単眼画像からの形状特徴を用いた動作認識法 ～人の位置・向きに頑健な動作認識器実現の試み～

下坂 正倫 †1　佐藤 真 †1　森 武俊 †1　佐藤 知正 †2　(東京大学)

1. 緒　　論

ロボットと人による円滑なコミュニケーションには人の動作の理解が重要である．人の日常生活のなかで用いるためには大掛かりな器具を人の体に装着するなどのわずらわしさがないことが必要である．本研究では，室内に備え付けられた一つのカメラから，「座っている」「立っている」などの人の日常動作を頑健に認識することを目的とする．

人の動作は姿勢や姿勢の動きとして発現されるものであるため，シルエットなどの形状情報は動作認識によく利用される．しかし，そのような形状情報はカメラの設置位置やカメラに対する人の位置などに依存して大きく変わってしまい問題となる．たとえば，同じ動作を映した画像でもカメラと人の位置関係が異なるときには同じ形状に見えないことがある．そこで，本研究では視点（カメラの設置位置）や人の位置・向きの変動に頑健な識別器の実現方法を述べる．

2. 形状特徴を用いた動作認識

2.1 認識器の構成

日常の動作は多様であり，同時に起こらないものや同時に起こりえるものなどさまざまであり，認識処理構成について考える必要がある．一般的なアプローチとして One-vs-All という方法を用いる．ある動作をしているかの2クラス分類認識器を並列に構成し，その結果を出力する方法である．この構成はいくつもの動作を複雑に認識する場合と比べ，各認識器の構造が単純であるというメリットがある．

2.2 Shape Context

形状情報を現す特徴量として Shape Context があり，Mori ら[1] をはじめとし，人の姿勢推定にも用いられている．Shape Context は，エッジ抽出などにより対象の形の代表点を算出し，ある一点を中心とする座標系でほかの点のヒストグラムを表したものである．Shape Context において各点を基準とするヒストグラムは図1のような距離と方向による log-polar ヒストグラムを用いて生成される．

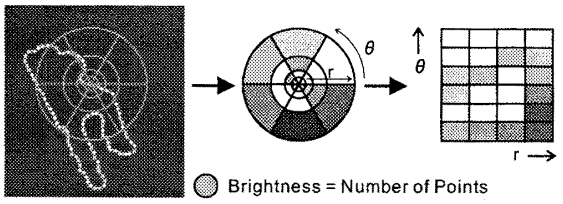

図1 Shape Context ヒストグラム

本手法では，シルエット画像の輪郭からその周に沿って均等に100点を代表点として，100点それぞれについて Shape Context ヒストグラムを生成し，特徴量とする．ヒストグラムは長さ方向に4，角度方向に6の4×6=24ビンとした．各代表点それぞれに24次元のヒストグラムが生成されるため，一枚の形状画像から生成される Shape Context ヒストグラム特徴量は24×100=2400次元である．本手法では，長さ方向の正規化として，シルエット画像が入る最小の円の二分の一をヒストグラムの外周とした．

2.3 Shape Context の変化に着目した特徴量

Shape Context ヒストグラムのテンプレートマッチングによる認識では代表点同士の対応の最適化，形状の回転に対応させることなどが必要となる．そのような対応付けの最適化は認識において計算処理量が大きくなるだけでなく，最適化をするか否かで表す情報が異なってしまう．具体的には，形状の回転に対応させる最適化をすることによって形状の向きという情報が失われてしまう．そこで，100点ある代表点からそのような対応関係の最適化を考えることなく形状情報の特性を表せるものとして Shape Context ヒストグラムの共分散行列を求めて認識に利用する．この Shape Context ヒストグラムの共分散行列は代表点すべてに対する各ビンの分散とビン間の相関を表す．代表点の対応付けの問題を避ける一方，輪郭形状の変化の様子を捉えた共分散行列を用いることで高い性能の認識器を構築することが，本研究の狙いである．全体もしくは一部のマッチングではなく，変化の相関を表す共分散行列を特徴とすることによってカメラと人の位置関係の違いにより形状が異なっている場合でも高い精度で認識できることを明らかにする．

2.4 ロジスティックモデルによる認識

その共分散行列を入力として，認識する動作ごとに2クラス分類認識器を構築する．認識器は確率のロジスティックモデルに基づいて認識を行う．Shape Context ヒストグラムの共分散行列である学習データを $\boldsymbol{X}_i \in \mathbb{R}^{b\times b}(i=1\cdots N$, N は学習データの数)，構築する認識器が認識する動作のラベルを，その動作をしている場合を1，していない場合を-1として，$y_i \in (-1,1)$ とする．各学習データ $\boldsymbol{X}_i$ での認識する動作ラベル y の確率 $p(y|\boldsymbol{X})$ は入力 $\boldsymbol{X} \in \mathbb{R}^{b\times b}$ からスカラ値への写像 $F(\boldsymbol{X}) : \mathbb{R}^{b\times b} \rightarrow \mathbb{R}$ を用いて，

$$p(y|\boldsymbol{X}) = \frac{1}{1+\exp(-yF(\boldsymbol{X}))} \quad (1)$$

と表す．これをロジスティックモデルといい，認識を行うときには，写像 $F(\boldsymbol{X})$ を適切に定めれば高性能な認識が実現できる．本研究では動画像の各フレームにおいて輪郭形状から Shape Context ヒストグラムを生成しその共分散行列を入力として $p(y|\boldsymbol{X})$ を算出して，逐次的に人の動作を認識していくものとする．

3. 学習による動作認識器の構築

写像 $F(\boldsymbol{X})$ を定める方法として，LogitBoost[2] による学習を用いる．LogitBoost は逐次的に生成した識別器の重み付け線形和で認識する認識器を構築する．前の識別器が不得意なサンプルを重点的に認識できるように生成されるため柔軟に識別することが可能である．逐次的に生成される識別器はその学習フェーズにおける最小二乗推定となるような識別器である．

3.1 LogitBoost

LogitBoost では全体の認識器を構成する認識器の数を K，初期値は $F(\boldsymbol{X}) \equiv 0$ とし，$F(\boldsymbol{X}) \rightarrow F(\boldsymbol{X}) + f_k(\boldsymbol{X})(k=$

†1 {simosaka,msato,tmori}@ics.t.u-tokyo.ac.jp
†2 tomomasasato@jcom.home.ne.jp

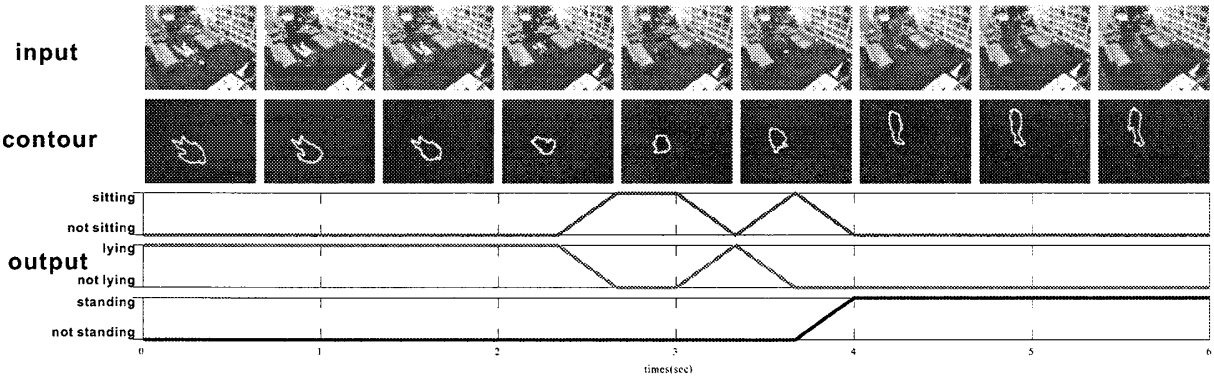

図 2　起き上がり動作の認識の様子

$1 \cdots K$) として $f_k(\cdot)$ を逐次的に生成してゆく．このとき $\omega_i = p(y=1|\boldsymbol{X}_i)p(y=-1|\boldsymbol{X}_i)$，$z_i = \frac{y_i}{p(y_i|\boldsymbol{X}_i)}$ として，

$$f_k = \arg\min_f \sum_{i=1}^{N} \omega_i (f(\boldsymbol{X}_i) - z_i)^2 \qquad (2)$$

にしたがって重み付き最小二乗法に基づいて f_k を最適化する．そして，得られた f_k から F を更新し，その後に z_i および ω_i も更新する．

3.2　リーマン多様体に基づく弱学習器

行列データと行列データの距離計量に着目したベクトルデータへの変換を施し，それに基づき重回帰分析法により (2) を実現することを考える．つまり，$\boldsymbol{h}_k(\boldsymbol{X}) : \mathbb{R}^{b\times b} \rightarrow \mathbb{R}^m$ および $g_k(\boldsymbol{x}) : \mathbb{R}^m \rightarrow \mathbb{R}$ となる関数を用いて $f_k(\boldsymbol{X}) = g_k(\boldsymbol{x})$，$\boldsymbol{x} = \boldsymbol{h}_k(\boldsymbol{X})$ として f_k を実現する．ここで，$\boldsymbol{h}_k(\boldsymbol{X})$ はリーマン多様体の考え方に基づき，このときのベクトルデータへの変換[3] を行うものとする．具体的には $\boldsymbol{h}_k(\boldsymbol{X}) : \mathbb{R}^{b\times b} \rightarrow \mathbb{R}^{\frac{b(b+1)}{2}}$ として，$\boldsymbol{h}_k(\boldsymbol{X}) = \mathrm{upper}(\log(\boldsymbol{M}_k^{-\frac{1}{2}} \boldsymbol{X} \boldsymbol{M}_k^{-\frac{1}{2}}))$ とする．$\mathrm{upper}(\boldsymbol{A})$ は $\boldsymbol{A}$ の上三角行列をベクトルにしたものであり，$\boldsymbol{M}_k$ は k 番目のフェーズにおいて共分散行列群 $\boldsymbol{X}_1 \cdots \boldsymbol{X}_N$ のリーマン多様体の枠組みにおける平均であり，$\boldsymbol{M}_k = \arg\min_{\boldsymbol{Y}} \sum_{i=1}^{N} \omega_i d^2(\boldsymbol{X}_i, \boldsymbol{Y})$ と定義する．ただし，$d^2(\boldsymbol{X}_i, \boldsymbol{Y})$ はリーマン多様体における $\boldsymbol{X}_i$ と $\boldsymbol{Y}$ の距離として定義する．k 番目の学習フェーズで i 番目のデータ $\boldsymbol{X}_i$ を変換したベクトル $\boldsymbol{x}_{i,k}$ は $\boldsymbol{\theta}_k$ をもちいて $g_k(\boldsymbol{x}_{i,k}) = \boldsymbol{\theta}_k{}^{\mathsf{T}} \boldsymbol{x}_{i,k}$ とする．

4.　実　　験

4.1　識別性能評価実験

提案手法の有効性を示すために性能評価実験を行った．入力として，リビングを想定した擬似居住空間で部屋の天井の隅に設置した IEEE1394 カメラで取得した 640pixel × 480pixel の画像を用いた．認識する動作は「座っている」「横になっている」「立っている」の三種類である．学習データとして各動作に対して立ち位置や姿勢などの異なる動画を 3fps で 6 秒間を 1 シーケンスとし，各動作に対して 12 シーケンス用意した．その 12 シーケンスを用いて leave one out 方式で交差検定をした．性能の評価には F 値を用いた．F 値は再現率 (対象動作の再現率) と精度 (対象動作を検知したときの正解率) の調和平均として表される．

提案手法の学習においては逐次的に生成する識別器の数 L を設定することが必要である．その数を大きくするほど性能が上がる可能性があるが，学習と認識にかかる時間も大きくなる．今回は $L = 10$ とした．

比較対象としては 1-NN 法を用いた．距離は一枚の画像から得られる 24 × 100 の Shape Context ヒストグラムの差の絶対値とした．距離の最適化として，$O(100!)$ のオーダーで対応点の全探索をするのではなく，$O(100)$ のオーダーですむように輪郭の開始点を探索してあとは周に沿って対応点とした．また，対応点のとり方を周に沿って変えていくのと同時に，Shape Context ヒストグラムの方向のとり方についても 60 度ごとに回転させた．提案手法および 1-NN 法の認識性能は次の表 1 のようにまとめられる．提案手法がカメラと人の位置関係が異なる場合でも識別性能の観点で提案手法が優れていることが確かめられた．

表 1　性能評価実験結果

Action	Sitting	Lying	Standing
提案手法 (F 値)	**0.94**	**0.90**	**0.99**
1-NN 法 (F 値)	0.88	0.78	0.82

4.2　遷移動作での認識の例

上述の実験で用いたすべての学習データを使用して学習させた認識器で「立ち上がる」という遷移動作に対して認識させた．結果は図 2 のようになった．立ち上がる際には少しぶれが生じているものの，立ち上がる経過中に座っているというラベルが付与されていることも含めて正確に認識できていることが確かめられた．

5.　結　　論

本研究では単眼画像からの画像を入力として日常動作の認識をするうえで，カメラの視点や人の位置・向きの変化に対して頑健に認識する手法を提案した．「座っている」や「立っている」などの簡単な動作に対して十分な性能があることや「立ち上がる」というような遷移する動作に対しても正確に認識できていることを確かめた．今後は認識する動作を増やしていくとともに，時系列や身体の一部に着目するなどの工夫により性能の高精度化をねらう．

参　考　文　献

1) G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV'02*, pp. 666–680.
2) J. Friedman et al. Additive logistic regression: A statistical view of boosting. In *Annals of Statistics*, Vol.28, pp. 337–407.
3) O. Tuzel et al. Human detection via classification on Riemannian manifolds. In *CVPR'07*.