ヒューマンインタフェース　95−9
コンピュータビジョンと
イメージメディア　129−9
（２００１．９．１３）

画素値分布特徴に基づく人物の検出

内海 章　　鉄谷 信二
ATR 知能映像通信研究所
〒 619-0288 京都府相楽郡精華町光台 2-2-2
{utsumi, tetutani}@mic.atr.co.jp

あらまし: 本報告では、画素値分布に基づく画像特徴を用いて画像内の人物を検出する手法を提案する。フレーム間差分やオプティカルフローなど、動きを前提にした人物検出処理はカメラの移動や背景の変動に対して脆弱であるという問題を持つ。提案手法では対象物体の動き情報を利用せず、人の見え方に関する統計モデルに基づいて検出を行う。このため、カメラの移動や背景の変動の影響を受けにくい。また、従来の統計モデルの多くが対象画像間の画素値そのものの近さを利用していたのに対し、本手法では物体の構造 (形状) によって生じる画素値分布の変化に着目して見え方をモデル化する。そのため、結果的に服装の違いや照明条件の変化の影響を受けにくい人物検出が可能となると考えられる。

Human Detection Method Using Statistical Model Based on Pixel Value Distributions

Akira Utsumi　　Nobuji Tetsutani

ATR Media Integration & Communications Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 JAPAN
{utsumi, tetutani}@mic.atr.co.jp

Abstract: We propose a statistical method to extract targets in images using common geometrical structures involved in the appearances of the target objects. Most appearance based methods focus on pixel values directly because the same classes of objects usually have similar pixel value distributions. However, this is not true for some particular objects. Humans are a good example. Human figures contain variety for different clothes, and pixel values (color, brightness) themselves can vary significantly from person to person. In this case, the geometrical structures observed as pixel value distances are essential for the recognition of object. In this paper, we propose a method to describe and recognize the appearances of objects based on geometrical structures. Using our method, objects having pixel value variety can be recognized with a small number of appearance models.

1　はじめに

物体認識はコンピュータビジョンにおける最も基本的な処理のひとつである。例えばランドマークの検出はロボットの誘導に重要であるし、各種追跡システムの安定性は目標の検出精度に大きく依存している。近年、画像を用いた人の動作認識・追跡についての研究が盛んであるが [1, 2, 3]、そこでも安定した目標検出は不可欠であるといえる。

しかしながら、従来提案されてきた人物追跡システムでは、計算コストを抑えるため、環境に対して強い仮定をおくことで検出処理を簡略化したものが多かった。そのような方法としてよく利用されるものに背景差分 (フレーム間差分) による動物体の抽出がある [3, 4, 5]。しかし、背景差分を利用するには、カメラ位置と背景の両方が固定されている必要があり、また追跡目標以外の物体移動は許されない。これらの点は、システムをより一般的な環境で動作させせようとした場合に大きな制約となる。

別のアプローチとして色情報の利用がある [6]。例えば、肌領域はほぼ一様な色をもつことから、顔や手の検出に色情報を使う提案もなされている。しかしながら、色情報は環境の変化に対して不安定で、また画像サイズが小さい場合にも領域検出は困難になる。

本報告では、これらの問題を解決するため、対象物体を動領域あるいは特定の色領域として抽出するのではなく、物体自身の見え方の特徴に基づいて検出する手法を検討する。そのために、対象物体の見え方に関する統計的な表現 (モデル) を作成し、検出処理に利用する。

物体の見え方の分布を統計的にモデル化する手法は物体認識において有用性が高く [7, 8]、例えば顔画像の検出についてもいくつかの提案が行われている。Schneiderman らは統計的なアプローチによって画像内の人の顔を検出する手法を提案している [9]。また、Osuna らはサポートベクターマシンを顔検出に応用している [10]。これら統計的な方法はしばしば大きな計算資源を要するが、この点も近年の計算機の性能向上によって克服されつつある。しかしながら、従来の統計モデルの多くは対象画像間の画素値そのものの近さを利用しており、人のように服装等により物体各部の色・明るさが変化する物体への適用は困難であった。

本報告では、幾何学的な構造 (形状) に基づいて物体の見え方をモデル化する手法を提案する。人のように様々なテクスチャを持つ物体では、物体各部に特定の画素値を仮定することはできないため、画素値そのものではなく画素値分布の変化に着目する。提案手法により、服装の違いや照明条件の変化の影響を受けにくい人物検出が可能となると考えられる。

次節では、提案手法の考え方について簡単に説明し、3 節ではモデルの計算方法および認識手法について述べる。4 節では提案手法の有効性を確認するために行った実験について述べる。5 節で今後検討すべき課題について整理し、6 節で本報告をまとめる。

2　画素値分布による人らしさの表現

提案手法の目的は、着衣の変化による色／テクスチャの違いなど、物体の画素値の多様性に対して共通な特徴を抽出し、これらの特徴を用いて物体認識 (検出) を行うことにある。

対象物体の色／明るさが一定であれば画素値は物体に関する重要な情報を与える。しかし、ある種の物体に関しては特定の画素値を持つことを期待できない。例えば、人物像は頭部、胴体、足、腕等のいくつかの部位からなり、その一部は衣服に覆われている。着衣の種類はほぼ無限にあると考えられるため、それらすべての色・模様をあらかじめモデルとして持つことは不可能である。

逆に、上記各部位の幾何学的な位置関係は人物像間で類似している。つまり、頭部、胴体、足などの部位間の相対的な位置、各部位の形状は服装の異なる人物間でも共通となり得る。提案手法では、画素値そのものではなく画素値の違いとして現れる形状情報に着目する。我々は対象物体の幾何学的構造をそのような画素値の違いとして抽出し、対象物体の見え方の統計的なモデルを構築することを目指す。すなわち、提案手法では画素値はある画像領域を他の領域から際立たせるためのラベルであると考えている。

画素値と独立な表現としては、輪郭 (エッジ) に基づく記述が知られている [11]。提案手法は、画素値そのものではなく画素値の違い (変化) を利用

するという意味で輪郭 (エッジ) に基づく認識手法の一種とも考えられる。しかしながら、従来の輪郭に基づく手法には一般に局所的な情報に敏感で、また画素値の持つ情報の多くが輪郭抽出の過程で失われるという問題があった。このことは、例えば人物検出処理において衣服の模様や背景ノイズの影響を受けやすいことを意味する。

我々は、提案手法によりこれら輪郭に基づく手法の持つ問題点を低減できると考えている。

3 アルゴリズム

3.1 距離マップの計算

提案手法では、前述の通り、画像領域間での画素値分布の近さ (距離) に基づいてモデルを構築する。利用可能な距離尺度としては種々のものが考えられるが、今回は計算の簡便さからマハラノビス汎距離を利用することとした。まず、入力画像をいくつかの領域 (ブロック) に分割し、全ての異なる 2 ブロック間の距離を計算する。各ブロック間について計算した距離値の集合をここでは「距離マップ」と呼ぶことにする。後で述べるように、距離マップの一部を認識処理に利用する。本節では、距離マップの計算方法について述べる。

まず、画像サイズを $m \times n$ とする (図 1 左)。ここで、各画素 $x_{s,t}(1 \leq s \leq m, 1 \leq t \leq n)$ は k 次元ベクトルとして記述できる。

$$\boldsymbol{x}_{s,t} = \begin{bmatrix} y_1 & y_2 & \cdots & y_k \end{bmatrix}'. \tag{1}$$

(例えば、濃淡画像であれば $k = 1$、カラー画像であれば $k = 3$ など。)

次に、入力画像を小ブロックに分割する。各ブロックは $p \times q$ 画素からなるものとし、水平方向に M、垂直方向に N 個のブロックが得られるとする (図 1 右)。ここでは、各ブロックに $1 \cdots MN$ の通し番号をつけて区別する (各ブロックを $\boldsymbol{X}_1 \cdots \boldsymbol{X}_{MN}$ と呼ぶ)。

続いて、各ブロック $\boldsymbol{X}_l$ についてブロック内の画素の平均ベクトルおよび共分散行列を計算する。

$$\bar{\boldsymbol{x}}_l = \frac{1}{pq} \sum_{(s,t) \in \boldsymbol{X}_l} \boldsymbol{x}_{s,t}, \tag{2}$$

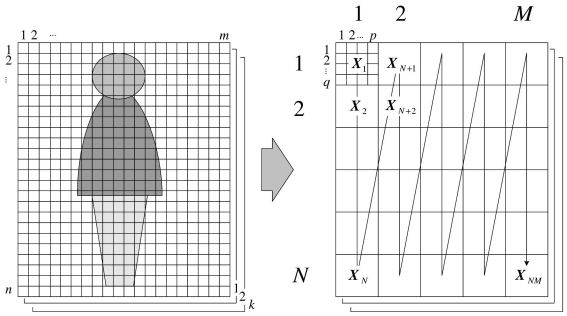


図 1: ブロック分割

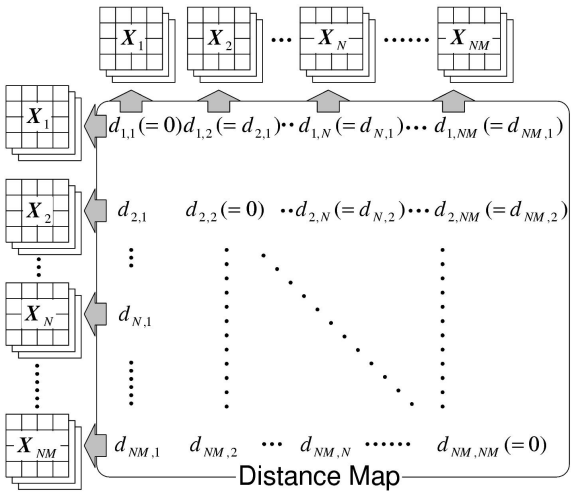


図 2: 距離マップの計算

$$\Sigma_l = \frac{1}{pq} \sum_{(s,t) \in \boldsymbol{X}_l} (\boldsymbol{x}_{s,t} - \bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{s,t} - \bar{\boldsymbol{x}}_l)'. \tag{3}$$

最後に各ブロック間のマハラノビス汎距離を計算し、距離マップ D を生成する (図 2)。

$$D = \begin{bmatrix} 0 & d_{(1,2)} & \cdots & d_{(1,MN)} \\ d_{(2,1)} & 0 & \cdots & d_{(2,MN)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{(MN,1)} & d_{(MN,2)} & \cdots & 0 \end{bmatrix},$$

ここで、

$$d_{(i,j)} = (\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_j)'(\Sigma_i + \Sigma_j)^{-1}(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_j). \tag{4}$$

図 3 に入力画像とその画像について計算した距離マップの例を示す。ここでは、60×90 画素から

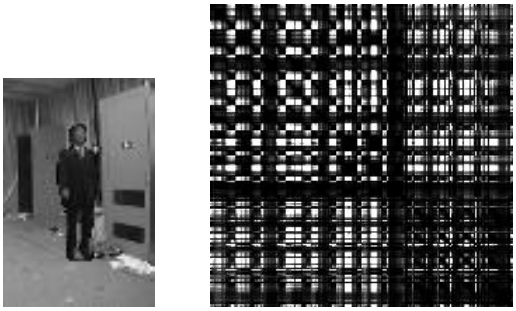


図 3: 距離マップの例 (左: 入力画像 (60 × 90 画素), 右: 距離マップ (228 × 228))

なる入力画像を 5 × 5 画素のブロックに分割した。結果として距離マップは 228 × 228 の大きさとなる。定義から明らかなように、距離マップは対角線 (左上-右下) に対して対称となる。

3.2 統計モデルの構築

前節で得られた距離マップをもとに、対象物体 (人物) 画像と背景画像を判別するための統計的なモデルを構築する。

まず、対象物体を含む場合 (人物) と含まない場合 (背景) の 2 種類について、サンプル画像を用意し、それらの画像について前節で述べた方法により距離マップを計算する。対象物体画像および背景画像各 K 枚から、(4) 式により対象物体画像についての距離マップ $D_1^{obj}, \cdots, D_K^{obj}$、背景画像についての距離マップ $D_1^{bck}, \cdots, D_K^{bck}$ を得る。ここでは、簡単のため両クラスのサンプル画像数は同数とした。図 4 に今回用いたサンプル画像の例を示す。なお、モデル構築のためには、各画像内で人物がほぼ同じ位置・大きさに現れる必要があることから、人物画像として背景画像に別途撮影した人のみの画像を合成したものを利用した。従って、人以外の背景部分については両者で共通である。今回は、人物・背景それぞれ 2000 枚の画像によってモデルを構築した。

まず、人物および背景それぞれの画像について、次式によりマップ内の各要素に関する平均 ($\bar{d}_{obj(i,j)}$, $\bar{d}_{bck(i,j)}$)・分散 ($\sigma^2_{obj(i,j)}$, $\sigma^2_{bck(i,j)}$) を求めた。

人物画像

背景画像

図 4: 学習用画像の例

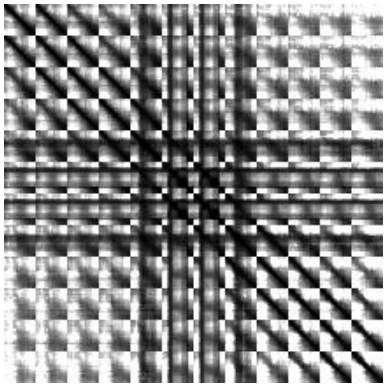
$$\bar{d}_{obj(i,j)} = \frac{1}{K}\sum_{k=1}^{K} d_{obj,k(i,j)}, \tag{5}$$

$$\bar{d}_{bck(i,j)} = \frac{1}{K}\sum_{k=1}^{K} d_{bck,k(i,j)}, \tag{6}$$

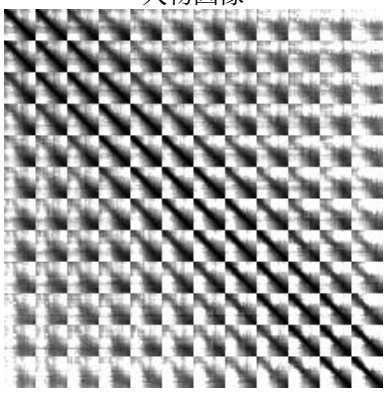
$$\sigma^2_{obj(i,j)} = \frac{1}{K}\sum_{k=1}^{K} \left(d_{obj,k(i,j)} - \bar{d}_{obj(i,j)}\right)^2, \tag{7}$$

$$\sigma^2_{bck(i,j)} = \frac{1}{K}\sum_{k=1}^{K} \left(d_{bck,k(i,j)} - \bar{d}_{bck(i,j)}\right)^2, \tag{8}$$

図 5 は、人物・背景それぞれについての距離マップの平均を示す。背景画像に対する距離マップの平均 (図 5 下) を見ると、距離値はマップ内で左上から右下に向かう対角線に近い要素では小さく、



人物画像



背景画像

図 5: 平均距離マップ

対角線から離れるに従い大きな値を持つことがわかる。これは、画像上で近い位置にある 2 つのブロックは似通った画素分布を持ち、2 ブロックが互いに遠く離れるほど異なる画素分布を持つ傾向が強いことを示している。なお、図 5 で縦方向および横方向に繰り返しパターンが見られるのは、本来縦横の 2 次元 ($N \times M$) に配置されているブロックを一次元に並べ替えたことに対応している (図 2、$\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{MN}$)。

一方で、人物画像に対する距離マップの平均についても背景と同様の傾向はあるものの、人の写っている領域に関連する中央部で背景画像とは異なることがわかる (図 5 上)。以下では、これらの違いに基づいてモデルを構築する。

ところで、距離マップはその全ての要素を認識 (検出) 処理に利用するとすると、数万次元 (前述の例では約 26000 次元) という膨大な次元数となり、学習・認識の両面で不都合が生じる。そのため、次元数の削減が必須となる。ここでは、以下に述べるような 2 段階の選抜法により、認識時に参照する要素数 (次元数) を削減する。

まず、図 5 の距離マップを各要素ごとに比較する。そのため、次式により両者の差分を求めた。

$$w_{(i,j)} = \frac{\left(\bar{d}_{obj(i,j)} - \bar{d}_{bck(i,j)}\right)^2}{\sigma^2_{obj(i,j)} + \sigma^2_{bck(i,j)}} \tag{9}$$

図 6 は (9) 式による比較の結果を示す。図中、明るい部分は人物画像と背景画像で距離マップ上の差が大きい要素を表しており、両者の識別に有効と考えられる。そこで、値の大きい順に r 個の要素を選択する ($(u_{r,1}, v_{r,1}), (u_{r,2}, v_{r,2}), \cdots, (u_{r,r}, v_{r,r})$)。この例では、全体の約 7.5%にあたる 2000 個の要素を選択した ($r = 2000$)。

以下では距離マップ D_k 内から r 個の要素を抜き出したものを r 次元距離値ベクトル $D'_{r,k}$ で表す ($D'_{r,k} = \begin{bmatrix} d_{k(u_{r,1},u_{v,1})} & \cdots & d_{k(u_{r,r},j_{r,r})} \end{bmatrix}'$)。

続いて、人物画像についての距離値ベクトル $D'^{obj}_{r,1}, \cdots, D'^{obj}_{r,K}$ および背景画像についての距離値ベクトル $D'^{bck}_{r,1}, \cdots, D'^{bck}_{r,K}$ の分布を r 次元正規分布で表現し、線形判別法により両者を判別する判別軸を求める。それぞれの距離値ベクトルの平均および共分散行列をそれぞれ $\bar{D'}^{obj}_r, \Sigma_{\bar{D'}^{obj}_r}$ (人物画像) および $\bar{D'}^{bck}_r, \Sigma_{\bar{D'}^{bck}_r}$ (背景画像) で表すとすると、r 次元での判別軸 A_r は次式によって得られる。

$$A_r = \left(\Sigma_{\bar{D'}^{obj}_r} + \Sigma_{\bar{D'}^{bck}_r}\right)^{-1} \left(\bar{D'}^{obj}_r - \bar{D'}^{bck}_r\right) \tag{10}$$

このようにして得られた判別軸から寄与の小さい要素を取り除き、判別軸を再度計算することを繰り返しながら順次次元数を削減する。

実験例では、(9) 式に基づいて得られた 2000 次元の分布から判別軸 A_{2000} 求め、寄与の小さい要素を順次取り除くことで、最終的に 10〜200 要素を選択した。次節で述べる認識部では、選択された要素に対応する判別軸により認識処理を行う。図 7 に 50 次元の場合について選択された要素を示す。(各要素は元の画像空間の 2 ブロックに相当するため、それらを中央に近いもの (右) と遠いもの (左) に分けて表示した。)

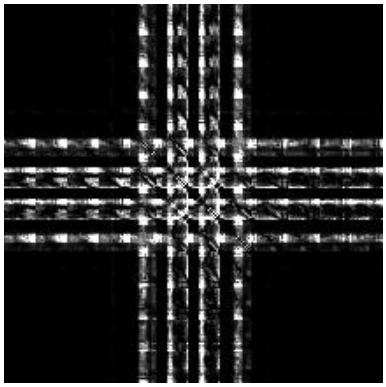


図 6: 平均距離マップ間の差

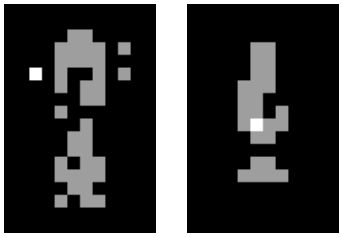


図 7: 選択された要素の例 (50 次元) (50 個の要素を元の画像空間に投影して表示。1 組のみハイライト表示。)

3.3 認識処理

認識処理においては、入力画像内の各部分について距離値ベクトルを計算し、前節で得た統計モデルと比較する。

計算コストを抑えるため、まず入力画像内の全ての可能なブロックについて画素値の平均ベクトル $\bar{\boldsymbol{x}}$ および共分散行列 Σ を計算する。その後、画像内の各部分について距離値ベクトル D'_r を計算する (図 8)。最後に、前節で求めた判別軸に基づいてスコアを計算し、次式のようにスコアが閾値を越えた場合に人物が検出されたと判断する。

$$A_r D'_r > \text{threshold} \tag{11}$$

同様の処理を解像度の異なる入力画像に対して行い、見かけの大きさの異なる人物への対応を図る。現在の実装では、4 段階の解像度 (元画像の 1,4/5,$(4/5)^2$,$(4/5)^3$ 倍) に対して認識処理を実行している (図 9)。

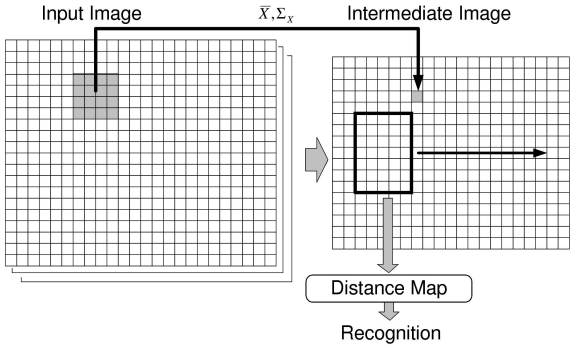


図 8: 認識処理の流れ

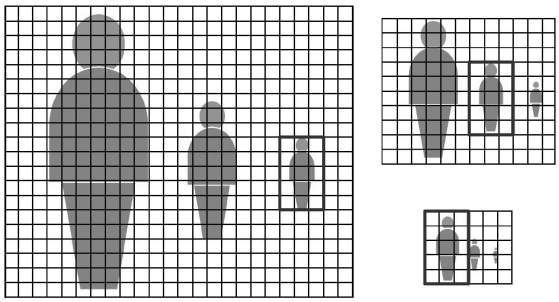


図 9: 多重解像度画像による認識

4 実験

人物画像、背景画像として各 4000 枚 (合計 8000 枚) の画像を用意して実験を行った。

まず、すでに述べたとおり半分の 4000 枚 (人物画像・背景画像各 2000 枚) の画像を使って統計モデルを構築し、残りの半分の画像によって本手法の認識率を評価した。

表 1 に認識結果を示す。ここで、次元数は認識に用いた距離マップ中の要素数を示す。ここでは 50 要素による正答率が最も高く 90%以上となったほか、10 要素のみによる実験でも 83%の正答率を得た。図 10 には各画像に対して式 (11) によって計算したスコアを示す。

図 11 にいくつかの画像に対する人物検出の例を

表 1: 実験結果 (認識率)

次元数	10	50	100	200
正答率 (%)	83	90	88	86

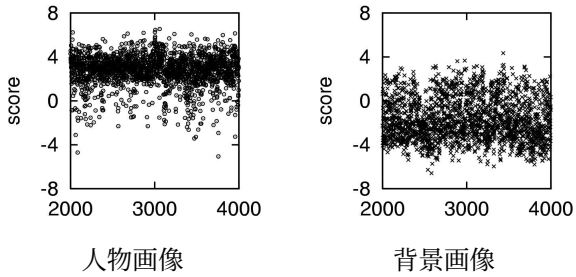


図 10: 実験結果 (スコア、50 次元)

示す。ここに見られるように、本手法により着衣の色や模様・体の向きの変化に関わらず人物を検出できた。また、図 12 に示すように、人物とカメラの距離によって人物像の大きさが変化した場合にも、解像度の異なる画像を同時に処理することである程度対応することができる。以上の結果から、環境条件にある程度変化があっても、提案手法により同一のモデルで人物検出が行えることが示されたといえる。

一方で、背景に置かれた物体が学習時と異なる場合に検出が極端に不安定になる場合があるなど、今回の実装では認識性能が学習データに依存する傾向が見られた。このことから、より多くの学習データの利用が必要と考えられるほか、次節で述べるように人の姿勢等に応じて複数のモデルを用意することも検討する必要がある。今後、これらの点を考慮しながら、より安定した検出結果が得られるよう認識手法の改善を進めていきたい。

図 11: 人物検出の例

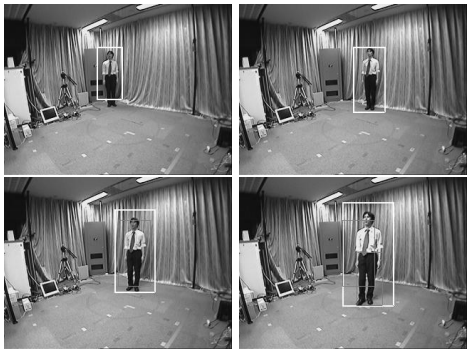


図 12: 撮影距離の変化への対応

5 考察

本報告では、画素値分布として観測される対象物体の幾何学的な構造 (形状) に基づいて物体を認識する統計的な方法を提案した。今回は、人物像に本手法を適用したが、本手法は人物像に限らず適用可能であると考えられる。本アルゴリズムは固定した背景を仮定しないため、手持ちカメラのようにカメラが移動する場合や背景や照明が動的に変化する場合にも動作することが期待できる。

現在の実装では、4 段階の解像度に対して認識処理を実行することで対象物体とカメラ間の距離の変化に対応している。これについては、最適な解像度の配分を含め今後さらに検討する必要がある。また、今回は全ての姿勢をひとつの統計モデルで表現しているが、認識精度を向上させるために対象物体の向き・姿勢ごとに異なるモデルを用意することも当然考えられる。計算コストの増加との関係を考慮しながら適応的なモデルを導入していく予定である。

また、提案手法では計算の簡便さからブロック間の画素値分布の近さの尺度としてマハラノビス汎距離を利用したが、マハラノビス汎距離には共分散のみが異なる場合を区別できない等の問題がある。今後他の距離尺度の利用についても検討するとともに、さらに距離マップに含まれる冗長性を利用してより効率的な認識を行う方法についても検討・実装を進めていきたいと考えている。

6 まとめ

画像内に現れる対象物体の構造 (形状) を画素値分布の変化を利用してモデル化し、画像内から対象物体を検出する統計的な手法を提案した。多数のサンプル画像から距離マップを計算し、人物像のモデルを構成した。また、実験により距離マップ中の少数の要素のみを用いて人物と背景を識別できることを示した。

今後、対象物の姿勢変化を効率的に表現できるモデルの構築、目的に応じた計算量の動的な制御等についてさらに検討していきたい。

参考文献

[1] Jakub Segen and Sarma Pingali. A camera-based system for tracking people in real time. In *Proc. of ICPR '96*, pages 63–67, 1996.

[2] Q. Cai, A. Mitiche, and J. K. Aggarwal. Tracking human motion in an indoor environment. In *Proc. of 2nd International Conference on Image Processing*, pages 215–218, 1995.

[3] Q. Cai and J. K. Aggarwal. Tracking human motion using multiple cameras. In *Proc. of ICPR '96*, pages 68–72, 1996.

[4] Akira Utsumi and Jun Ohya. Multiple-camera-based human tracking using non-synchronous observations. In *Proc. of ACCV2000*, pages 1034–1039, 2000.

[5] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *Proc. of CVPR '96*, pages 73–80, 1996.

[6] G. Yang and T.S. Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53–63, 1994.

[7] Mathew Turk and Alex Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[8] Hiroshi Murase and Shree K. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995.

[9] Henry Schneiderman and Takeo Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *Proc. of CVPR '98*, pages 45–51, 1998.

[10] E. Osuna and F. Girosi R. Freund. Training support vector machines: an application to face detection. In *Proc. of CVPR '97*, pages 55–60, 1997.

[11] Daniel P. Huttenlocher, Gregory A. Klanderman, and William J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Pattern Anal. Machine Intell.*, 15(9):850–863, 1993.