

修士論文

# 時間変化を考慮したHOG特徴量による 動作認識



平成 23 年度修了  
三重大学大学院 工学研究科  
博士前期課程 情報工学専攻

佐々木拓真

## 目次

はじめに	1
第1章 序論	2
1.1 研究の背景と目的	2
1.2 既存研究と用いられる特徴量について	2
第2章 特徴量と識別器について	4
2.1 HOG 特徴量	4
2.1.1 HOG とは	4
2.1.2 HOG の抽出手順	5
2.2 サポートベクターマシン	6
2.2.1 サポートベクターマシンとは	6
2.2.2 線形 SVM	6
2.2.3 非線形 SVM	7
2.2.4 多クラス SVM	8
第3章 提案手法	10
3.1 平均画像による特徴量抽出	10
3.2 画素のばらつき頻度を導入した手法	11
第4章 動作認識手法	13
4.1 認識手順	13
4.2 人物抽出手法	13
4.2.1 背景差分	13
4.2.2 ラベリング処理	14
4.2.3 画像正規化	15

第5章 実験	16
5.1 実験環境について	16
5.2 評価方法について	17
5.3 実験結果	19
5.3.1 実験1について	19
5.3.2 実験2について	20
5.3.3 総合結果	21
5.4 考察	21
5.5 実験のまとめ	22
おわりに	23
謝辞	24
付録	27
6.1 背景画素の出現数を考慮した手法	27
6.2 全画像列から抽出したHOGと多数決を用いた手法	29

# はじめに

動作認識とは、その名の通り動画像に映っている人や物体の動作をコンピュータに認識させるものである。動作認識は、ビデオサーベイランス、マーケティング、スポーツ解析、パーソンサーティフィケーションなど幅広い分野での実用化が期待されており、近年大きな注目を集めている。このような動作認識において、本研究では人の動作に着目し、動作の自動識別システムを目標とする。そのシステムの実現に向け、与えた動作に対する認識率の向上を目指し、研究に取り組んだ。

本研究の新規性として、動作の指標を表す特徴量に、従来とは異なる特徴量を用いた。さらに、その特徴量に動きの情報を取り入れることにより、人の動きを強く表した特徴量の取得を目指した。そして、独自に撮影した動画像を用いた動作実験により、提案する特徴量の有効性を吟味した。

本稿では、まず1章で研究の背景と目的を述べる。次の2章では、本研究で使用した特徴量と識別器について言及する。3章には本研究で用いる特徴量について提案し、4章では動作認識手法について述べる。5章で提案手法を用いた実験について記述する。



# 第1章

## 序論

### 1.1 研究の背景と目的

近年、カメラから得た動画像から、人や物体の動作をコンピュータに認識させる研究が盛んに行われている。高精度な動作認識は幅広い分野での応用が可能である。例えば、駅やコンビニに設置された防犯カメラ映像を用い、そこに映る人の異常行動をコンピュータに自動検知させることによって、不審者の早期発見や防犯に役立てることができる。また、スポーツ中継映像を用い、相手チームの戦略分析や、選手の自動フォーム解析に用いることができる。その他にも、客の動向からマーケティング調査を行うことや、歩行動作から個人認証を行うことなどが、動作認識の活用場面は多数存在する。このような動作認識において、本研究では人の動作に着目し、自動動作識別システムの構築を目標とする。自動動作識別システムの実装により、上にあげた幅広い場面での応用が可能となる。このシステムの実現に向け、まずは与えられた動作の認識率の向上を目指す。

### 1.2 既存研究と用いられる特徴量について

動作認識の研究は数多く行われているが、用いられる特徴量も様々である。

先行研究として、オプティカルフローを用いて動物体の領域を切りだし、その領域の特徴から動作の認識をするもの [1][2] や、動作の変化を表現した MHI, MEI という画像を用いて動作を表現し、テンプレートマッチングにより動作を認識するもの [3] が存在する。

また近年の動作認識では、文献 [4][5] など、特徴量に CHLAC を用いた動作認識も存在する。立体高次局所自己相関特徴 (Cubic Higher-order Local Auto-Correlation(CHLAC))[6] は、Kobayashi, Otsu らによって、2004 年に提案された特徴量である。顔検出等の物体検出に用いられている 2 次元画像を対象とする高次局所自己相関特徴に時間軸を加えることで 3 次元に拡張したものであり、動画像中に出現する動物体の「動き」の「形」を表現することができる特徴量である。位置不変性やロバスト性に長けていることを特徴とし、人の異常行動を検出する動作認識などでよく使用されている。

さらに、Space-Time Patch(ST-patch) と呼ばれる特徴量を用いた研究 [7][8] も存在する。ST-patch 特徴量 [9] は Shechtman らによって、発見された特徴量である。ST-patch 特徴量は物体の「アピアランス」と「モーション」の 2 つの情報を持つ。ST-patch 特徴量はテクスチャが異なる対

象物体の動きや、非剛体の物体のように複雑に動く物体に対しても動きの評価を行えるのが特徴である。St-patch 特徴量は、背景が動的な場合などでよく使用されている。

このように多種の特徴量が動作認識に有効とされているが、一概にどの特徴量が優秀であるとは判断できない。そこで本研究では、特徴量の 1 つである HOG 特徴量 [13] に着目し、その有効性を検討する。HOG 特徴量は、勾配方向と勾配強度を用いて算出される特徴量であり、局所的な幾何学変化や照明変化に頑健であると言われている。また HOG 特徴量とサポートベクターマシン (SVM) を組み合わせた人物抽出や車両検出は、高精度な性能を持つことが報告されている [10][11][12]。そこで、本研究も HOG 特徴量に SVM を組み合わせることによって、動作認識を行う。次の章では、その HOG 特徴量とサポートベクターマシンについて述べる。

## 第2章

# 特徴量と識別器について

## 2.1 HOG 特徴量

### 2.1.1 HOG とは

HOG(Histograms of Oriented Gradients)[13] とは, N. Dalal により 2005 年に発表された特徴量である。HOG は, 隣接する画素の勾配を局所領域毎にヒストグラム化し, 正規化を行うため, 局所的な幾何学変化や照明変化に頑健であると言われている。また, HOG と類似した特徴量として SIFT[14] があげられる。SIFT は特徴点に対して特徴量を記述するものであるが, それに対し, HOG ではある一定領域に対する特徴量の記述を行う。そのため, HOG は大まかな物体形状を表現することが可能であり, 人検出や車検出等の一般物体認識等に用いられている。

HOG はセルと呼ばれる局所領域において勾配方向ヒストグラムを作成することで特徴量を抽出する。局所領域セルの範囲は  $3 \times 3$  ピクセル,  $5 \times 5$  ピクセルなど研究によって様々であるが, 論文 [13] では  $6 \times 6$  の合計 36 ピクセルが最適とされている。本研究でも, 実験的検討を行った結果, 1 セルの大きさを  $6 \times 6$  ピクセルの領域が最適と判断する。また,  $3 \times 3$  セルの領域を 1 ブロックとして扱う。

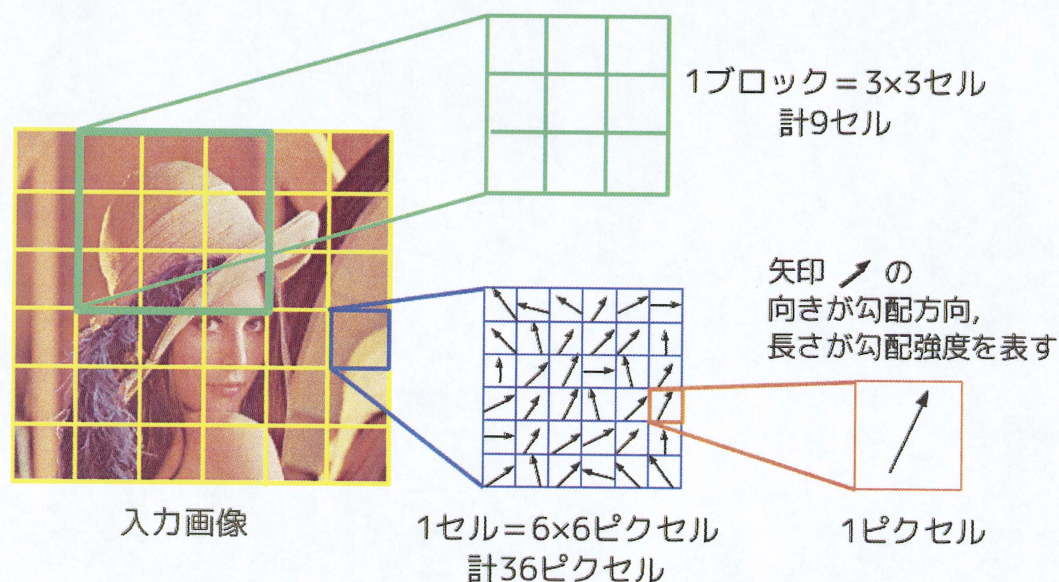


図1 HOG のセルとブロックについて



### 2.1.2 HOG の抽出手順

HOG の抽出手順について述べる．まず勾配方向ヒストグラムを作成するため，1 セル，36 ピクセルの領域における各ピクセルの輝度から，勾配強度  $m(x, y)$  と勾配方向  $\theta(x, y)$  を次式より算出する．

$$f_x(x, y) = L(x + 1, y) - L(x - 1, y) \quad (1)$$

$$f_y(x, y) = L(x, y + 1) - L(x, y - 1) \quad (2)$$

$$m(x, y) = \sqrt{f_x(x, y)^2 + f_y(x, y)^2} \quad (3)$$

$$\theta(x, y) = \tan^{-1} \frac{f_x(x, y)}{f_y(x, y)} \quad (4)$$

$L(x, y)$  は画像上の位置  $(x, y)$  における輝度値を表す．

この算出された勾配強度  $m(x, y)$  と勾配方向  $\theta(x, y)$  から 1 次元のヒストグラムを作成する．ヒストグラムの  $x$  軸に勾配方向を， $0^\circ$  から  $180^\circ$  の領域において  $20^\circ$  ずつに分割し，9 方向に配置する．ヒストグラムの  $y$  軸には勾配強度  $m$  を配置し，勾配方向の該当する強度が加算されたヒストグラムが作成される．（ここで算出された勾配方向は  $0^\circ$  から  $360^\circ$  の領域となるが，前景の明るさと隣接する背景領域の明るさの大小関係が逆転しても勾配方向が不変となるように  $0^\circ$  から  $180^\circ$  の領域に変換して用いる．）

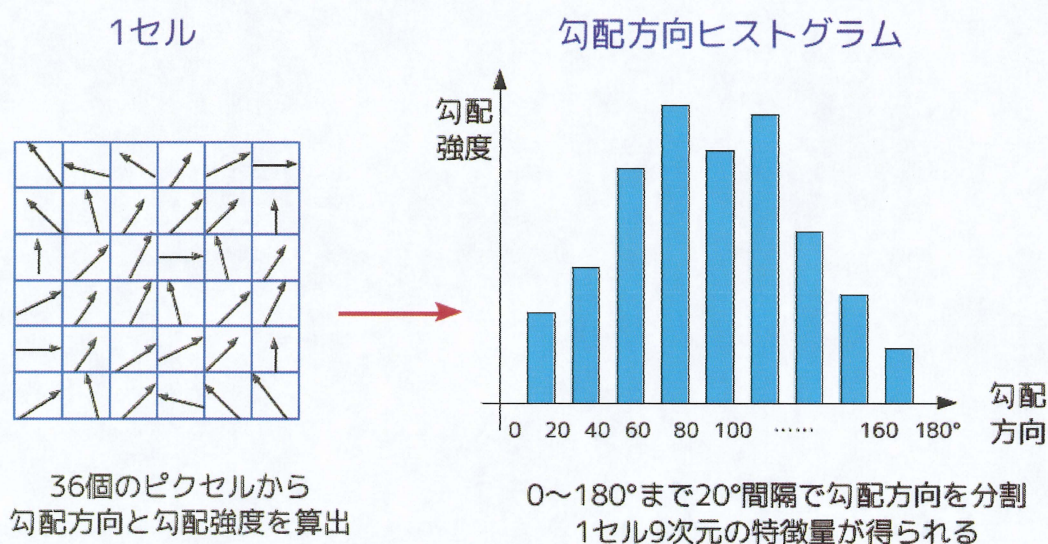


図2 ヒストグラムの作成

各セルで作成した輝度の勾配方向ヒストグラムを  $3 \times 3$  セル (1 ブロック) ごとに算出する．そして，ブロック単位で正規化を行う． $i$  行  $j$  列のセル  $(i, j)$  の特徴量 ( $0^\circ$  から  $180^\circ$  の

9 方向勾配なため 9 次元) を  $F_{ij} = [f_1, f_2, \dots, f_9]$  とすると,  $k$  番目のブロックの特徴量は  $V(k) = [F_{ij}, F_{i+1j}, \dots, F_{i+2j+2}]$  として 81 次元で表すことができる. あるブロックの  $i$  行  $j$  列のセルの特徴量を  $I(i, j)$  としたとき, 次式より正規化する.

$$I'(i, j) = \frac{I(i, j)}{\sqrt{\sum_{i=1}^3 \sum_{j=1}^3 I(i, j)^2 + \epsilon}} \quad (5)$$

ただし  $\epsilon$  は分母が 0 の場合に計算不能になるのを防ぐ係数であり, ここでは  $\epsilon = 1$  とする. 正規化は, ブロックを 1 セルずつ移動させることによって正規化を行う. そのため, 特徴量  $I(i, j)$  は異なるブロックの領域によって何度も正規化される. 入力画像を  $60 \times 120$  ピクセルとした場合, 横方向に 8 ブロック, 縦方向に 18 ブロック, 合計 144 ブロックに対して正規化を行う. 各ブロックごとに正規化された HOG 特徴量は,  $144 \text{ ブロック} \times 81 \text{ 次元} = 11664 \text{ 次元}$ となる.

## 2.2 サポートベクターマシン

### 2.2.1 サポートベクターマシンとは

サポートベクターマシン (SVM)[15] は 1960 年代に Vapnik らが考案した Optimal Separating Hyperplane を起源とし, 1990 年代になってカーネル学習法と組み合わせた非線形の識別手法へ拡張された手法である. カーネルトリックにより非線形の識別関数が構成できるように拡張された SVM は, 現在知られている手法の中で最もパターン認識性能の優秀な学習モデルの一つである.

SVM には様々なライブラリが存在する. その代表例として, LIBSVM[16], SVMlight[17] が考えられる. 本研究では HOG 特徴量を SVM に与えることによって, それがどのような動作を表したものを判定させため, SVM による多値分類問題を取り扱うこととなる. そのため, 本研究では, SVMmulticlass[18] と呼ばれる多クラス SVM に対応したライブラリを使用する.

以下で, 線形 SVM, 非線形 SVM の順に述べ, 最後に多クラス SVM について記述を行う.

### 2.2.2 線形 SVM

SVM は, 基本的に 2 クラスのパターン認識器を構成する手法であるため, 2 クラスのパターン認識器を構成するということは, 学習パターンに対して最適な識別境界を決定することと同じ意味合いを持つ. 以下の図 6 の (a) ような赤と青のサンプルを持つ 2 クラスのパターン分類問題を考える.



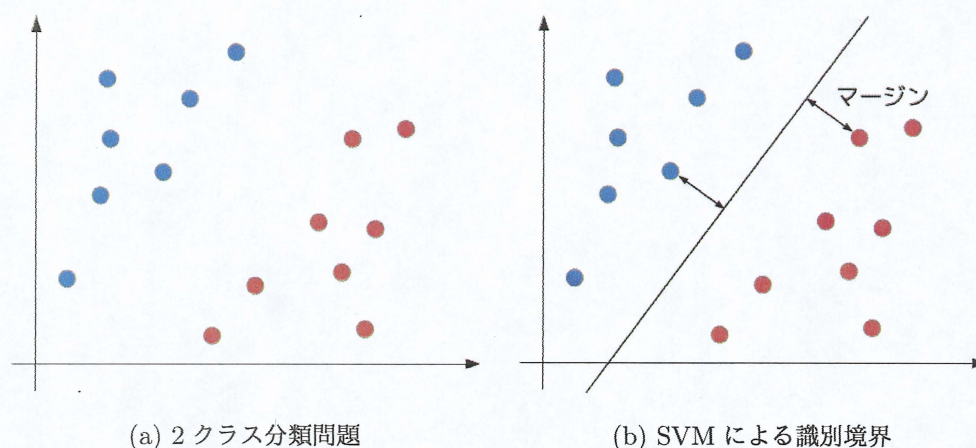


図3 SVM の考えかた

このような場合、赤と青のサンプル群の間にどのような識別境界を引くかが問題となる。SVM は図 6 の (b) のように、マージンと呼ばれるサンプル群から識別境界への距離が最も大きくなる位置に識別境界を引く。これによって、サンプルと実データの差を飲み込むため、認識性能が高くなる。このような実データに対する頑健性を汎化性能といい、SVM は最も優れた汎化性能を持つ。

### 2.2.3 非線形 SVM

SVM の基本的な構造は、図 6 に示すような線形しきい素子である。しかし、これでは線形分離不可能なデータに適用することができず、SVM の応用範囲は非常に限られたものになってしまう。データが完全に分離されない限り、マージンは負の値をもってしまうためである。そこで、SVM によって非線形な分類を可能にする方法として、高次元化が挙げられる。これは、非線形写像によって、元の入力データを高次元特徴空間に写像し、特徴空間において線形分離を行うカーネルトリックと呼ばれる手法である。そうすることによって、結果的に元の入力空間においては非線形な分類を行っていることになる。



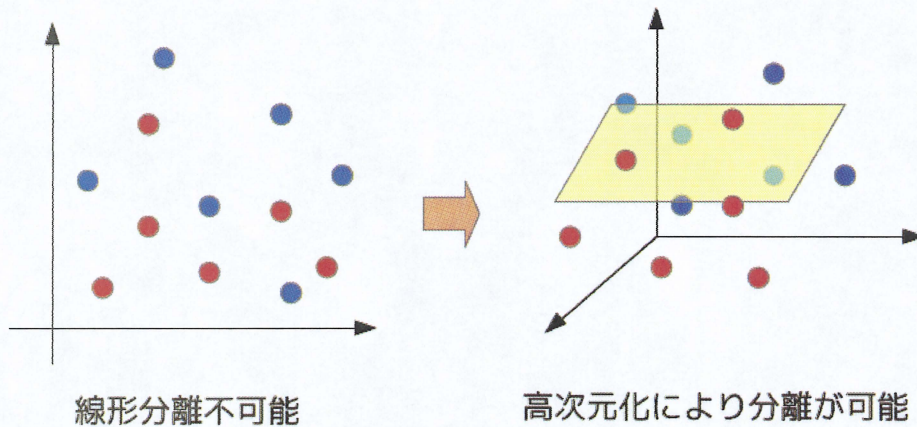


図4 SVM の非線形変換

非線形 SVM の識別関数  $f(x)$  は次のように定義される。

$$f(x) = \sin(g(x)) \quad (6)$$

$$g(x) = \sum_{i=1}^m w_i K(\tilde{x}_i, x) + b \quad (7)$$

$x$  は入力ベクトル,  $w_i$  および  $b$  は識別関数を決定するパラメータ,  $\tilde{x}_i$  はサポートベクター,  $K(x_1, x_2)$  はベクトル  $x_1, x_2$  を引数とする関数でカーネル関数と呼ばれる。カーネル関数は、通常次のような多項式型カーネル,

$$K(x_1, x_2) = (1 + x_1^t x_2)^p \quad (8)$$

またはガウシアン型カーネル,

$$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|}{2\sigma^2}\right) \quad (9)$$

が用いられる。 $p$  は多項式型カーネル関数の次数を決定するパラメータであり,  $\sigma$  はガウシアン型カーネル関数の拡張を決定するパラメータで、いずれもユーザが事前に値を決定する。実際には、標本の分布に対応したものを実験的に求めて設定することが多い。

## 2.2.4 多クラス SVM

多クラス SVM には、One Against All(OAA), One Against One(OAO), DAGSVMS などの手法が提案されてきた。本研究は其中最も良い精度を持つ OAO[19] を採用した。今クラス数  $k$ , 要素数  $l$  のトレーニングデータが与えられたとする。この入力を  $(x_1, y_1), \dots, (x_l, y_l)$ ,  $x_i \in R^n$ ,  $i = 1, \dots, l$ ,  $y_i \in 1, \dots, k$  と表すと、OAO は  $k(k-1)/2$  個の 2 クラス識別器を構成する。入力データがクラス  $i$  か  $j$  のどちらかに属するかを判断する 2 クラス識別器を構成するために、OAO

は以下の式を用いる.

$$\min_{w^{ij}, b^{ij}, \xi^{ij}} \frac{1}{2} (w^{ij})^T w^{ij} + C \sum_t \xi_t^{ij} \quad (10)$$

$$(w^{ij})^T \phi(x_t) + b^{ij} \geq 1 - \xi_t^{ij}, \quad (\text{if } y_t = i) \quad (11)$$

$$(w^{ij})^T \phi(x_t) + b^{ij} \leq -1 + \xi_t^{ij}, \quad (\text{if } y_t \neq j) \quad (12)$$

$$\xi_t^{ij} \geq 0, \quad (j = 1, \dots, l) \quad (13)$$

これらの式を全てのクラス組み合わせにおいて解くことにより,  $k(k-1)/2$  個の決定関数  $\sin((w^{ij})^T \phi(x) + b^{ij})$  が得られる. あるデータについてこの決定関数値を計算すると, そのデータがクラス  $i$  かクラス  $j$  のどちらに属するかが求められる.  $OAO$  は  $k(k-1)/2$  個全ての決定関数値を計算し, そのクラスに属すると判断された数を数え, 最も回数が多かったクラスであると予測を行う.



## 第3章

# 提案手法

2章で述べたように, HOG 特徴量は 1 枚の画像から抽出する特徴量であり, 動画像を扱う動作認識で HOG を用いるためには, フレーム画像 1 枚 1 枚から全て HOG を抽出する必要がある. しかしこの場合, 膨大な計算コストがかかるという問題点があり, さらに静止画一枚で動作を識別させることが困難であることは明白である. そこで, 本研究では, HOG 特徴量を抽出する際に工夫を加え, 時間的変化を考慮した HOG 特徴量の抽出方法を提案する.

### 3.1 平均画像による特徴量抽出

まず, HOG 特徴量を抽出する際, フレーム画像 1 枚を用いるのではなく, 複数枚のフレーム画像から作成した平均画像から HOG を抽出することとする. 平均画像は以下の式 (14) で計算される画素  $\bar{L}(x, y)$  を各画素の要素としている画像である.

$$\bar{L}(x, y) = \frac{1}{N} \sum_{t=1}^N L_t(x, y) \quad (14)$$

$L_t(x, y)$  は  $t$  枚目フレームにおける位置  $(x, y)$  の画素値を意味し,  $N$  は動画像内のフレーム枚数を意味する. 上の式によって, 時系列上に並んだ画素の平均が得られ, 大まかな動きの体型を表した画像を作成することができる. そのため, 人の「走る」動作などの激しい動作は, 画素の変動が大きくなり, 平均画像は画像全体に薄く画素が散らばりやすくなる. また, 人の「歩く」などの動きの少ない動作においては, 足などの局所的な部分に画素が散らばりやすく, 変動の少ない上半身部分などは人の形状をはっきりと残した平均画像が生成される. このように, この平均画像を用いることによって, 動きの変動の差が画像に現れやすく, おおよその動きの識別が可能となる. 5章で述べる実験において, 実験対象とした 4 動作による平均画像は以下の図 5 のようになる.

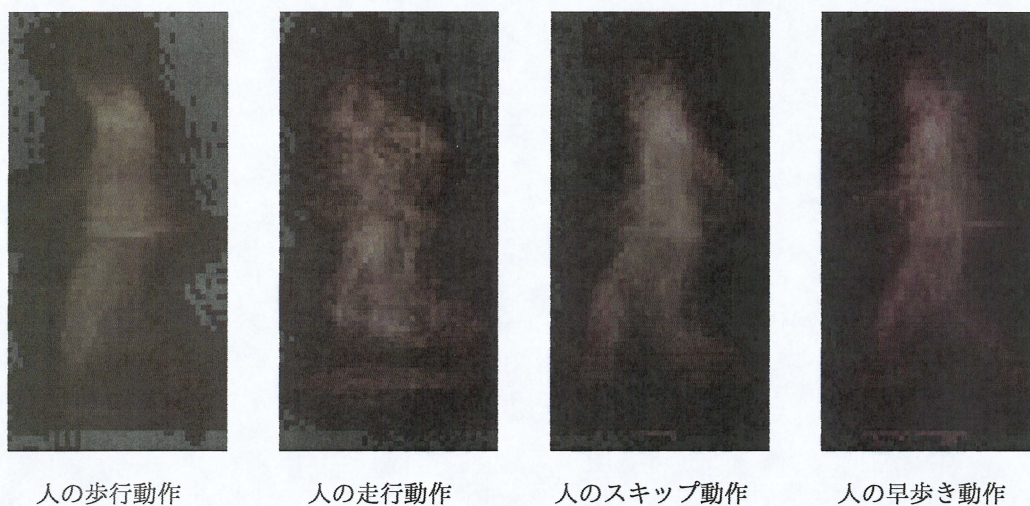


図5 実験で使った平均画像例

### 3.2 画素のばらつき頻度を導入した手法

本研究では、上記の平均画像に加え、さらに動きの指標を HOG に加える。画素値の平均  $\bar{L}(x, y)$  を算出する際、その平均画素を用い、 $\sigma(x, y)^2$  を以下の式 (15) のように求める。

$$\sigma(x, y)^2 = \frac{1}{N} \sum_{t=1}^N (L_t(x, y) - \bar{L}(x, y))^2 \quad (15)$$

$\sigma(x, y)^2$  は、画素値の時間軸上での変化のばらつき頻度を表す分散である。 $\sigma(x, y)^2$  の値が大きいことは、画素  $(x, y)$  における画素の変動が激しいことを意味する。したがって、 $\sigma(x, y)^2$  が動作の変動の激しさを表す指標になることが期待できる。

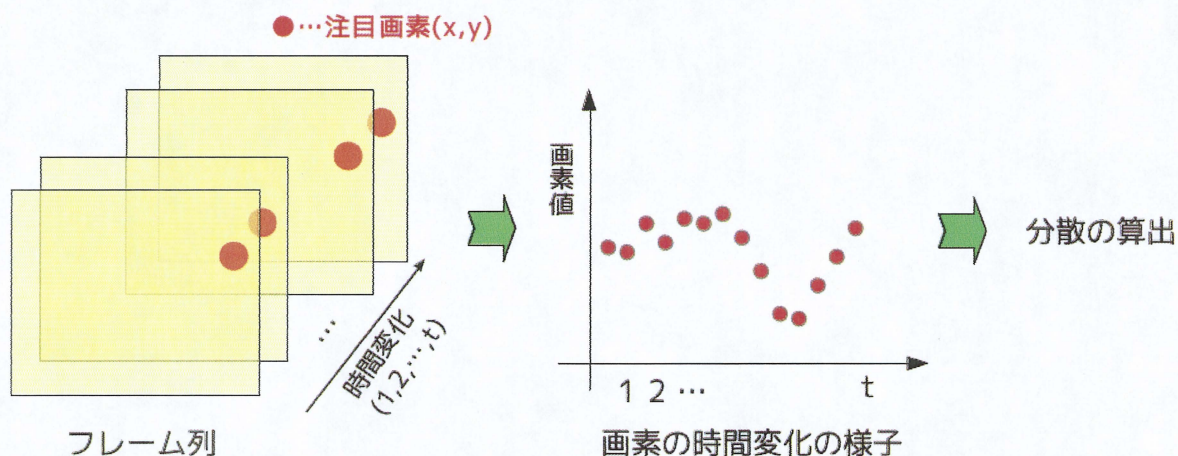


図6 分散  $\sigma(x, y)^2$  の抽出手順

この  $\sigma(x, y)^2$  を用い, HOG 特徴量の拡張を行う. HOG の勾配強度算出式 (3) を以下の式 (16) ように拡張する.

$$\hat{m}(x, y) = \sqrt{f_x(x, y)^2 + f_y(x, y)^2 + k\sigma(x, y)^2} \quad (16)$$

定数  $k$  の値は  $\sigma(x, y)^2$  をどれだけ勾配強度に加算するかを表す重みである.  $k$  は実験的に検討されるべきである. この拡張により, 動きが激しい画素ほど勾配強度  $m(x, y)$  が大きくなり, 動きの情報をより取り入れることが可能となる. また, 勾配方向  $\theta$  をあえて拡張させないことによって, 次元数を増加を防いでいる.

人の歩行動作と早歩き動作などの類似動作では平均画像によるシルエットでは, 差が現れにくい. しかし, この  $\sigma(x, y)^2$  を用いることによって, 局所的な動きの変化の差が現れ, 歩行動作と早歩き動作などの類似シルエット動作においても識別が可能になることが期待できる.

## 第4章

# 動作認識手法

### 4.1 認識手順

本研究の動作認識手順は以下のようになる。

1. 1つの動画像を読み込み、フレームごとに人物領域を抽出する。
2. 人物領域が抽出されたフレーム列から画素値の平均を求め、平均画像を作成する。
3. 平均画像から時間的变化を考慮した HOG 特徴量を抽出する。
4. 算出した HOG 特徴量をあらかじめ学習させておいた SVM に識別させる。
5. 手順 1 に戻り、全ての動画像においてこの処理を繰り返す。

この認識手順において、HOG 特徴量抽出手法や SVM についてはすでに述べているため、以下では人物抽出について詳しく述べる。

### 4.2 人物抽出手法

入力された動画像において、1枚1枚のフレーム画像からそれぞれ人物抽出を行う。人物抽出は以下の手順で行う。

1. 動画像から1枚のフレームを読み込み、背景差分により、人物領域を大まかに抽出する。
2. 抽出された人物領域にラベリング処理を施し、人物領域を確定させる。
3. 人物領域を矩形で切り取り、正規化する。
4. 手順 1 に戻り、全てのフレームにおいてこの処理を繰り返す。

以下に背景差分、ノイズ処理、画像正規化について順に述べる。

#### 4.2.1 背景差分

背景差分には文献 [20] で報告されている背景画像の時間的な変化を考慮した動的背景更新によるロバストな注目物体の検出手法を用いた。これによって、対象フレームから背景領域が差分され、対象フレームの背景が黒く塗りつぶされる画像が得られる。以下にその例を示す。



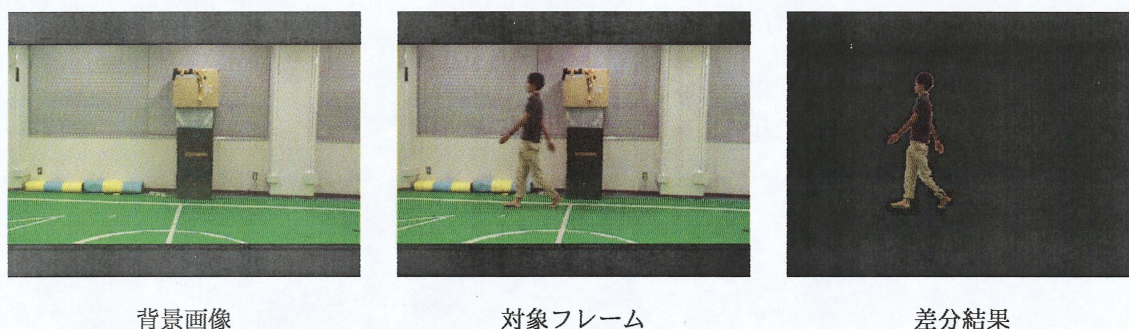


図 7 背景差分の適用例

#### 4.2.2 ラベリング処理

背景差分で抽出された領域には多くのノイズや、背景領域が残されている。それらの部分を除去するためにラベリング処理を施す。

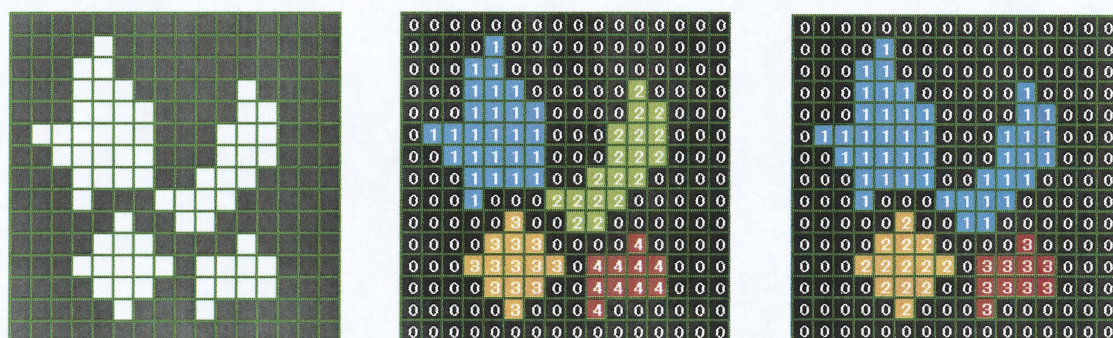
まず背景差分後の画像をグレースケール化し、さらに二値化処理を行う。それによって、背景差分で抽出された領域が白、背景領域が黒で統一される。



図 8 ラベリング前処理

この二値画像にラベリング処理を適用する。ラベリング処理は、二値化画像処理された画像において、白の部分（または黒の部分）が連続した画素に同じ番号を割り振る処理を指す。ラベリング処理には、縦、横方向に連続している部分を同じラベルにする 4 近傍を見る手法と、縦、横、斜め方向に連続している部分を同じラベルにする 8 近傍を見る手法の 2 種類がある。本手法では、白画素に着目し、8 近傍による処理を適用させ、ラベル付けを行う。そして一番面積の大きい白の領域を人物領域と確定される。この人物領域の高さ、幅がある値を満たしていない場合、そのフレームには人物が写っていないと判定する。





入力した二値画像

4 近傍によるラベリング結果

8 近傍によるラベリング結果

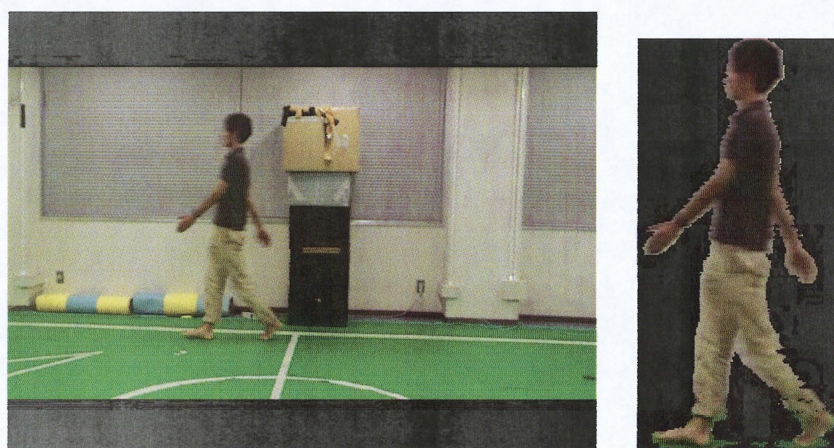
図 9 入力画像に対するラベリング結果 (参照画像 [21])

### 4.2.3 画像正規化

HOG 特徴量抽出時に画像サイズを統一させる必要があるため、人物抽出を行った時点で、画像の幅と高さの比率を全フレームで統一される。比率は画像を目視で確認した結果、幅：高さを 1：2 が適当とし、固定することとした。

画像正規化の手順として、まずラベリング処理による一番面積の大きな領域の重心を算出する。次にその重心を中心としたラベル領域を全て囲む幅：高さが 1：2 の矩形を画像に描く。このとき矩形が全て画像内に収まっている場合、人物が画像に映っていると判定し、画像から矩形領域を切り出し、人物領域の抽出を完了させる。矩形が画像からはみ出しているときは、人物がまだ画像に現れていないと判定し、切り出しは行わない。

以上の処理により、以下の画像のように人物領域が抽出される。



対象フレーム

人物抽出画像

図 10 人物領域の抽出結果



## 第5章 実験

時間的变化を考慮した HOG 特徴量と SVM を組み合わせた動作認識の有効性を実験によって検証する。

### 5.1 実験環境について

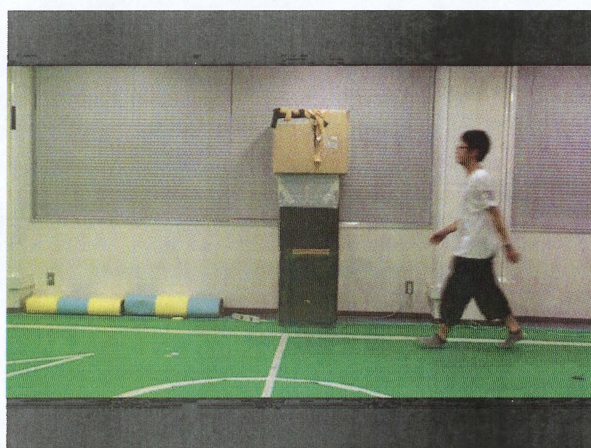
4 種類の動作識別実験を行う。対象とした動作は人の、「歩行」「走行」「スキップ」「早歩き」の 4 動作である。これらの動作について、独自に撮影し、得たデータを実験に用いる。撮影に用いたカメラは SONY 製の HDR-CX12 で、三脚を用い、カメラを固定することによって撮影を行った。



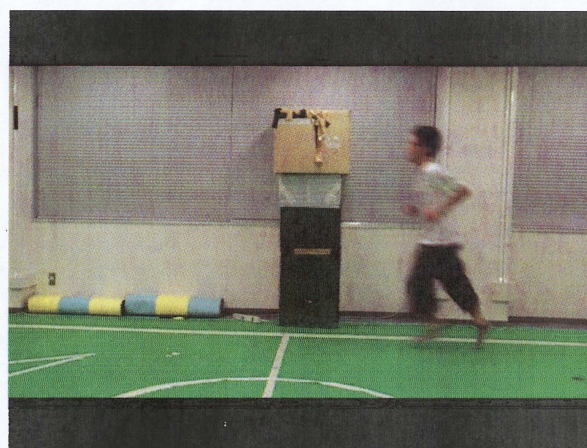
図 11 撮影に用いたカメラ

撮影は、本研究室のメンバーの 9 人を対象とし、4 種類の動作を行っている場面を撮影した。動作はどれも直線的に行ってもらい、その様子を設置したカメラで真横から撮影した。また各動作につき、1 人 5 回撮影し、9 人  $\times$  4 動作  $\times$  5 回 = 180 個の動画を得た。動画像のフレームレートは 30fps、画像サイズは 640 $\times$ 480 とする。なお、画像サイズは人物抽出時に幅と高さの割合が 1:2 の大きさに正規化され、その後 HOG 特徴量抽出時に 42 $\times$ 84 のサイズに修正される。この画像サイズは実験的に検討を行い、最適な値を定めたものである。

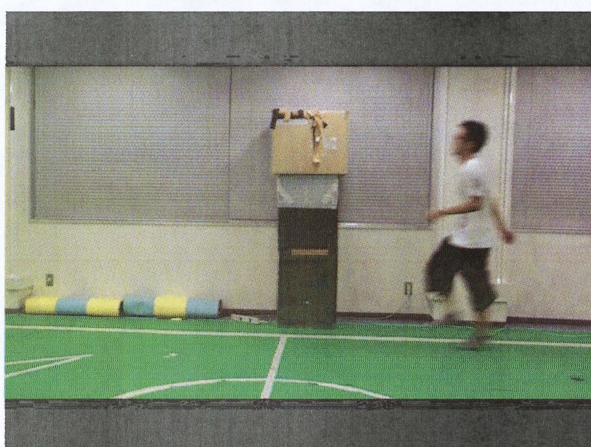




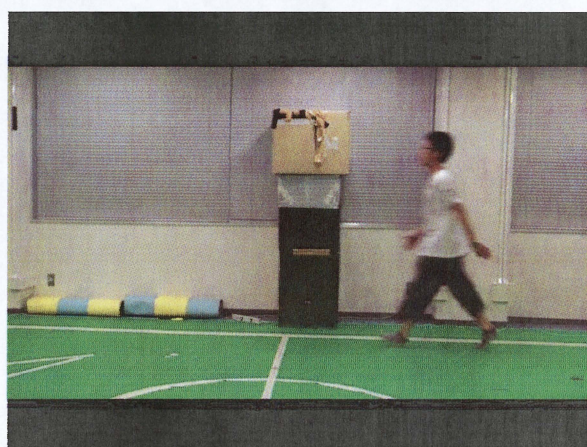
歩行



走行



スキップ



早歩き

図 12 撮影で得た動画像の 1 フレーム

また実験に使用したコンピュータは DELL 製の vostro 220s である。CPU は Intel Core2 Duo 3.0GH, OS は Vine Linux 5.0, メモリは 4GB のものを使用した。

## 5.2 評価方法について

あらかじめ学習させる動画像列のデータを定めておき、その動画像列から抽出した特徴量を正解動作とともに SVM に学習させる。そして、学習では使用していない評価用データから抽出した特徴量を学習させた SVM に与え、その動作が何の動作を表しているかを SVM に自動識別させる。それが与えた評価用データのうち何 % 正解しているかによって認識率を算出する。

また今回は実験 1, 実験 2 の 2 種類の動作識別実験を行う。実験 1 は、学習した被験者に対する実験であり、実験 2 は未学習の被験者に対する実験である。



まず、実験1では、全4動作を9人が5回ずつ行っている場面を撮影したデータのうち、被験者9人の2回分のデータを学習に用いる。そして、残りの3回分のデータを評価用に用いる。つまり学習に  $9 \times 4 \times 2 = 72$  個分の動画データを用い、残りの  $9 \times 4 \times 3 = 108$  個の動画データを評価用に用いることとなる。これにより識別に成功したデータ数を観測する(最大108個)。また、学習・評価に用いたデータをそのまま入れ替え、学習に108個、評価用に残りの72個を用いた場合の正解数(最大72個)も観測する。これらの合計2回分の正解数を合計し、全体の何%認識に成功しているかを計測し、これを認識率とする。この学習と評価に用いるデータをそのまま入れ替えて2回実験を行う手法はジャックナイフ法と呼ばれる。ジャックナイフ法は計算量が少ないことを長所とした1種の評価法である。

実験2では、全4動作を被験者9人分撮影したデータのうち、被験者4人分のデータを学習に用いる。そして、残りの5人分のデータを評価に用いる。つまり学習に  $4 \times 4 \times 5 = 80$  個分の動画データを用い、残りの  $5 \times 4 \times 5 = 100$  個の動画データを評価用に用いることとなる。実験2でも、実験1同様にジャックナイフ法を適用させ、学習と評価に用いるデータをそのまま入れ替え、計2回分の正解数から認識率を算出する。

実験1、実験2ともに、平均画像のみを用いた場合、分散のみを用いた場合、平均画像と分散をともに用いた場合の3手法でそれぞれ認識率を求め、比較する。なお、分散のみを使用した場合には、HOG特徴量は抽出しておらず、画素ごとの変化量の分散だけをそのまま特徴量として使用した場合を指す。また、平均画像と分散をともに用いた場合の実験において、算出した分散に対する重み $k$ は $k=1$ としている。

## 5.3 実験結果

実験 1, 実験 2 についてそれぞれ 3 手法で認識率を算出し, 比較を行う。

### 5.3.1 実験 1 について

実験 1 による結果は, 3 手法それぞれで以下の表 1, 2, 3 のようになった。

表 1 平均画像のみを用いた実験結果

与えた動作	与えた動作に対する判定 (動画像数)				認識率 (%)
	歩行と認識	走行と認識	スキップと認識	早歩きと認識	
歩行	43 個	0 個	0 個	2 個	95.6%
走行	0 個	45 個	0 個	0 個	100.0%
スキップ	0 個	0 個	44 個	1 個	100.0%
早歩き	3 個	0 個	0 個	42 個	93.3%

表 2 分散のみを用いた実験結果

与えた動作	与えた動作に対する判定 (動画像数)				認識率 (%)
	歩行と認識	走行と認識	スキップと認識	早歩きと認識	
歩行	44 個	0 個	0 個	1 個	97.8%
走行	0 個	45 個	0 個	0 個	100.0%
スキップ	1 個	0 個	41 個	3 個	91.1%
早歩き	4 個	0 個	1 個	40 個	88.9%

表 3 平均画像と分散を用いた実験結果

与えた動作	与えた動作に対する判定 (動画像数)				認識率 (%)
	歩行と認識	走行と認識	スキップと認識	早歩きと認識	
歩行	43 個	0 個	0 個	2 個	95.6%
走行	0 個	44 個	0 個	1 個	97.8%
スキップ	0 個	0 個	43 個	2 個	95.6%
早歩き	4 個	0 個	2 個	39 個	86.7%

この結果の表において, 一番左の列は, SVM に入力した評価用データの動作を示している。その動作において, SVM が識別を行った結果の動画数をその行に記述している。1 つの動作に対する動画像数は, 9 人  $\times$  5 回 = 45 個となっており, 識別結果が 45 個正解している場合が最も認識率が良い場合となる。例えば, 表 1 において, 入力した動作が歩行動作であるとき, 歩行であると判

定された動画数は 43 個、スキップと判定された動画は 2 個となっている。この場合の認識率は  $43 \div 45 \times 100 = 95.6\%$  となる。

実験 1 において、全体の認識率は、平均画像のみを用いた場合 96.7%、分散のみを用いた場合 94.4%、平均画像と分散をともに用いた場合 93.9% という結果になった。

### 5.3.2 実験 2 について

以下の表 4, 5, 6 のようになった。

表 4 平均画像のみを用いた実験結果

与えた動作	与えた動作に対する判定 (動画画像数)				認識率 (%)
	歩行と認識	走行と認識	スキップと認識	早歩きと認識	
歩行	37 個	0 個	1 個	7 個	82.2%
走行	0 個	43 個	2 個	0 個	95.6%
スキップ	5 個	2 個	37 個	1 個	82.2%
早歩き	8 個	1 個	0 個	36 個	80.0%

表 5 分散のみを用いた実験結果

与えた動作	与えた動作に対する判定 (動画画像数)				認識率 (%)
	歩行と認識	走行と認識	スキップと認識	早歩きと認識	
歩行	32 個	0 個	2 個	11 個	71.1%
走行	0 個	41 個	4 個	0 個	91.1%
スキップ	0 個	5 個	33 個	7 個	73.3%
早歩き	4 個	1 個	7 個	33 個	73.3%

表 6 平均画像と分散を用いた実験結果

与えた動作	与えた動作に対する判定 (動画画像数)				認識率 (%)
	歩行と認識	走行と認識	スキップと認識	早歩きと認識	
歩行	38 個	0 個	1 個	6 個	84.4%
走行	0 個	44 個	1 個	0 個	97.8%
スキップ	2 個	2 個	39 個	2 個	86.7%
早歩き	1 個	3 個	2 個	39 個	86.7%

全体の認識率は、平均画像のみを用いた場合 85.0%、分散のみを用いた場合 77.2%、平均画像と分散をともに用いた場合 88.9% という結果になった。

### 5.3.3 総合結果

実験 1, 実験 2 について 3 つの手法についてまとめた結果を以下の表 7 に示す。

表 7 実験の総合結果

実験	使用した 特徴量	認識率				
		歩行	走行	スキップ	早歩き	全体
実験 1	平均画像のみ	95.6%	100.0%	100.0%	93.3%	96.7%
	分散のみ	97.8%	100.0%	91.1%	88.9%	94.4%
	平均画像 + 分散	95.6%	97.8%	95.6%	86.7%	93.9%
実験 2	平均画像のみ	82.2%	95.6%	82.2%	80.0%	85.0%
	分散のみ	71.1%	91.1%	73.3%	73.3%	77.2%
	平均画像 + 分散	84.4%	97.8%	86.7%	86.7%	88.9%

## 5.4 考察

表 7 より、どの手法でも比較的高い認識率を達成していることがわかる。動作別に結果を見てみると、「走行」に対する認識率が高いことがわかる。これは走行動作が他の動作に比べて、体の動きが大きな変動をするため、平均画像に差が現れやすいためではないかと考えられる。他の 3 動作については、認識率はとても似通った形となっている。これは、それぞれの動作が平均画像に差が現れにくく、識別の難しさがわかる。

また実験ごとに結果を見てみると、実験 1 は、実験 2 に比べて認識率が高い。実験 1 は学習済みの人物に対する実験であるため、平均画像に学習したものと類似したデータが評価にも与えられ、容易に識別できたのではないかと考えられる。実験 1 において、3 手法は全て高い認識率を達成しているが、平均画像だけを用いた場合の認識率が最も良いものとなっている。分散のみを用いた場合でも 94.4% という高い認識率を達成してため、これを平均画像に組み合わせた平均画像 + 分散の手法が良い認識率を達成できそうだが、結果に影響がなかった。問題点がどこにあるかや、分散との組み合わせ方などをさらに考慮する必要がある。

実験 2 においては、平均画像に分散を加えた手法が最も高い認識率 88.9% を達成した。これは、うまく動きの情報を HOG に加えられたためではないかと考えられる。しかし、分散のみを用いた手法は、実験 1 に比べて認識率が大きく低下してしまっている。これは、HOG 特徴量を使用していない分散のみでは、人の体型を表現できていない可能性が考えられる。

## 5.5 実験のまとめ

時間的变化を考慮した HOG 特徴量を使用した手法により、学習済みの被験者に対する認識率が 96.7%，未学習の被験者に対する認識率 88.9% と高い認識率を達成することができた。この認識率について既存研究と結果を比較する、文献 [5] で行われている異常行動の研究では、正常動作とし歩行動作を学習させ、歩行、スキップ、早歩きの 3 動作を正常か異常動作かの識別をしている。認識率は、学習済みの被験者に対して約 98%，未学習の被験者に対して約 85% の認識率を達成している。本研究とは、実験内容や評価方法も異なるが、4 種類の動作での実験で同等の認識率を達成できているため、有効性を主張できるのではないかと考える。

今後は、実験 1、実験 2 とともに平均画像だけを用いた場合より良い認識率を達成できる動作の指標を目指す。

## おわりに

本研究は、HOG 特徴量と SVM を組み合わせ、新しい動作認識手法を考案した。また、動作認識に有効な動きによる画素の変化に着目し、それを HOG 特徴量に取り込むことによって、有効性を検証した。実験では、4 種類の類似動作に対して、学習済みの被験者に対する認識率が 96.7%、未学習の被験者に対する認識率 88.9% と高い認識率を達成することができた。

本研究には以下のような課題点も考えられる。

まず、既存研究との比較の問題である。本稿で HOG 以外の特徴量を使用した動作認識の研究が多く存在することを示したが、それらの手法を本実験データに適用し、認識率を算出する必要がある。それによって算出された認識率と本手法による認識率の比較を行うことによって、本研究の本当の有効性が得られると考える。

次に、実験量の問題である。本研究では特徴量の拡張方法や SVM への特徴の与え方を重視したため、ジャックナイフ法のための評価しか行っていない。評価の方法として、leave-one-out 法などの別の評価方法を採用し実験を繰り返す必要がある。実験で使用した変数についても検討が必要である。画素の変化のばらつきを表す  $\sigma^2$  に掛ける  $k$  の値に変動値を用いる場合や、さらに細かく倍率を変動させると今より良い認識率が達成できる可能性がある。使用したフレーム枚数についても、最適な値を実験を重ねることによって、求めるべきであると考え。

それらを検討後、動作の種類や被験者数を増やし、さらに高性能な自動動作識別システムを目指す。

## 謝辞

日ごろから多くの御指導を頂きました鈴木秀智准教授，太田義勝教授に深く感謝いたします。そして，日頃何かとお世話になりました落合美子事務員に感謝いたします。また，本論文作成にあたって特にお世話になりました鈴木秀智准教授に深く感謝いたします。最後に，日頃から熱心に討論して頂いた研究室の諸氏に感謝いたします。

## 参考文献

- [1] S. X. Ju and M. J. Black and Y. Yacoob, "Cardboardpeople: A parameterized model of articulated motion", International Conference on Automatic Face and Gesture Recognition, 38—44, 1996.
- [2] Yaser Yacoob and Michael J. Black, "Parameterized Modeling and Recognition of Activities", Computer Vision and Image Understanding, 73, 2, 232-247, 1999.
- [3] Romer Rosales, "Recognition of Human Action Using Moment-Based Features," 1998-020, 1998.
- [4] 南里卓也, "複数人動画像からの異常動作検出", 情報処理学会論文誌. コンピュータビジョンとイメージメディア 46(SIG\_15(CVIM\_12)), 43-50, 2005-10-15.
- [5] 児玉吉晃, "ベクトル長を考慮した相互部分空間法に基づく動作認識", Technical report of IEICE. PRMU 109(182), 7-12, 2009-08-24.
- [6] Takumi Kobayashi, Nobuyuki Otsua, "Action and Simultaneous Multiple-Person Identification Using Cubic Higher-order Local Auto-Correlation", International Conference on Pattern Recognition (ICPR' 04).
- [7] 吉川拓弥, "First Person Vision のための ST-patch 特徴を用いた自己動作識別", 電子情報通信学会 パターン認識・メディア理解研究会 (PRMU), pp. 53-58, 2010.
- [8] 村井泰裕, "Space-Time Patch を用いた物体の移動方向識別とセグメンテーション", 情報処理学会論文誌 CVIM, Vol. 1, No. 2, pp. 21-31, 2008.
- [9] Eli Shechtman and M. Irani, "Space-Time Behavior Based Correlation" Computer Vision and Pattern Recognition, vol.1, pp. 405-412, 2005.
- [10] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol 2, pp. 524 - 531, 2005.
- [11] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection", Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 886-893, 2005.
- [12] F. Han, Y. Shan, R. Cekander, "A Two-Stage Approach to People and Vehicle Detection with HOG-Based SVM", PerMIS, pp. 133-140, 2006.
- [13] N. Dalal et al, "Histograms of Oriented Gradients for Human Detection", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR' 05).
- [14] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International



- Journal of Computer Vision, Vol.60 No.2 pp.91—110, 2004.
- [15] V.N.Vapnik, "Statistical Learning Theory", John Wiley & Sons, 1999.
  - [16] LIBSVM, "<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>".
  - [17] SVMlight, "<http://svmlight.joachims.org/>".
  - [18] SVMmulticlass, "[http://svmlight.joachims.org/svm\\_multiclass.html](http://svmlight.joachims.org/svm_multiclass.html)".
  - [19] Chih-Wei Hsu, Chih-Jen Lin, "A Comparison of Methods for Multi-class Support Vector Machines", Department of Computer Science and Informaiton Engineering National Taiwan University Taipei 106, Taiwan.
  - [20] 森田 真司, 山澤 一誠, 寺沢 征彦, 横矢 直和: "全方位画像センサを用いたネットワーク対応型遠隔監視システム", 電子情報通信学会論文誌 (D-II), Vol. J88-D-II, No. 5, pp. 864-875, (2005.5).
  - [21] 参照画像, "<http://imaging-solution.blog107.fc2.com/blog-entry-193.html>".

## 付録

本稿に記載した提案手法は、平均画像と分散  $\sigma$  を用いた手法であるが、他にも多くの手法を試行している。この付録にはそのような今後発展する可能性がある試行した手法を記述する。

### 6.1 背景画素の出現数を考慮した手法

画素の時間上での変化のばらつき頻度  $\sigma^2$  を利用した手法とは別の、背景画素数を考慮した手法を合わせて提案する。この手法は、画像上のある注目画素が背景画素であるかを、時間軸上において計測し、利用する手法である。背景画素になる頻度が高い画素は動作による体の動きによって、背景画素になっている可能性が高い。つまり、そのような画素における特徴量を強く抽出することによって、より動きの情報を考慮した特徴量になることが期待できる。ある画素  $(x, y)$  における背景画素の計測回数を  $n(x, y)$  とし、 $n(x, y)$  の値が以下の式、

$$M_{in} < n(x, y) < M_{ax} \quad (17)$$

を満たした場合、勾配強度  $m$  に重み  $k$  を掛ける。

$$m = k \sqrt{f_x(x, y)^2 + f_y(x, y)^2} \quad (18)$$

$k$  は実験により  $k = 2$  が適当であると判断した。また、 $M_{in}$ 、 $M_{ax}$  は  $n(x, y)$  の範囲を表す定数である。フレーム数が  $t$  枚の動画像列における、ある点  $(x, y)$  の画素値に背景画素が出現する回数は最多で  $t$  回であり、最少で 0 回となる。 $n(x, y) = t$  のとき、画素  $(x, y)$  はずっと変化のない背景であり、 $n(x, y) = 0$  のとき、画素  $(x, y)$  は人物領域の内側である可能性が高いと言える。従って、 $n(x, y)$  が  $\frac{t}{2}$  前後であるとき、最も背景画素になる頻度が高いと推測できる。 $\frac{t}{2}$  の前後で  $M_{in}$ 、 $M_{ax}$  の値を調整し、最適な値を検討する。

この手法を用いて実験を行う。実験環境については 5 章で述べたものと同じで、実験 2 のみ検証を行った。結果は以下の表 8 ようになった。

表 8 出現頻度を変更した実験 (実験 2)

$M_{in} \sim M_{ax}$	認識率 (%)				
	歩く	走る	スキップ	早歩き	全体
10~11	82.2%	95.6%	84.4%	93.3%	88.9%
10~12	82.2%	97.8%	93.3%	100.0%	93.3%
10~13	82.2%	97.8%	82.2%	93.3%	88.9%
10~14	82.2%	95.6%	82.2%	91.1%	87.8%
9~11	82.2%	95.6%	86.7%	86.7%	87.8%
9~12	84.4%	95.6%	82.2%	88.9%	87.8%

表 8 が示す 10~11 などの範囲を表す値は,  $M_{in} = 10$ ,  $M_{ax} = 11$ であることを示しており, 背景画素の出現回数  $n(x, y)$  が  $M_{in} \sim M_{ax}$  の範囲にあれば, 勾配強度  $m$  を  $k$  倍 (2 倍) する手法を意味している.

結果を見ると, 全体的に高い認識率を達成できていることがわかる. また,  $M_{in} \sim M_{ax}$  の範囲が 10~12 のときには, 全体の認識率 93.3% を達成している. これは, 未学習の人に対する実験として, 最も高い認識率を達成することができた. そこで,  $M_{in} \sim M_{ax}$  の範囲が 10~12 のときに限定し, 実験 1 も行ってみたところ, 以下の表 9 のような認識率を達成した.

表 9  $M_{in}$ ,  $M_{ax}$  の範囲が 10~12 としたときの実験 1 の結果

認識率 (動画数)				
歩く	走る	スキップ	早歩き	全体
95.6%(43)	100.0%(45)	100.0%(45)	91.1%(41)	96.7%(174)

実験 1 においても, 高い認識率を達成していることがわかる. この手法にも大きな有用性がある可能性が高い. ただ,  $M_{in}$ ,  $M_{ax}$  の値が最適かどうかを判断するには実験不足であり,  $M_{in}$ ,  $M_{ax}$  の範囲が 10~12 であるとき, 偶然認識率がよくなっただけとも考えられるため, 最終的な提案手法として採用するには至らなかった.

## 6.2 全画像列から抽出した HOG と多数決を用いた手法

画素のばらつき頻度  $\sigma^2$  を用いた手法や、背景画素の出現数を考慮した手法とはかなり異なった手法である。これらの 2 つの手法との大きな違いは以下の 2 つである。

- 平均画像を使用しない。
- フレーム画像 1 枚 1 枚を SVM にかけて、フレームごとに動作の判定を行う。

この手法による動作認識手順は以下のようになる。

- (1) コンピュータに 1 つの動画画像における 1 枚の注目フレームを読み込ませる。
- (2) 注目フレームにおいて、背景画像差分により人物領域を抽出し、正規化し、切り抜く。
- (3) 正規化した人物領域から、HOG 特徴量を抽出する。
- (4) あらかじめ学習させておいた SVM に抽出した HOG 特徴量を与え、注目フレームがどの動作を示したものを識別させる。
- (5) 最終フレームまで (1)~(4) までの処理を繰り返す。
- (6) 一つの動画画像内の各フレームの識別結果を集計し、多数決を採ることによって、その動画画像における動作を決定する。

SVM での識別結果は、入力した 1 枚の画像のみの判定となる。そのため、入力した動画画像が正解動作であるかの判定は、1 枚 1 枚のフレーム画像 SVM 判定を多数決したものとなる。

動画画像のある 1 フレームでは、そのフレームがどの動作を表しているかの判定を人間による目視でも困難な場合は多い。そのため、コンピュータに静止画で動作を識別することの難易度は高いと予測できる。そのため、この手法にも HOG に時間的な変化を考慮した特徴を加える。HOG の勾配強度算出式を以下のように拡張する。

$$f_x = L_t(x+1, y) - L_t(x-1, y) \quad (19)$$

$$f_y = L_t(x, y+1) - L_t(x, y-1) \quad (20)$$

$$f_t = L_t(x, y) - L_{t-n}(x, y) \quad (21)$$

$$m_t = \sqrt{f_x^2 + f_y^2 + f_t^2} \quad (22)$$

$L_t(x, y)$  は  $t$  枚目フレームにおける画素  $(x, y)$  の輝度値を表す。 $f_x, f_y$  の算出は従来通りであるが、それに加えて  $f_t$  を算出することとする。 $f_x, f_y$  が近傍の画素との差分を計算するのに対し、 $f_t$  は、現在フレームと過去フレームとの差分を計算する。そのため、 $f_t$  は動きの変化があった場合に強く抽出されることが期待できる。これを勾配強度  $m_t$  の算出式に加えることによって拡張を行う。なお、 $n$  は過去何枚前とのフレームと比較を行うかの値であり、値を変更しながら実験を行う。

この手法を用いて実験を行う。実験方法は今までに述べたものと多く異なるが、実験環境については5章で述べたものと同じである。HOGの拡張を行う場合と、拡張を行わないそのままのHOGでそれぞれ認識率を算出した。実験1について結果は以下の表10, 11のようになる。

表10 実験1の結果：HOGを拡張させていない場合

与えた動作	与えた動作に対する判定 (動画像数)				認識率 (%)
	歩行と認識	走行と認識	スキップと認識	早歩きと認識	
歩行	45 個	0 個	0 個	0 個	100.0%
走行	0 個	45 個	0 個	0 個	100.0%
スキップ	1 個	2 個	42 個	0 個	93.3%
早歩き	32 個	0 個	1 個	12 個	26.7%

表11 実験1の結果：HOGを拡張させた場合 ( $t=3$ )

与えた動作	与えた動作に対する判定 (動画像数)				認識率 (%)
	歩行と認識	走行と認識	スキップと認識	早歩きと認識	
歩行	45 個	0 個	0 個	0 個	100.0%
走行	0 個	45 個	0 個	0 個	100.0%
スキップ	1 個	3 個	41 個	0 個	91.1%
早歩き	28 個	0 個	0 個	17 個	37.8%

全体の認識率は、HOGを拡張させていない場合80.0%、HOGを拡張させた場合82.2%となった。また実験2について結果は以下の表12, 13のようになる。

表12 実験2の結果：HOGを拡張させていない場合

与えた動作	与えた動作に対する判定 (動画像数)				認識率 (%)
	歩行と認識	走行と認識	スキップと認識	早歩きと認識	
歩行	45 個	0 個	0 個	0 個	100.0%
走行	1 個	44 個	0 個	0 個	100.0%
スキップ	2 個	3 個	40 個	0 個	88.9%
早歩き	33 個	0 個	2 個	10 個	22.2%

表 13 実験 2 の結果 : HOG を拡張させた場合 ( $t = 3$ )

与えた動作	与えた動作に対しての判定 (動画像数)				認識率 (%)
	歩行と認識	走行と認識	スキップと認識	早歩きと認識	
歩行	43 個	0 個	0 個	2 個	100.0%
走行	0 個	45 個	0 個	0 個	100.0%
スキップ	6 個	4 個	35 個	0 個	77.8%
早歩き	28 個	0 個	1 個	16 個	35.6%

全体の認識率は、HOG を拡張させていない場合 77.2%、HOG を拡張させた場合も 77.2% となった。

結果を見ると、拡張させた HOG を用いた手法が拡張していない手法に比べ、実験 1 においては若干有効であった。ここに有意な差があると言い切ることはできないが、時間変化の指標が認識率の向上につながっている。

また、この実験結果において一番注目すべき点は、どの手法のどの実験においても早歩きの認識率が極端に悪い点である。特に時間軸方向の勾配を利用していない場合の早歩きの認識率は 20% 台となっている。これは、大量の画像データを SVM に学習させることによって、識別境界がだいぶ偏ったものになっている可能性が高い。そのため、4 動作の中間動作であるような早歩きに対して、識別境界が狭くなり、認識率が下がってしまったものではないかと考えられる。

平均画像を用いた結果と比較すると、実験 1、実験 2 とともに認識率で劣っており、計算コストも大きくかかってしまうため、採用するには至らなかった。しかし、今後の拡張方法によっては、さらに高い認識率を達成できる可能性はあると考えられる。