

Haar-like 特徴と Mean-Shift を用いた複数視点人物追跡

伊東祐* (静岡大学大学院), 山下淳, 金子透 (静岡大学)

Multi-Viewpoint Human Tracking Based on Face Detection Using Haar-like Features and Mean-Shift

Yu Ito* (Shizuoka University), Atsushi Yamashita, Toru Kaneko (Shizuoka University)

Abstract

Visual surveillance systems using vision sensor increase in various environments by the rise of security needs. Human tracking is an important function for an automatic surveillance system using a vision sensor. However, it is difficult to detect a human in an image with stability under general environments and to track a human exactly in an image due to the variety of poses. This paper describes a method for automatic human tracking based on face detection using Haar-like features and mean-shift tracking. The method increases its trackability by using multi-viewpoint images. Experimental results show the validity of the method.

キーワード：人物追跡，複数視点，Haar-like 特徴，Mean-Shift
(Human Tracking, Multi-Viewpoint, Haar-like Features, Mean-Shift)

1. 序 論

セキュリティニーズの高まりにより，カメラを用いて様々な環境を監視する機会が増加している．しかし，監視作業に用いることができる人的資源には限りがあるため，この作業を自動化するための研究が盛んに行われている．近年はコンピュータビジョンの分野で画像を認識する技術が進展し，視覚情報を用いることが監視システムの自動化に際して有効な手段となりつつある．

カメラを用いた監視の目的として人物の行動の解析がある．例えば人物の動いた軌跡をたどることでその行き先を記録する，環境内の危険な領域への接近を感知して自動的にその人物に対して警報を出す等が挙げられる．

人物の行動を知るためには人物追跡を行う必要がある．自動的に人物の追跡を行うためには追跡の前段階として，画像を人物領域と人物以外の背景領域に分離（人物検出）する必要がある．

近年 Viola ら⁽¹⁾ のフレームワークを元に，Lienhart ら⁽²⁾ によって改良された Haar-like 特徴（図 1）を用いた物体検出法が提案されている．これは統計学習⁽³⁾ を使って作成した判別器を用いた物体検出手法で，高速かつ安定した物体検出を実現している．その他にも，人物の監視を行う上で多くの情報を持つ，顔の検出や認識に関する研究もある⁽⁴⁾⁽⁵⁾．

また，移動体を追跡する手法の 1 つとして Mean-Shift を用いたもの⁽⁶⁾⁽⁷⁾ がある．これらは画像内で追跡対象モデルの色情報と類似した領域を計算し，求められた類似度の分布に対して探索を行うものである（図 2）．色情報が類似していれば追跡が可能のため，追跡対象の形状変化や部分



図 1 Haar-like 特徴の例

Fig. 1. Example of Haar-like features

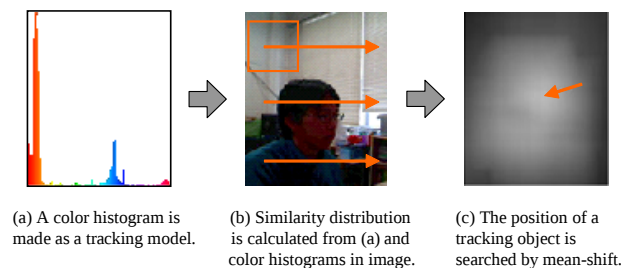


図 2 Mean-shift トラッカ

Fig. 2. Mean-shift tracker

的なオクルージョンに対して頑健である．

上に挙げた手法の応用例として，Jaffre の Video Content Indexing に関する研究⁽⁸⁾ では Lienhart らの物体検出手法⁽³⁾ で顔を検出し，Mean-Shift トラッカを用いたコストチュームの追跡を行っている．この研究では，テレビ番組のようなカメラに対して正面を向けている人物を追跡している．

また，画像認識技術に関する研究では複数視点に関するものも多い．複数視点を用いる目的は認識すべき対象を様々な角度から捉え，認識精度を向上させようというアイデアに基づいている⁽⁹⁾．

2. 研究目的

Haar-like 特徴と統計学習を用いて構成した判別器を使った顔検出は複雑な背景下でも適用でき，監視に用いる場合の検出精度に問題はない．しかし人物を追跡する際，上記の Haar-like 特徴を使った検出手法のみでは不十分である．顔を検出する場合，対象の向きや角度によって見え方は変化する．一般的に Haar-like 特徴を使った物体検出は特定の姿勢に限定され，例えば正面顔用の判別器は側面顔を検出することはできない．監視の場合は記念写真のように人物の正面顔ばかりを得られるとは限らず，この手法のみでは

顔を見失ってしまう可能性がある．そこで顔を検出できない間を補完する方法として Mean-Shift トラッカを用いる．

一方、物体追跡手法の1つである Mean-Shift トラッカは、形状変化に頑健という性質があるが、事前に追跡対象モデルが必要になる．そこで、追跡対象モデルには Haar-like 特徴により検出した顔領域の色情報を用いる．

以上のことを踏まえて、本研究では Haar-like 特徴を使った顔検出と Mean-Shift トラッカによる人物追跡を組み合わせ、安定した人物の追跡を行う．しかし、Haar-like 特徴を使った顔検出と Mean-Shift トラッカによる人物追跡では顔が検出されるまでは追跡が開始されないという問題がある．これは実際の環境では人物の顔がカメラの方向を向いているとは限らず、従って単視点では画像に人物が出現してから正面顔が得られるまでに時間がかかる可能性が大きいのである．

そこで我々は、複数視点を用いることで顔を様々な角度から捉え、単一のカメラから得られた画像では検出できない姿勢の顔も、複数視点を用いて別方向から捉えることで検出を行う手法を提案する．

また、複数視点を導入することにより Mean-Shift トラッカの追跡精度を向上させる．これは、Mean-Shift ではトラッカのサイズが変更されないため、追跡中の人物がカメラの光軸方向に移動する際、追跡に失敗することがある．そこで、人物がカメラの光軸方向に移動する際は、それに伴ってトラッカのサイズを変更することが必要になる．トラッカのサイズ変更も考慮した Mean-Shift トラッカとして CAMSHIFT 法が提案されている⁽¹⁰⁾．この手法は、画像内の情報だけでトラッカのサイズを変更しているため、追跡対象と同色の物体がトラッカに隣接すると正しく変更されないことがある．

それに対して本研究では、複数台のカメラから算出した追跡対象の3次元座標を利用し、より正確にトラッカのサイズを変更することを目指す．そして、トラッカサイズを適切に変更することで Mean-Shift トラッカの追跡精度の向上を図る．

3. 処理の流れ

始めに、単視点人物追跡処理の流れを説明する(図3)．本研究は Haar-like 特徴を使った顔検出と Mean-Shift トラッカの2つの手法の組み合わせで人物追跡を行う．まずカメラから得られた画像に対して Haar-like 特徴を使って人物の正面顔領域を検出する．検出された顔が追跡中の人物の顔でない場合は新しく追跡対象にし、顔領域の位置とサイズを記憶する．すでに追跡中の人物の顔であれば、顔領域の位置とサイズを更新する．検出した顔が追跡中の人物の顔と判断するのは、検出した顔領域の位置の近傍に前フレームで追跡している人物の顔領域が存在するときである．前フレームで追跡している人物のうち、正面顔が検出できずに未更新でいる人物が存在すれば Mean-Shift トラッカを用いて近傍の色情報が類似した領域を探索し、顔領域の位置を更新する．つまり基本的には Haar-like 特徴を用いた

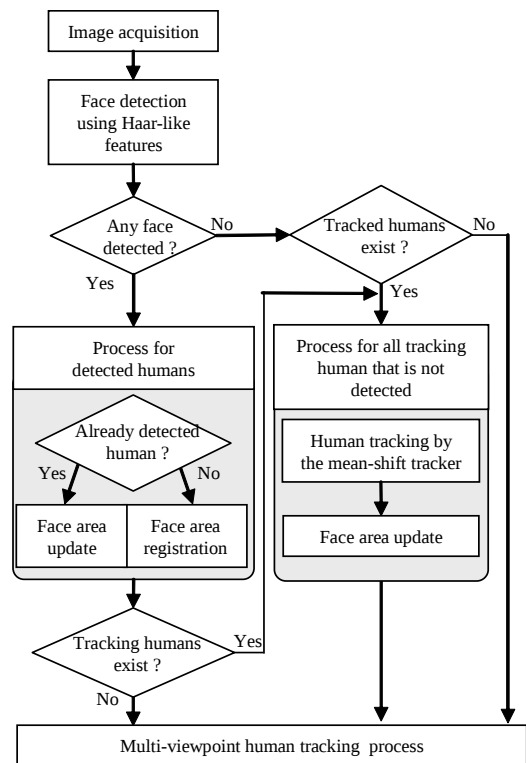


図3 単視点人物追跡処理

Fig.3. A human tracking process from a single viewpoint

顔検出で人物の位置とサイズを更新するが、それが失敗した場合は Mean-Shift トラッカを用いて人物の位置を更新する．そして、追跡中の人物のトラッカが画像端に接触した場合は、追跡を終了する．これは、Mean-Shift トラッカによる誤追跡を避けるためである．

また、Mean-Shift トラッカを使用する際、1つ前および2つ前のフレームで人物の3次元座標が算出されている場合は、人物の移動方向が予想できる．これにより Mean-Shift で使用するトラッカのサイズの変更と Mean-Shift トラッカで利用する類似度分布の作成範囲を限定する．単視点での処理が終了後、複数視点の処理を行う．

次に複数視点人物追跡処理の流れを説明する(図4)．単視点人物追跡により正面顔が検出されたら、画像中での正面顔の位置とサイズの情報を他視点に送信する．他視点から画像中での正面顔の位置とサイズの情報を受け取ると、その情報を基に正面顔に対応した後頭部を探索し、検出された後頭部を新規に登録する．検出した後頭部の領域(以後説明する上で顔領域と呼ぶ)は次のフレームから各視点で追跡する．また、視点間の対応付けを行った後、その情報を元に視点間の誤対応を削減する．

4. 単視点人物追跡

4.1 顔検出処理 入力された画像全てに対して文献(2)の Haar-like 特徴を使った顔検出手法で正面顔矩形領域を検出する．正面顔を検出した結果は図5(a)のようにな

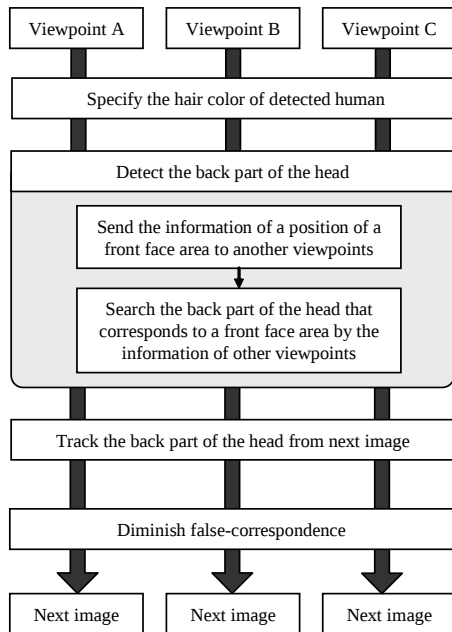


図 4 複数視点人物追跡処理

Fig. 4. A human tracking process from a multi-viewpoint

り、この矩形領域を追跡対象として新規に登録、または過去のフレームで近傍に追跡中の矩形領域が存在していれば位置とサイズの更新を行う。本研究では顔が検出されると、顔領域の色相分布を計算する。ここで求められた色相分布は Mean-Shift トラッカを使う際、追跡対象モデルの色情報として用いられる。

また、Haar-like 特徴を使った顔検出はグレースケールの画像を用いているため、図 5(b) のように明暗のパターンが似ていれば顔でないものも顔として検出される。誤検出の多くは図 5(d) の色相分布のように図 5(c) の正面顔領域の色相分布と類似していないものが多い。そのため検出された正面顔候補のうち、肌色が主でないものをこの段階で除去することで検出精度を向上させる。

4.2 Mean-Shift トラッカ 4.1 節の顔検出処理で顔を検出できず、顔領域の位置の更新ができない場合は人物の追跡を続けることができない。このような追跡中の人物については、Mean-Shift トラッカを用いて近傍の色情報が類似した領域を探索する。類似度は正面顔の検出時に求めた色相分布と現在の画像内の矩形領域の色相分布の間で、式 (1) の Bhattacharyya 係数⁽¹¹⁾を用いて求められる。ここで式 (1) において p, q は比較対象となる正規化された色相分布 ($\sum_{u=1}^m p_u = 1, \sum_{u=1}^m q_u = 1$)、 u は色相の成分番号、 m は色相の成分数を表す。

$$\rho = \sum_{u=1}^m \sqrt{p_u q_u} \dots \dots \dots (1)$$

色相分布を用いた追跡手法は文献 (7) に基づいている。この手法では顔を追跡するために彩度や明度の低い領域の色相を無視する。これは実環境において彩度や輝度の低い部

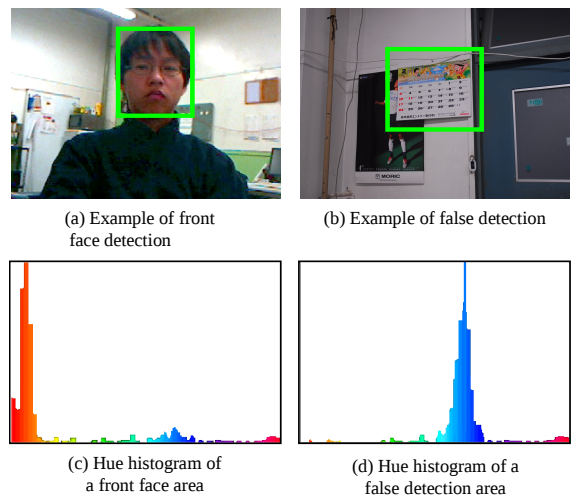


図 5 検出された顔領域

Fig. 5. Detected face region

分は照明の影響を受けやすく、追跡するときの不安定要因となるためである。しかし、追跡中の人物はカメラに対して反対方向に向くことがあり、その場合に追跡中の顔領域の大部分は毛髪が占め、黒などの明度の低い髪は追跡できなくなる。そこで本研究では明度の低い部分に限っては本来の色相分布に割り当てず、色相分布の外側に拡張した特別な色として評価することで、カメラに対して反対方向を向いた人物の顔も追跡できるようにしている。

以上の手順で色の類似度分布が求められた後、Mean-Shift によって最も追跡対象らしい場所を探索し、この結果を用いて顔領域の位置の更新を行う。

4.3 過去の移動情報の利用 5 章で説明する複数視点人物追跡により、追跡中の人物の 3 次元座標が 1 つ前と 2 つ前のフレームで算出された場合、人物の過去の移動情報が算出できる。過去の移動情報を利用するにあたって、現在のフレームで人物が存在するであろう位置 $p(t)$ を式 (2) により予測する。

$$p(t) = p(t-1) + v(t-1, t-2)\Delta t \dots \dots \dots (2)$$

ここで、 $p(t-1)$ は 1 つ前のフレームでの人物の 3 次元座標を、 $v(t-1, t-2)$ は 1 つ前のフレームと 2 つ前のフレームの人物の 3 次元座標から求めた速度ベクトルを、 Δt は 1 つ前のフレームから現在のフレームまでの時間を表す。また、処理を行うにあたって人物の 3 次元座標は、画像の縦方向、横方向、奥行方向 (カメラの光軸方向) を軸とするように変換する。

始めに、カメラの光軸方向の移動情報により、Mean-Shift で使用するトラッカのサイズを変更する。これは、2 章でも述べたように、Mean-Shift により追跡している人物がカメラの光軸方向に移動した際に正しく追跡できないことがあるために行う。式 (2) により求めた現在のフレームでの人物の予測位置 $p(t)$ から、現在のフレームで Mean-Shift による探索に用いるトラッカのサイズを求める。1 つ前の

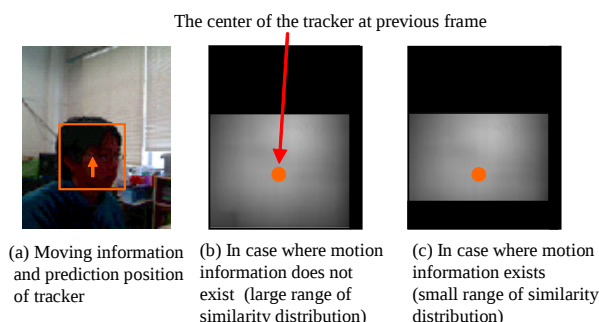


図 6 類似度分布の作成範囲
Fig. 6. Making range of similarity distribution

フレームにおいて光軸方向のカメラから人物までの距離を $p_z(t-1)$ 、現在のフレームでの人物の予測位置において光軸方向のカメラから人物までの距離を $p_z(t)$ 、1 つ前のフレームでのトラックのサイズを $S(t-1)$ 、現在のフレームで Mean-Shift による探索に用いるトラックのサイズを $S(t)$ とすると、トラックのサイズ $S(t)$ は式 (3) により算出できる。

$$S(t) = \frac{p_z(t-1)}{p_z(t)} S(t-1) \dots \dots \dots (3)$$

次に、画像の縦および横方向の移動情報により、Mean-Shift トラックで使用する類似度分布の作成範囲を限定する。これは、類似度分布の作成範囲を限定することにより Mean-Shift トラックによる追跡の失敗を減らすために行う。例えば、追跡対象に対して、その周辺の背景が追跡対象と同じ色分布の場合は追跡に失敗する可能性がある。このような場合に、人物の過去の移動情報を用いることで追跡に失敗する可能性を削減することができる。移動情報が存在しないとき、類似度分布は 1 つ前のフレームでのトラックの中心から左右上下に一定範囲で作成する。逆に、縦および横方向の移動情報が存在するときは、移動情報に基づいて作成範囲の限定を行う。作成範囲の限定は、移動情報が存在しないときの類似度分布の作成範囲に対して、人物の移動方向とは反対側の領域を削減する（図 6）。削減の度合いは、人物の移動量により変化する。移動量が大きい場合ほど削減の割合は大きくなり、移動量が小さい場合は削減の割合も小さくなる。

トラックのサイズを変更や類似度分布の作成範囲を限定した後、4.2 節の Mean-Shift トラックにより人物の追跡を行うことで、Mean-Shift トラックの追跡精度を向上させる。

5. 複数視点人物追跡

5.1 毛髪の色の特 5.2 節で説明する視点間の同一人物の対応付けを行うためには、対応付けを行う人物の毛髪の色が必要となる。視点間の同一人物の対応付けは正面顔を検出した際に行うため、正面顔領域から毛髪の色を抽出する。正面顔の領域は 4.1 節で説明した Haar-like 特徴を用いた顔検出手法により検出される（図 7 の緑枠）。検出した正面顔の領域において、その領域の上部に毛髪が写っていると予想できるため、正面顔領域の上部を毛髪

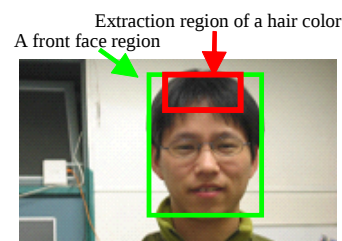


図 7 毛髪の色抽出
Fig. 7. Extraction of a hair color

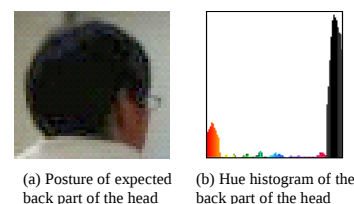


図 8 斜め後方から見た顔領域とそのヒストグラム
Fig. 8. Posture of expected back part of the head and its hue histogram

色特定領域に指定する（図 7 の赤枠）。次に、毛髪の色特定領域において最も頻度の高い色を特定し、毛髪の色とする。頻度の高い色の特定は、毛髪の色特定領域に存在する全ての色を HSV 空間に投票することで算出する。

5.2 視点間の対応付け 顔検出精度の向上を目的とした複数視点人物追跡は、位置関係が既知である 3 台のカメラを用いて行う。3 台のカメラの配置は、カメラを配置する環境により異なるが、今回は三角形に配置し、カメラの光軸が三角形の中心に向くようにする。これにより、人物の顔が 3 台のカメラから全く見えなくなる可能性を少なくする。

あるカメラで人物の正面顔を検出した場合、その他の画像中での同一人物の顔の位置をエッジボーラ拘束により限定し、推定する。ただし、画像中に映る正面顔の位置情報だけでは、その他の画像で同一人物が存在する位置は特定できない。しかし、3 台のカメラを三角形に配置しているため、1 台のカメラで正面顔を検出すると、その他のカメラには同一人物を斜め後方から見た画像が映っていると予想できる（図 8）。そこで、5.1 節で求めた毛髪の色から人物の斜め後方から見た画像のヒストグラムを作成し、そのヒストグラムと類似度の高い領域を探索することで顔領域を検出する。

5.3 誤対応の削減 5.2 節で説明した視点間の同一人物の対応付けだけでは、視点間で間違っ

て対応付けすることがある。そこで、あるカメラで人物の正面顔検出し、残り 2 台のカメラにおいても正面顔に対応した領域を検出できた場合、視点間で行う同一人物の対応付けが適切かどうか判断する。

視点間の対応付けが適切かどうかの判断は、正面顔を検出したカメラとその他 2 台のカメラにより同一人物の 3 次

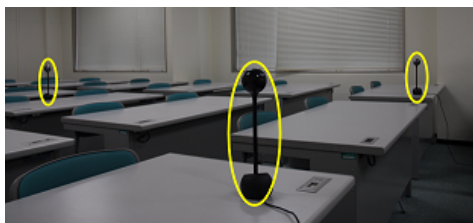


図 9 実験環境

Fig. 9. The experimental environment



図 10 人物追跡結果

Fig. 10. The experimental result

元座標値を2組算出し、算出した3次元座標値を比較することで行う。視点間の対応付けが適切な場合は、2組の3次元座標値は同じになる。もし、2組の3次元座標値が異なる場合は、誤った対応付けを行ったことになる。そして、誤対応が存在する場合は、誤対応を検出・除去する。

誤対応の検出は、2台のカメラにより算出した3次元座標値を残りのカメラで撮影した画像に投影し、投影した位置を中心とした矩形領域のヒストグラムが顔(後頭部)ヒストグラムかどうかで判断する。矩形領域のヒストグラムが顔(後頭部)ヒストグラムの場合、投影した3次元座標値は正しいと判断する。もし、矩形領域のヒストグラムが顔(後頭部)ヒストグラムでない場合は、投影した3次元座標値が間違っていることを示し、3次元座標値を算出した画像で誤対応があることが分かる。

誤対応の修正は、誤対応の検出により正しいと判断された3次元座標値を投影した点を中心とした矩形領域を正しい顔領域とすることで行われる。

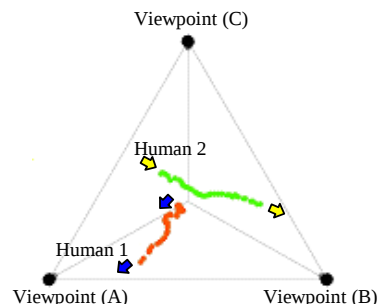


図 11 人物の移動軌跡

Fig. 11. Trajectory of tracking humans

6. 実験

実験は先に述べたように3台のカメラを正三角形に配置して行った(図9)。

まず、3視点での人物追跡を行った結果を図10に示す。画像中の四角形は追跡中の人物の顔領域の検出結果を示している。また、人物の移動軌跡を図11に示す。

図10(a-1)では正面顔が検出され、人物1とラベル付けされている。図10(a-1)に対応したFig.10(b-1)・(c-1)では、正しく視点間の対応付けができていることが分かる。図10(a-1)で検出した正面顔が時間の経過に伴って見えなくなる場合もMean-Shiftにより追跡できていることが分かる(図10(a-2)・(a-3))。次に、図10(b-4)で正面顔が検出され、人物2とラベル付けされている。そして、図10(b-4)に対応した図10(a-4)・(c-4)では正しく視点間の対応付けが行われている。以後、各視点で人物2が見えている限り正しく追跡ができていることが分かる。

また、図11に、図10で検出した人物1の移動軌跡を橙色で、人物2の移動軌跡を緑色で表示した結果を示す。図11では、図10により3次元座標値が正確に算出されていることが分かり、人物1・人物2の移動軌跡も正しいことが分かる。

次に、3視点人物追跡の精度の検証を行った。精度は評価を行うにあたって、3視点人物追跡結果を9つ用意した。9つの実験結果に対して2つの項目で検証を行った。1つ目は検出対象(顔)を正確に検出した割合を示す対象検出率であり、2つ目は検出対象以外の物を誤って検出した割合を示す誤検出率である。また、それぞれの項目とも、各実験結果に対して3台のカメラそれぞれで検証した。

1つ目の対象検出率は、3視点のいずれかで正面顔が写った後、顔が写っているフレーム数に対して検出できた割合を示す。その結果を表1に示す。

表1を見ると、対象検出率は90%~100%の間に結果が集中していることが見て取れ、平均も97%と非常に高くなっている。検出率が100%にならなかった主な原因としては、Mean-Shiftトラッカによる誤追跡が挙げられる。これは、トラッカの内部に含まれる背景の割合が追跡するにつれて大きくなってしまったためである。解決策としては、類似

表 1 対象検出率

Table 1. Object detection accuracy

Data number	Camera 0	Camera 1	Camera 2
1	99	100	96
2	100	100	70
3	100	100	100
4	100	100	100
5	86	100	100
6	100	100	100
7	100	100	100
8	100	78	100
9	100	100	100

(Unit : %)

表 2 誤検出率

Table 2. False-detection rate

Data number	Camera 0	Camera 1	Camera 2
1	3	0	0
2	0	0	21
3	0	0	0
4	0	2	0
5	6	58	0
6	0	52	0
7	0	0	0
8	1	12	1
9	0	0	0

(Unit : %)

度分布を作る際に肌や毛髪の色も考慮に入れることが考えられる。

2 つ目の誤検出率は、撮影したフレーム数に対して誤検出が存在したフレーム数の割合を示す。誤検出率の結果を表 2 に示す。

誤検出率は、0%～10%の間に結果が集中していることが見て取れ、概ね良好な結果と言えるが一部で 58%や 52%と非常に高い値になっている。この原因は大きく 2 つある。1 つは表 1 でも述べた Mean-Shift トラッカによる誤追跡である。もう 1 つは、Mean-Shift トラッカにより追跡中の人物が画像の外に出た際に、トラッカが消えずに画像内に残ってしまうことがある。提案手法は、画像端にマージンを設けて画像端にきたトラッカを消している。しかし、人物の移動速度やマージンの大きさによって画像内にトラッカが残ってしまう。従って、画像内に残ったトラッカを消す手法が今後必要になる。

2 つの検証項目ともに、いくつかの課題が残っているが良好な追跡精度が出ていることが分かる。そして、今後課題を解決することで追跡精度の向上が期待できる。

以上の結果より、提案手法の有効性が確認できた。

7. 結 論

Haar-like 特徴を使った顔検出と Mean-Shift トラッカの組み合わせによる人物追跡を、3 視点で行うことにより顔検出精度の向上を図る手法を構築し、その有効性を確認した。また、過去の移動情報を利用することで、Mean-Shift

トラッカの追跡精度の向上を図り、その有効性も確認した。しかし、追跡対象の位置や動きにより Mean-shift トラッカによる追跡に課題が残った。

今後の課題として、正面以外の顔の検出や人物の追跡範囲の設定や Mean-Shift トラッカの追跡精度の向上を行う必要がある。

謝 辞

本研究の一部は、財団法人スズキ財団科学技術研究助成の補助を受けた。

参考文献

- (1) P. Viola and M. J. Jones: "Rapid Object Detection Using a Boosted Cascade of Simple Features", *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.511-518, (2001).
- (2) R. Lienhart and J. Maydt: "An Extended Set of Haar-like Features for Rapid Object Detection", *Proceedings of the 2002 IEEE International Conference on Image Processing*, Vol.1, pp.900-903, (2002).
- (3) Y. Freund and R. E. Schapire: "Experiments with a New Boosting Algorithm", *Proceedings of the 13th International Conference on Machine Learning*, pp.148-156, (1996).
- (4) M. A. Turk and A. P. Pentland: "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, Vol.3, No.1, pp.71-86, (1991).
- (5) M. A. Turk and A. P. Pentland: "Face recognition using eigenfaces", *Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.586-591, (1991).
- (6) D. Comaniciu, V. Ramesh and P. Meer: "Real-Time Tracking of Non-Rigid Objects Using Mean Shift", *Proceedings of the 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.2, pp.142-149, (2000).
- (7) G. R. Bradski: "Real Time Face and Object Tracking as a Component of a Perceptual User Interface", *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision*, pp.214-219, (1998).
- (8) G. Jaffre and P. Joly: "Costume: A New Feature for Automatic Video Content Indexing", *Proceedings of RIAO2004*, pp.314.325, (2004).
- (9) P. Braud, M. Dhome, J. T. Laprest and N. Daucher: "Modelled Object Pose Estimation and Tracking by a Multi-Camerassystem", *Proceedings of the 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.976-979, (1994).
- (10) G. R. Bradski: "Computer Vision Face Tracking For Use in a Perceptual User Interface", *Intel Technology Journal*, No.Q2, p15, (1998).
- (11) G. Xuan, P. Chai and M. Wu: "Bhattacharyya Distance Feature Selection", *Proceedings of the 13th International Conference on Pattern Recognition*, Vol.2, pp.195-199, (1996).