Recognizing multitasked activities from video using stochastic context-free grammar”, Proceedings of the National Conference on Artificial Intelligence, pp. 770–776 (2002).

[2] R. Hamid, A. Y. Johnson, S. Batta, A. F. Bobick, C. L. Isbell and G. Coleman: “Detection and explanation of anomalous activities: Representing activities as bags of event n-grams”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. I: 1031–1038 (2005).

[3] K. M. Kitani, Y. Sato and A. Sugimoto: “Recovering the basic structure of human activities from a video-based symbol string”, Proceedings of the IEEE Workshop on Motion and Video Computing, pp. 9–9 (2007).

[4] 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).

[5] 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).

[6] 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).

[7] K. Nigam, A. McCallum, S. Thrun and T. Mitchell: “Text classification from labeled and unlabeled documents using EM”, Machine Learning (1999).

[8] T. Hofmann: “Probabilistic latent semantic analysis”, Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp. 289–29 (1999).

[9] D. M. Blei, A. Y. Ng and M. I. Jordan: “Latent Dirichlet allocation”, Journal of Machine Learning Research, **3**, pp. 993–1022 (2003).

[10] Y. W. Teh, M. I. Jordan, M. J. Beal and D. M. Blei: “Hierarchical Dirichlet processes”, Journal of the American Statistical Association, **101**, 476, pp. 1566–1581 (2006).

[11] A. H. Fagg and M. A. Arbib: “Modeling parietal–premotor interactions in primate control of grasping”, Neural Networks, **11**, 7-8, pp. 1277–1303 (1998).

[12] C. Fanti, L. Zelnik-Manor and P. Perona: “Hybrid models for human motion recognition”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1 (2005).

[13] 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).

[14] D. G. Lowe: “Object recognition from local scale-invariant features”, Proceedings of the International Conference on Computer Vision, p. II:1150 (1999).

[15] O. Boiman and M. Irani: “Detecting irregularities in images and in video”, Proceedings of the International Conference on Computer Vision, pp. I:462–469 (2005).

[16] 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).

[17] I. Laptev: “On space-time interest points”, International Journal on Computer Vision, **64**, 2, pp. 107–123 (2005).

[18] D. D. Lee and H. S. Seung: “Learning the parts of objects by non-negative matrix factorization”, Nature, **401**, pp. 788–791 (1999).

[19] W. Buntine: “Variational extensions to EM and multinomial PCA”, Proceedings of the European Conference on Machine Learning, pp. 23–34 (2002).

[20] O. Okun and H. Priisalu: “Fast nonnegative matrix factorization and its application for protein fold recognition”, EURASIP J. Appl. Signal Process., **2006**, 1, pp. 62–62 (2007).

[21] J. Pearl: “Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference”, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1988).

[22] C. Schuldt, I. Laptev and B. Caputo: “Recognizing human actions: A local SVM approach”, Proceedings of the International Conference on Pattern Recognition, pp. 32–36 (2004).

[23] A. Vinokourov and M. Girolami: “A probabilistic framework for the hierarchic organisation and classification of document collections”, Journal of Intelligent Information Systems, **18**, 2-3, pp. 153–172 (2002).