視覚的文脈を用いた人物動作のカテゴリー学習*

木谷　クリス　真実[†a)]　　岡部　孝弘[††b)]　　佐藤　洋一[††c)]　　杉本　晃宏[†††d)]

Using Visual Context for Action Category Discovery*

Kris M. KITANI[†a)], Takahiro OKABE[††b)], Yoichi SATO[††c)], and Akihiro SUGIMOTO[†††d)]

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．しかし，映像に基づく動作カテゴリー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ーを教師なしで学習する手法を提案する．具体的には，動作カテゴリーを学習するための，（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ーが自動的に得られることを示す．更に，複雑な背景をもつシーンから動作カテゴリー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，教師なし学習，bag of features，潜在変数モデル

1. まえがき

本論文では，大量の人物ビデオデータベースから動作カテゴリー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みを提案する．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，本のページをめくるなど，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（以後，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，教師なしでビデオ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（カテゴリー）を教師なしで学習した．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おらず，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いわれている [11]．例えば，ピアノを弾くときの手の動きとキーボードを使うとき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

† 電気通信大学大学院情報システム学研究科，調布市
Graduate School of Information Systems, University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan
†† 東京大学生産技術研究所，東京都
Institute of Industrial Science, The University of Tokyo, 4–6–1 Komaba, Meguro-ku, Tokyo, 153–8505 Japan
††† 国立情報学研究所，東京都
National Institute of Informatics, 2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430 Japan
a) E-mail: kitani@is.uec.ac.jp
b) E-mail: takahiro@iis.u-tokyo.ac.jp
c) E-mail: ysato@iis.u-tokyo.ac.jp
d) E-mail: sugimoto@nii.ac.jp
* 本論文は第 12 回画像の認識・理解シンポジウム推薦論文である．

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合せとして考える．

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，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歩行者の動きを記述しているものの，認識の手掛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，本論文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ーが自動的に得られることを示す．更に，複雑な背景をもつシーン（図 1）でも，背景の視覚的ノイズに影響されず，動作カテゴリーを正しく学習することにより本手法の有用性を示す．

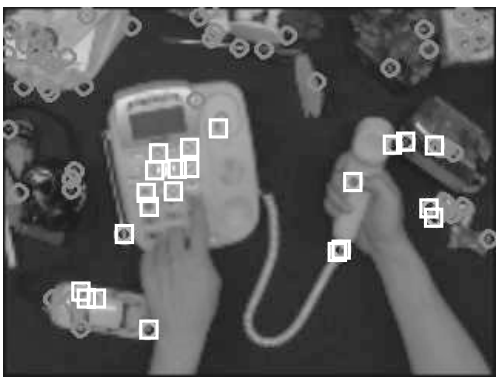

図 1　視覚的文脈を用いた動作学習：動作と関係のある視覚的特徴（□）と動作と関係のない特徴（○）を区別し動作カテゴリーを学習する．

Fig. 1　Leveraging visual features for action recognition: Relevant visual features (squares) induced by using the telephone and irrelevant features (circles) produced by unrelated background objects.

2. 提案手法

本研究の目的は人物動作を含むビデオデータベースから人物動作のカテゴリーを学習することである．この目標を達成するため，まずビデオデータベースを短い（数秒間）ビデオセグメントに分割し，各ビデオセグメントから視覚的文脈（空間的特徴）と動き（時空間的特徴）を抽出することにより（**2.1**），動作の特徴を得る．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ヒストグラムとして表現する（**2.2**）．最後に二つのモードをも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データベースに含まれる動作カテゴリーを教師なしで学習する（**2.3**）．

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ビデオフレームから SIFT 特徴点 [14] を抽出し，正規化された 128 次元ベクトルの集合を得る．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 7×7×4（7×7 の画像平面を 4 フレーム分）の時間こう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．

なお，対象動作の見え及び動きの変形に耐性をもたせるために本手法では局所特徴を利用し，bag-of-features（ヒストグラム）表現を採用している．

2.2　特徴の 2 段階クラスタリング

2.2.1　オンラインクラスタリング

次は，各ビデオセグメントから抽出された特徴集合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す．ただし，ビデオデータベースから抽出される大量の特徴を処理するために全データを記録する K 平均法のようなバッチ処理は避けたい．ここで，本手法はクラスタリングの第 1 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

Algorithm 1 – 最近代表点クラスタリング

```
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_k to x_di
    if L_2(x_di, c_k) > θ then
      Create new representative point c_j ← x_di
      Set count of cluster v_dj = 1
    else
      Increment count v_dk of nearest cluster c_k
    end if
  end for
end for
```

ム作成を同時に行う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．なお，本手法は leader-follower clustering [18] と異なり，クラスタ中心点の更新は行っていないためクラスタ数が増加する反面，データの順序に影響され難いという性質がある．

まず，ビデオデータベース（ビデオセグメントの集合）$\mathbf{d}$ からビデオセグメント $d \in \mathbf{d}$ 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，しきい値 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もつヒストグラム $\mathbf{v}_d = (v_{d1}, \cdots, v_{dn})^T$ 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 1 段階では高速なオンライン処理を行った反面，ビデオデータベース全体の特徴を考慮しなかった．第 2 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）[19] を用いる．本研究では人物の行動を特徴ヒストグラムとして表しているため，出現する特徴の数は必ず正である（若しくは零）ということから，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ふさわしいといえる．

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 $\mathbf{H}$ に分解するために以下の距離を最小化する．

$$\arg\min_{\mathbf{W},\mathbf{H}} L(\mathbf{V}, \mathbf{WH}) \tag{1}$$

ここで使用する距離 L は以下の Kullback-Leibler divergence である．

$$L(\mathbf{A}, \mathbf{B}) = \sum_{ij} \left(\mathbf{A}_{ij} \log \frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}} \right) \tag{2}$$

最小化の結果としてデータ行列 $\mathbf{V}$ が基底行列 $\mathbf{W}$ と係数行列 $\mathbf{H}$ に分解される．

$$\mathbf{V} \approx \mathbf{WH}. \tag{3}$$

ここで得られる符号行列 $\mathbf{H}$ は，データ $\mathbf{V}$ を r 次元空間に射影したものとなる．ただし，r は事前に与える必要がある．

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ようなものであり，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 [20] で示されている．また，Kullback-Leibler divergence を使用することにより NMF は PLSA と等しいことが証明できる [21]．したがって，第 2 段階の NMF を各モードに基づく動作カテゴリー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ー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，動作カテゴリー

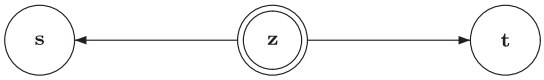

図 2　2 モード潜在変数モデル：モデルは動作カテゴリー **z** と空間特徴 **s** と時空間特徴 **t** から構成される
Fig. 2　Bi-modal latent variable model defined by the latent topic **z**, spatial features **s** and temporal features **t**.

変数 **z** により空間特徴変数 **s** と時空間特徴変数 **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．

$$p(d) = \sum_z p(d|z)p(z) \tag{4}$$

$$p(d|z) \propto \prod_{s\in d} p(s|z) \prod_{t\in d} p(t|z) \tag{5}$$

$$= \prod_s p(s|z)^{n(s,d)} \prod_t p(t|z)^{n(t,d)} \tag{6}$$

動作カテゴリー z が与えられたときの空間特徴 s と時空間特徴 t の条件付き独立性の仮定により，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 (6)）．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．

提案手法の動作モデルでは二つの観測モードを仮定しているため，関連物体による空間的特徴と動きによる時空間的特徴の両方が観測される必要がある．したがって，物体や背景に依存しない動き（例えば「走る」という動き）を学習するためには提案モデルは適していない．

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 **d** のゆう度を最大にするパラメータ $\theta = \{p(s|z), p(t|z), p(z)\}$ を求める．

$$\log p(\mathbf{d}) = \sum_d \log \sum_z p(d|z)p(z) \tag{7}$$

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ゆう度 $E[\mathcal{L}^c]$ を最大にすることによりデータゆう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 $\hat{\theta}$ を求める．

$$\hat{\theta} = \arg\max_\theta E[\mathcal{L}^c] \tag{8}$$

$$E[\mathcal{L}^c] = \sum_{d,z} p(z|d) \log p(d|z)p(z) \tag{9}$$

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．

$$p(z|d) = \frac{p(d|z)p(z)}{\sum_{z'} p(d|z')p(z')} \tag{10}$$

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 と動作カテゴリーの次元 q が等しいため，一つのモードの正規化された符号行列 **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，完全データゆう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．データゆう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．

$$\hat{p}(s|z) \propto \sum_d n(s,d)p(z|d) \tag{11}$$

$$\hat{p}(t|z) \propto \sum_d n(t,d)p(z|d) \tag{12}$$

$$\hat{p}(z) \propto \sum_d p(z|d) \tag{13}$$

以後 E ステップと M ステップを繰り返し，対数ゆう度が最大値に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．結果的に各ビデオセグメント d の動作カテゴリー分布 $p(z|d)$ が学習される．

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ともに動作の視覚的文脈を利用し，より正確な動作カテゴリー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とおりである．

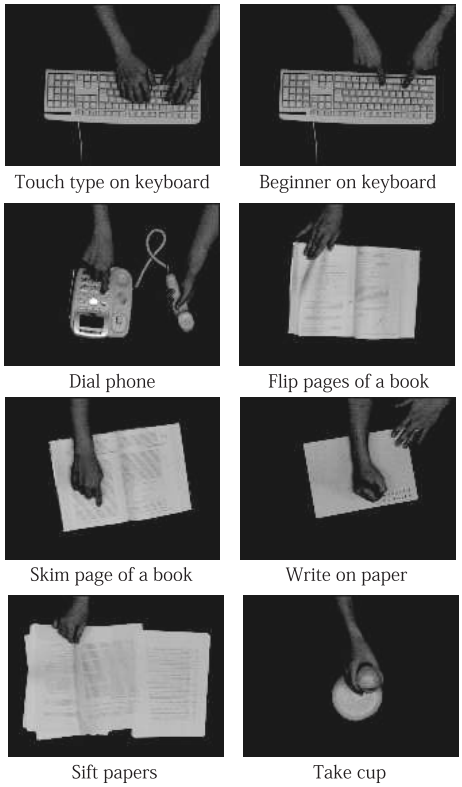

図 3　動作と物体のデータベース C_{OBJ}：8 種類の動作と物体の組が含まれている
Fig. 3　Key frames for the corpus C_{OBJ} with 8 desktop actions involving objects.

(1)　キーボードをブラインドタッチで打つ (touch type on 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)　本のページをめくる (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五つのセグメントを利用し，データベースを合計 $m_{OBJ} = 40$ 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．

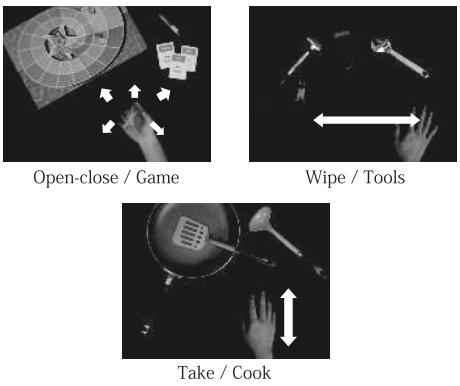

図 4　動作と背景データベース C_{BG}：3 種類の動きと 3 種類の背景で 9 種類の動作を含む．白い矢印は動きを示す．
Fig. 4　Examples from corpus C_{BG} with 9 actions including 3 different motions and 3 different background objects. Direction of motion is shown in white.

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とおりである．

(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)

一つの動作に対し五つのセグメントがあり，データベースは合計で $m_{BG} = 45$ 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．

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，セグメントごとに内容が異なる．C_{OBJ} と同様に合計 $m_{BGOB} = 40$ 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．

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ーの学習を行う．次に，三つのデータベースに対して三つの実験を行い，提案手法による学習結果の改善を示す．

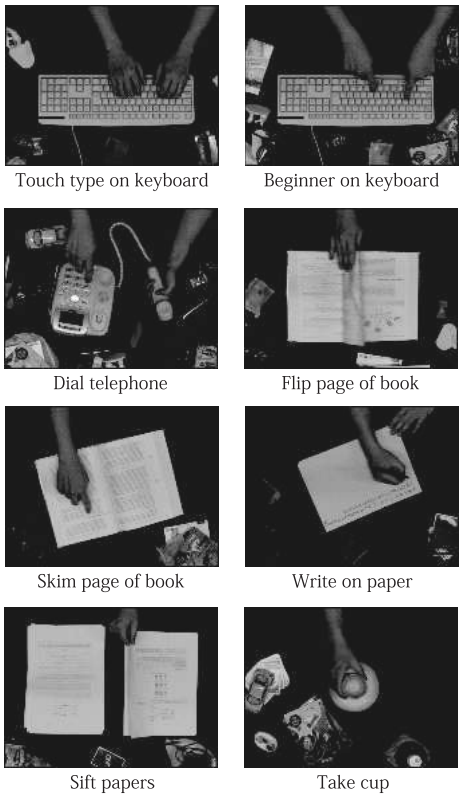

図 5　データベース C_{BGOB}：8 種類の動作と物体が含まれ，各ビデオセグメントの背景（周辺物体）が異なる

Fig. 5　Examples from corpus C_{BGOB} with 8 different actions with objects and varied random background objects for each video segment.

表 1　従来手法で学習した動作カテゴリーの平均確率行列

Table 1　Average posterior probabilities for each action category using only temporal features.

	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**

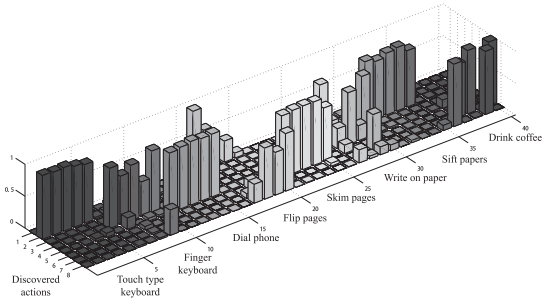

図 6　従来手法の解析結果：従来手法 [4] で動作と物体データベース C_{OBJ} を解析した結果（時空間特徴のみ）．

Fig. 6　Baseline results: pLSA using only temporal features for corpus C_{OBJ}. The horizontal axes gives the ground truth for each video d and the discovered action category z. The vertical axis is the posterior probability $p(z|d)$.

最近代表点クラスタリングの距離のしきい値は $\tau_t = 0.02$ と $\tau_s = 0.01$ に設定した．ヒストグラム行列 $\mathbf{V}$ の八つの主成分（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ー数 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（時間こう配のみ）を使用し動作カテゴリー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ー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，ペンで紙に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ー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} から動作カテゴリーを行い，PCC は 99.6%であった（図 7）．動作の数 $r = 8$ とし与え，クラスタリングの結果として $n_t = 471$ 次元と $n_s = 1577$ 次元のコードブックが得られた．すべての動作カテゴリーが学習され，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ーの学習結果

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があるといえる．ここでは，動作と背景の

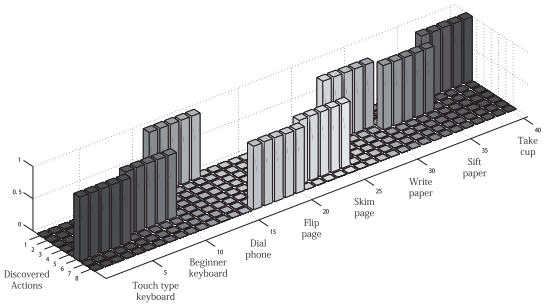

図 7 動作と物体データベース C_{OBJ} の解析結果：提案手法で八つの動作カテゴリーが正しく分類されている
Fig. 7 The resulting posterior $p(z|d)$ of corpus C_{OBJ} containing 8 different actions.

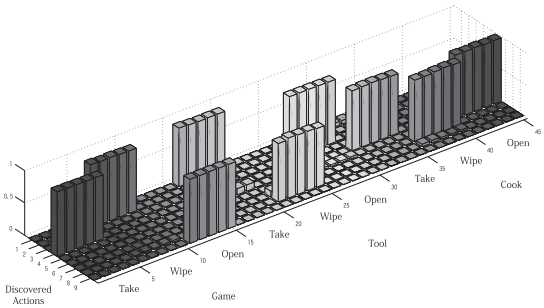

図 8 動作と背景データベース C_{BG} の解析結果：C_{BG} に含まれている 9 種類の動作カテゴリーが学習されている
Fig. 8 The resulting posterior $p(z|d)$ of corpus C_{BG} containing 3 motions and 3 backgrounds.

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（$r = 9$）の動作が学習されるか検証する．動作の数 $r = 9$ とし与え，クラスタリングの結果として $n_t = 100$ 次元と $n_s = 507$ 次元のコードブックが得られた．

結果は，図 8 で示すとおり，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．

前述のとおり，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．

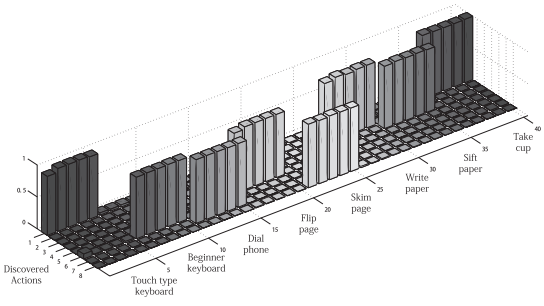

図 9 動作と物体と背景データベース C_{BGOB} の解析結果：散らかっている机上環境でも，8 種類の動作カテゴリーが学習されている
Fig. 9 The resulting posterior $p(z|d)$ of corpus C_{BG} containing 3 motions and 3 backgrounds.

3.2.4 視覚的文脈を用いた動作カテゴリーの学習結果

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，動作カテゴリーを正しく学習できることを示す．動作の数 $r = 8$ とし与え，クラスタリングの結果として $n_t = 388$ 次元と $n_s = 2794$ 次元のコードブックが得られた．

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ごとに違うので，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ーを正しく学習している．

4. むすび

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な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 1 段階では最近代表点クラスタリングを行い，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 2 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，二つのモードをも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，動作カテゴリーを学習した．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ーを教師な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,” Proc. 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,” Proc. 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,” Proc. IEEE Workshop on Motion and Video Computing, p.9, 2007.

[4] J.C. Niebles, H. Wang, and L. Fei-Fei, “Unsupervised learning of human action categories using spatial-temporal words,” Proc. 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,” Proc. 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,” Proc. 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,” Mach. Learn., vol.39, no.2-3, pp.103–134, 2000.

[8] T. Hofmann, “Probabilistic latent semantic analysis,” Proc. Conference on Uncertainty in Artificial Intelligence, pp.289–29, 1999.

[9] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol.3, pp.993–1022, 2003.

[10] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, “Hierarchical Dirichlet processes,” J. American Statistical Association, vol.101, no.476, pp.1566–1581, 2006.

[11] A.H. Fagg and M.A. Arbib, “Modeling parietal–premotor interactions in primate control of grasping,” Neural Netw., vol.11, no.7-8, pp.1277–1303, 1998.

[12] C. Fanti, L. Zelnik-Manor, and P. Perona, “Hybrid models for human motion recognition,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol.1, pp.1166–1173, 2005.

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

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

[15] O. Boiman and M. Irani, “Detecting irregularities in images and in video,” Proc. 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,” Proc. 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,” Int. J. Comput. Vis., vol.64, no.2, pp.107–123, 2005.

[18] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, Wiley-Interscience Publication, 2000.

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

[20] W. Buntine, “Variational extensions to EM and multinomial PCA,” Proc. European Conference on Machine Learning, pp.23–34, 2002.

[21] E. Gaussier and C. Goutte, “Relation between PLSA and NMF and implications,” Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.601–602, 2005.

[22] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local SVM approach,” Proc. 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,” J. Intelligent Information Systems, vol.18, no.2-3, pp.153–172, 2002.

（平成 20 年 10 月 15 日受付，21 年 3 月 2 日再受付）

木谷　クリス　真実

1999 南カリフォルニア大・工・電子卒．2000 ケーエルエー・テンコール（株）入社．2005 東京大学大学院情報理工学系研究科電子情報学専攻修士課程了．2008 同大学院同研究科同専攻博士課程了．現在，電気通信大学大学院情報システム学研究科助教．MIRU2008 学生優秀論文賞受賞．

岡部　孝弘　（正員）

1997 東大・理・物理卒．1999 同大大学院理学系研究科物理学専攻修士課程了．2000 同博士課程中退．2001 より東京大学生産技術研究所技官，助手を経て，現在同研究所助教．コンピュータビジョン，コンピュータグラフィックス，画像パターン認識に関する研究に従事．平 17 度本会論文賞，MIRU2004 優秀論文賞，平 16 年度 PRMU 研究奨励賞，平 19 年度情報処理学会山下記念研究賞などを受賞．情報処理学会，IEEE 各会員．

佐藤　洋一　（正員）

1990 東大・工・機械卒．1997 カーネギーメロン大学計算機科学部ロボティクス学科博士課程了．Ph.D in Robotics．同年より東京大学生産技術研究所研究機関研究員，講師，助教授を経て，現在同大大学院情報学環准教授．コンピュータビジョン，ヒューマン・コンピュータ・インタラクション，コンピュータグラフィックスに関する研究に従事．本会論文賞（平 17 年度，平 19 年度），平 11 年情報処理学会山下記念研究賞，平 11 年日本バーチャルリアリティ学会論文賞等を受賞．情報処理学会，日本バーチャルリアリティ学会，ACM，IEEE 各会員．

杉本　晃宏　（正員）

1987 東大・工・計数卒．1989 同大大学院工学系研究科修士課程了（数理工学専攻）．日立製作所基礎研究所，ATR，京都大学を経て，2002 より国立情報学研究所．現在，同研究所教授．総合研究大学院大学複合科学研究科教授併任．2006〜2007 Paris-Est 大学客員教授．博士（工学）．視覚情報処理や離散システム・アルゴリズムなどに広く興味をもち，数理的手法に基づいた手法を確立する研究に従事．2001 情報処理学会論文賞．IEEE，ACM，日本応用数理学会，等各会員．