

# シルエットを用いた Tree-Based Filtering による人体の姿勢推定

岡田 隆三<sup>†</sup> シュテンガ ビヨン<sup>††</sup>

<sup>†</sup> (株) 東芝 研究開発センター

〒 212-8582 神奈川県川崎市幸区小向東芝町 1

<sup>††</sup> 東芝欧州研究所

1 Guildhall St, Cambridge CB2 3NH, UK

E-mail: <sup>†</sup>ryuzo.okada@toshiba.co.jp, <sup>††</sup>bjorn.stenger@crl.toshiba.co.uk

あらまし 関節物体の効率的な姿勢推定手法として、木構造に基づくフィルタリング (Tree-based filtering) が提案されており、三次元形状モデルを用いた手の姿勢推定に応用されている。本論文では、単眼カメラを用いてマーカー等を必要としない人体全身の姿勢推定を行うため、木構造をシルエット距離に基づいて生成し、各部位の隠蔽を考慮した予測を行うことにより、効率的で安定な姿勢推定手法を提案する。画像特徴としては、背景差分に基づくシルエットを用い、人体の三次元モデルから得られるシルエットと画像から観測されるシルエット間のコスト関数を、シルエットの太さを考慮して正規化することにより、手や足などの細い部位の推定を安定化する。実験によって提案手法の有効性を示すとともに、ハイエンド PC を用いた場合約 127 [ms/frame], Cell ブロードバンドエンジンを用いた場合約 86 [ms/frame] のオンラインモーションキャプチャシステムが実現できることを示す。

キーワード 関節物体追跡, 姿勢推定, モーションキャプチャ, 木構造に基づくフィルタリング, Cell ブロードバンドエンジン

## Human Posture Estimation using Silhouette-Tree-Based Filtering

Ryuzo OKADA<sup>†</sup> and Björn STENGER<sup>††</sup>

<sup>†</sup> Corporate Research and Development Center, Toshiba Corporation

1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki-shi, Kanagawa, 212-8582 Japan

<sup>††</sup> Toshiba Research Europe LTD

1 Guildhall St, Cambridge CB2 3NH, UK

E-mail: <sup>†</sup>ryuzo.okada@toshiba.co.jp, <sup>††</sup>bjorn.stenger@crl.toshiba.co.uk

**Abstract** This paper presents a method for marker-less human motion capture using a single camera. It is based on tree-based filtering, a method for posture estimation of articulated objects. In our case we define the tree structure based on a silhouette based distance of human postures and use a dynamic model which takes self-occlusions into account by increasing the variance of occluded body-parts. A new normalized cost function between the observed silhouette and a model silhouette, obtained from a 3D human body model, is proposed, which improves the estimation with respect to thinner body parts such as arms and legs. Experimental results show the effectiveness of the method. An online motion capture system is realized at a frame rate of 127 [ms/frame] and 86 [ms/frame] using a high-end-class PC and a Cell broadband engine, respectively.

**Key words** articulated object tracking, posture estimation, motion capture, tree-based filtering, Cell broadband engine

### 1. はじめに

画像による人体の姿勢推定は、コンピュータインターフェース, ゲーム等のエンターテイメント, 画像監視, モーションキャプチャなど様々な応用が考えられ, 盛んに研究が行われてい

る [1] ~ [11].

人体は多数の関節からなる高い自由度を持った関節物体であり, インターフェースやゲーム等実時間性が要求されるアプリケーションでは, 高次元の姿勢状態空間を効率よく探索して最適な姿勢を求めることが重要な課題となる. フレーム間の姿勢

変化を検出し、前のフレームの姿勢に変化を加算することによって姿勢推定を行う手法 [6], [7] は、前のフレームの姿勢の近傍のみを考慮することにより計算量を低減することができる。しかし、手動で正確な初期姿勢やキーフレームの姿勢を与えなければならないといった問題や、推定誤差が蓄積するという問題がある。体の各部位を画像から抽出 [10], [11] することにより、初期化の問題や計算コストの問題を解決できるが、マーカー等を用いずに画像上の人体の部位の位置を検出することは難しい問題であり、カメラの数が少ない場合には、部位の隠蔽も問題となる。姿勢パラメータと画像から観測されるシルエットやエッジ等の画像特徴の間の非線形性も安定な推定を難しくする要因である。このような問題に対して、パーティクルフィルタに基づく手法 [8], [9] や画像特徴と姿勢の関係の学習に基づく手法 [12] が提案され、人体の姿勢追跡にも応用されている。これらの手法も、手動の初期化が必要であったり、計算コストが高いといった問題がある。

このような問題に対して、姿勢の状態空間の分割密度を変化させることによって得られる姿勢の木構造を利用し、ベイズ推定の枠組みに基づいて、効率的かつ安定に高次元の状態空間を探索する手法 (Tree-based filtering) [13] が提案され、手の姿勢推定に応用されている。この手法では、あらかじめ下層に向かって姿勢状態空間の分割が細くなるような木構造を生成しておく。認識時には、画像特徴であるエッジと、木構造の各ノードを代表する姿勢から得られる輪郭とのマッチング、および運動モデルによる予測を行って最適な姿勢を探索する。このとき、上位層で得られた結果を用いて下位層の枝刈を行いながら、最適な姿勢を粗密探索することにより効率的な探索を行うことができる。状態空間の分割方法としては、関節角度の差に基づく手法 [13] と画像特徴の類似度に基づく手法 [14] ~ [16] がある。後者は、木構造のノードが関節角度の大きく異なる姿勢から構成されることもあるため運動モデルの導入が容易ではないが、モデルと画像特徴のマッチングにおいて無駄がなく効率的である。本論文では、画像特徴の類似度に基づいて構成した木構造に運動モデルを導入し、効率的で安定な姿勢推定を行う。

カメラの数が少ない場合には、部位の隠蔽が問題となるが、一般にベイジアンフィルタに基づく追跡手法は、運動モデルによって次のフレームの姿勢を予測して事前確率を生成するため、隠蔽された部位の運動が運動モデルに従っていれば追跡が可能である。しかし、隠蔽が起こっている間に、部位の運動がこの運動モデルから大きく外れている場合には、追跡が難しくなる。そこで提案手法では、部位の隠蔽を考慮して予測モデルを切り替えることによって、隠蔽に対して安定な姿勢推定を行う。

ベイジアンフィルタに基づく追跡手法では、状態空間中の姿勢と観測値である画像を関係づける尤度を、姿勢と画像のマッチングによって決定するので、マッチングに用いる画像特徴の選択と評価関数の定義も重要な課題である。人体全身の追跡には、画像特徴として、シルエット [3] ~ [5] や奥行き [2], エッジ [1], 動き [17] 等が用いられており、これらはアプリケーションに応じて適切に選択する必要がある。本論文では、カメラの

設置の容易さやシステム規模を小さくすることを念頭において、固定の単眼カメラを使用する。また、背景や照明条件はある程度制御できるという前提の下、背景差分によるシルエットを用いる。評価関数は姿勢探索中に多くの回数評価されるため、実時間性を考慮する場合には、計算コストが小さくかつ安定な評価関数が要求される。本論文では 2 つのシルエット間の類似度の評価値として、シルエットの中心に近いほど重みを大きく設定した排他論理和 (core-weighted XOR) [18] を、部位の太さを考慮して正規化した、正規化中心重み付排他論理和を使用する。この評価値により、腕など細い部位と胴体など太い部位を等価に扱うことができ、細い部位の推定の安定性が向上する。

以下では、まず 2. 節で従来の Tree-based filtering と提案手法の違いについて概要を述べる。3. 節で画像特徴に基づく木構造の生成手法、4. 節で隠蔽を考慮した姿勢予測モデルについて述べ、5. 節でシルエットの評価値およびこの評価値を用いた尤度を定義する。6. 節で画像特徴に基づく Tree-based filtering について述べ、7. 節で実験結果を示す。さらに、8. 節で提案手法を用いた実時間モーションキャプチャシステムについて述べる。

## 2. 木構造に基づくベイジアンフィルタリング

まず、Tree-based filtering [13] について簡単に述べる。

姿勢追跡の問題は、以下のようにベイジアンフィルタを用いて定式化できる。ある時刻  $t$  の姿勢を、人体の三次元位置および全ての関節の角度を要素とする状態ベクトル  $x_t$ 、時刻  $t$  に観測値として画像から得られた画像特徴を  $z_t$ 、時刻  $t$  までに得られた観測値の列を  $z_{1:t}$  とする。各時刻について姿勢の事後確率分布を求めることにより、姿勢推定を行う。現在の観測値  $z_t$  は過去の観測  $z_{1:t-1}$  と独立であると仮定すると、事後確率分布はベイズ則により

$$p(x_t | z_{1:t}) = c_t p(z_t | x_t) p(x_t | z_{1:t-1}) \quad (1)$$

となる。ここで、 $c_t$  は姿勢  $x_t$  と独立な正規化定数、 $p(z_t | x_t)$  は尤度、 $p(x_t | z_{1:t-1})$  は事前確率分布である。事前確率分布は、以下のように計算される。

$$p(x_t | z_{1:t-1}) = \int p(x_t | x_{t-1}) p(x_{t-1} | z_{1:t-1}) dx_{t-1} \quad (2)$$

つまり、初期事前確率分布  $p(x_1)$  を既知として、式 (2) による予測と観測値を用いて、式 (1) による確率分布の更新を繰り返すことにより、各時刻の姿勢  $x_t$  が推定される。 $p(x_t | x_{t-1})$  は、フレーム間の姿勢遷移確率分布であり、運動モデルに相当する。

これらを直接計算することは難しいため、姿勢の状態空間を複数の解像度で分割することによって生成した木構造を用い、上位層では少ないノードによって荒く事後確率分布を近似し、その結果を用いて事後確率が低い下位層の枝刈を行い効率的に事後確率を計算する [13], [19]。木構造を用いた粗密探索手法では、画像特徴の類似度によって木構造を構成する手法 [14], [15] も多く用いられているが、木構造の各ノードが類似した姿勢からのみ構成されているとは限らず、運動モデルの導入が困難である。そのため、Tree-based filtering [13], [19] では状態空

間内の距離に基づいて状態空間を分割した木構造を用いている。

部位の隠蔽が起こった場合、Tree-based filtering では隠蔽された部位の運動が運動モデルに従っていれば、追跡が可能であると考えられる。しかし、姿勢の状態空間内の距離に基づいて状態空間を分割しているため、画像特徴が変化しない隠蔽されている部位の姿勢の変化に対しても、計算量の大きい画像と姿勢の評価関数の計算を何度も行うことになり、無駄が多い。また、人体の腕のように動きの自由度が高い部位は、隠蔽されている間の運動が運動モデルに従わないこともある。このような場合、確率密度分布で表される運動モデルの分散を大きくすることが考えられるが、以下のような問題がある。(1) 部位の隠蔽を検出することが困難、(2) 隠蔽が起こっていない部位の分散も大きく設定すると、計算量の削減のために行う枝刈りの効果が弱くなり、探索範囲が大きくなるため効率が悪い、(3) 運動モデルは姿勢推定結果の時間的な連続性を保つ役割も果たし、画像特徴だけでは決定できない姿勢のあいまい性<sup>(注1)</sup>を解決するという働きもあるが、分散を大きくするとこの効果も弱くなり、推定結果が不安定になる場合がある。

提案手法では、画像特徴の類似度に基づいて木構造を構成することにより、画像特徴に変化を与えない隠蔽部位などの姿勢変化に対する無駄な評価関数の計算を削減することができる。また、安定な姿勢推定のために運動モデルは不可欠なので、画像特徴の類似度に基づいて構成した木構造に、運動モデルを導入する。シルエットを計算する際に求まる部位の隠蔽情報を利用して、隠蔽が起こっている部位に関してのみ運動モデルの分散を大きくすることで、隠蔽部位の姿勢変化に対応することができる。

### 3. 画像特徴に基づく姿勢の木構造

認識を行う姿勢の状態空間は非常に広いため、市販のモーションキャプチャシステムを用いて姿勢のサンプルを多数収集し、これら姿勢サンプルの集合を姿勢の状態空間  $R$  と定義する。

木構造の階層  $l-1$  において、状態空間が  $N_{l-1}$  個の重ならない部分  $\{S^{i(l-1)}\}_{i=1}^{N_{l-1}}$  に分割されているとする。つまり、 $R = \cup_{i=1}^{N_{l-1}} S^{i(l-1)}$  である。 $S^{i(l-1)}$  を、下位の階層  $l$  において以下のように分割する。ただし、最上位階層  $l-1=0$  においてはノード数  $N_0=1$  で、その下の階層  $l=1$  では状態空間全体が分割対象の空間となる ( $S^{10}=R$ )。

(1)  $S^{i(l-1)}$  に含まれる全ての姿勢サンプルの平均値に最も近い姿勢サンプルを、木構造のノードを代表する姿勢(代表姿勢)として、新しいノードを生成する。

(2)  $S^{i(l-1)}$  に含まれる姿勢の中で、現在までに選択されている全ての代表姿勢からの画像特徴距離(5. 節で定義するシルエット間の距離)が閾値  $T_l$  より大きく、代表姿勢から平均距離が最も大きい姿勢を選択する。このような姿勢が存在すれば、新たなノードを生成してその代表姿勢とし、2へ戻る。存在しなければ、3へ進んで代表姿勢の選択を終了する。

(注1): 例えば正面を向いているか背面を向いているかは、輪郭やシルエットからだけでは判断できない。

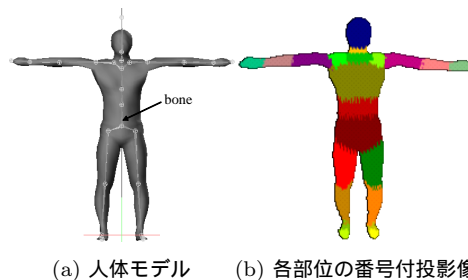


図1 人体の三次元モデル(男性洋服のJIS規格(JIS L4004)に基づくMAサイズの人体モデル。ポリゴン数は4,500、関節数は27。)

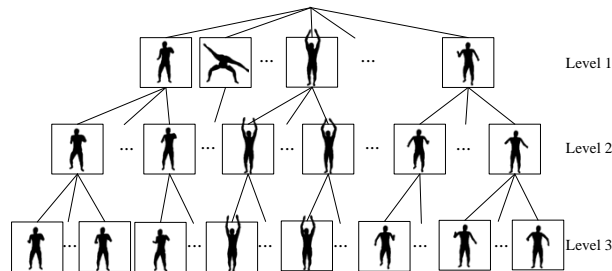


図2 生成された木構造の例(3階層で、総姿勢数は57,136姿勢。各階層のノード数は上位階層から順に7,828, 26,805, 44,108姿勢。)

(3)  $S^{i(l-1)}$  の中で、代表姿勢に選ばれていない姿勢について、最も画像特徴距離が小さい代表姿勢を選択し、代表姿勢の属するノードに登録する。各ノードに属する姿勢サンプルの集合が分割された状態空間  $S^{jl}$  となる。

(4) 各ノードに属する姿勢サンプルの平均値に最も近い姿勢サンプルを、代表姿勢として再定義する。以上をあらかじめ定められた階層に達するまで再帰的に実行し、木構造を生成する。

#### 3.1 画像特徴

次に、姿勢サンプルから木構造を作る際に用いる画像特徴であるシルエットを抽出する方法について述べる。

様々な姿勢のシルエットを生成するために、あらかじめ姿勢推定を行う人物の体の三次元形状モデルを次のように取得しておく(図1(a)参照)。

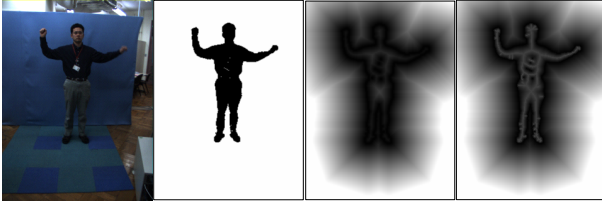
(1) 追跡対象人物の表面形状をポリゴンで近似する<sup>(注2)</sup>。画像への投影にかかる計算量の削減のためのポリゴン数は数千程度に削減する。

(2) 姿勢を変化させるための骨格構造を埋め込む。

(3) 表面形状を変形させるため骨格構造の各部位にポリゴンを分類する。

人体モデルを、木構造の各ノードの代表姿勢に変形させ、全ポリゴンを隠蔽を考慮しながら画像に投影する。このとき、各ポリゴンが属している部位の一意に定められた番号を画素値とすることにより、各部位の番号付投影像を得ることができる(図1(b)参照)。シルエットはこの投影像を背景とそれ以外で2値化すれば得られる。

(注2): 本論文では、7. 節で述べるように、あらかじめ用意しておいた典型的な体型の標準人体モデルから、追跡対象人物の体型に近いモデルを選択する。



(a) Original (b) Silhouette (c) wXOR (d) nwXOR

図3 シルエット抽出およびシルエット間距離に用いる重み ((c) と (d) はそれぞれ core-weighted XOR と正規化 core-weighted XOR に用いる重みで、暗いほど重みが小さいことを表す.)

2つの姿勢から抽出されたシルエット間の距離については、5.節で述べる。

#### 4. 隠蔽を考慮した運動モデル

3.1節で述べたように、人体の三次元モデルの画像への投影によって部位の番号付投影像が得られる。この投影像を用いて、各部位の面積(画素数)が得られる。ある姿勢  $x$  を持つ人体モデルの部位  $j$  の投影面積を  $A_j$  とすると、投影面積が閾値  $T_A$  より小さい部位を隠蔽部位とする。隠蔽情報は、以下のようになる。

$$F_{occ}(x) = \{f_j\}_{j=1}^{N_b}, \quad f_j = \begin{cases} 1 & \text{if } A_j < T_A \\ 0 & \text{else} \end{cases} \quad (3)$$

ここで、 $N_b$  は人体の部位の数である。

運動モデルは、過去のフレームの姿勢から現在のフレームの姿勢への遷移確率分布で与えられ、一次マルコフモデルを仮定して  $p(x_t|x_{t-1})$  となる(式(2))。本論文では、認識する動きに対する制約を少なくするため、運動モデルとしてランダムウォークモデルを採用する。すなわち、前フレームの姿勢  $x_{t-1}$  を平均値とする正規分布を運動モデルとして用いる。

$$p(x_t|x_{t-1}) \sim N(x_{t-1}, \Sigma) \quad (4)$$

ここで、 $\Sigma$  は各部位の運動の速さを考慮して実験的に定めた。

提案手法では、正規分布の分散を部位の隠蔽情報  $F_{occ}(x)$  を用いて、隠蔽が起きている部位に対しては、分散の大きい運動モデルを選択する。

$$\Sigma = \text{diag}(\sigma_j'^2), \quad \sigma_j' = \begin{cases} \sigma_j & \text{if } f_j = 0 \\ m\sigma_j & \text{if } f_j = 1 \end{cases} \quad (5)$$

ここで、 $\sigma_j$  は、各関節に関する運動モデルの標準偏差、 $m > 1$  は隠蔽が起きている場合の標準偏差の拡大率を表すパラメータである。本論文の実験では、 $m = 5$  を用いている。これにより、隠蔽部位についてのみ運動モデルに合致しない姿勢変化を許容する。

#### 5. 尤度

本節では、画像から得られる観測値であるシルエットと、姿勢を関係づける尤度を定義する。

##### 5.1 画像からのシルエット抽出

本論文では、照明条件や背景はある程度制御できると仮定し

て、カラー画像の背景差分によって、シルエットを抽出する。

まず、追跡対象の人物がカメラに映っていない状態で、背景モデルを学習する。多数の背景画像を撮影し、各ピクセルの色ベクトル (R, G, B) を大きさで正規化し、平均ベクトルと各要素の分散を対角成分とする分散行列を計算する。これらの平均値と分散をパラメータとする正規分布を背景の各ピクセルに関する尤度関数とする。この尤度関数と、現在の画像の正規化色ベクトルを用いて尤度を計算し、尤度が閾値より小さいピクセルの集合をシルエットとする。

ただし、大きさを正規化した色ベクトルを用いているため、影や照明条件の変動にはロバストになっているが、背景と輝度値のみが異なる物体は検出できない。そこで、背景と現在の画像の間で、正規化していない色ベクトルの差が十分大きい明らかに背景ではない画素もシルエットとして抽出する。このようにして画像から観測されたシルエットを、 $P_o$  とする。

##### 5.2 シルエット間の距離

次に、2つのシルエット間の距離を定義する。最も単純なシルエット間の距離は、シルエット部分の画素値を1、背景部分を0として、両者の排他的論理和 (XOR) が真 (1) であるピクセルの密度が考えられる。2つのシルエット画像を  $P(x, y)$ 、 $Q(x, y)$  とすると、

$$d_{sil}(P, Q) = \frac{1}{N_W} \sum_{(x,y) \in W} XOR(P(x, y), Q(x, y)) \quad (6)$$

ここで、 $W$  は二つのシルエットを内包する領域で、 $N_W$  は  $W$  に含まれる画素数である。

これに対して、シルエットの輪郭から距離に比例して重みを付ける手法 [18] が提案されている。この手法では、上式の XOR の代わりに、シルエット画像  $P(x, y)$  のシルエット輪郭からの距離  $d_P(x, y)$  で重み付けした、次の  $wXOR$  (core-weighted XOR) を用いる。

$$wXOR(P(x, y), Q(x, y)) = d_P(x, y) XOR(P(x, y), Q(x, y)), \quad (7)$$

$$d_P(x, y) = D_P(x, y) + wD_{\bar{P}}(x, y) \quad (8)$$

ここで、 $D_P$  はシルエット画像  $P$  のシルエットに関する距離変換画像、 $D_{\bar{P}}$  はシルエット画像の背景とシルエットを反転させた距離変換画像である。輪郭から離れるほど重みが大きくなるため、輪郭付近の微妙な変化より、2つのシルエット間の腕や脚の中心部が重なっていないといった、構造的な違いに対して距離が大きくなるよう設計されている。

しかし、重みが輪郭からの距離に基づいて定義されているため、腕や脚といった細い部位の中心部の重みが、胴体などの太い部位の中心部の重みに比べて小さくなる。提案手法では、部位の太さに関係なく、中心部が重なり合っていない場合に均一に距離が大きくなるよう、以下のように重み  $d_P(x, y)$  を定義する。

$$d_P = \frac{D_P}{D_P + D_{P_{med}}} + w'D_{\bar{P}} \quad (9)$$

ここで、 $P_{med}$  はシルエット画像  $P$  を細線化処理した画像であ

り、 $D_{P_{med}}$  は、細線化画像の距離変換画像である。また、上式では  $(x, y)$  を省略している。つまり、シルエット内部の輪郭からの距離を、輪郭からシルエットの細線 (中心線) までの距離で正規化することにより、太い部位でも細い部位でもそれらの中心部の重みは等しくなり、細い部位の姿勢推定が安定に行えるようになる。なお、 $w'$  は、シルエット内部と外部でのシルエットの不一致に対する重みを調整する係数である。本論文では、姿勢の木構造に含まれる全ての姿勢を含む矩形を外部領域の外側境界と定義している。この場合、内部領域に比べて外部領域が大きく、内部領域の重み  $\frac{D_P}{D_P + D_{P_{med}}}$  が外部領域の重み  $D_{\bar{P}}$  に比べて小さい数値をとる。そのため、内部領域の重み  $\frac{D_P}{D_P + D_{P_{med}}}$  と外部領域の重み  $w' D_{\bar{P}}$  の最大値が等しくなるように  $w'$  を設定している。

### 5.3 尤度関数

次に、このシルエット距離を用いて観測値であるシルエット  $P_o$  に関する尤度を求める。

3.1 節で述べたように、ある姿勢  $x_t$  に変形した三次元人体モデルから、シルエット  $P_m$  を得ることができる。観測シルエット  $P_o$  とモデルシルエット  $P_m$  の距離は、前節で述べたように  $d_{sil}(P_o, P_m)$  である。よって、尤度  $p(z_t = P_o | x_t)$  は、 $d_{sil}(P_o, P_m)$  の関数となり、本論文ではこれを正規分布で近似する。

$$p(z_t | x_t) \sim \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{d_{sil}(P_o, P_m)^2}{2\sigma^2}\right) \quad (10)$$

ここで、 $\sigma^2$  は分散で本論文では実験的に定めた。

Tree-based filtering では、木構造のノードに関する尤度を求める必要がある。各ノードは代表姿勢にシルエットが似ている姿勢で構成されているので、階層  $l$  のノード  $j$  に属する姿勢に関して尤度は一定と近似し、ノードの代表姿勢  $\bar{x}_t^{jl}$  に関する尤度とする。

$$p(z_t | \theta_t^{jl}) \sim p(z_t | \bar{x}_t^{jl}) \text{vol}(S^{jl}) \quad (11)$$

ここで、 $\theta_t^{jl}$  は、階層  $l$  のノード  $j$  に属する姿勢の集合、 $\text{vol}(S^{jl})$  はノード  $j$  に含まれる姿勢集合  $\theta_t^{jl}$  によって形成される分割された状態空間  $S^{jl}$  の体積を表す。各ノードに対応する分割された状態空間  $S^{jl}$  は、ノードに含まれる姿勢集合  $\theta_t^{jl}$  に応じて大きさが異なるが、本論文では単純化のためその体積  $\text{vol}(S^{jl})$  を一定とする。

## 6. 画像特徴に基づく Tree-based filtering

画像特徴に基づく Tree-based filtering においても、3. 節で生成した木構造を用いて、最下層の各ノードの事後確率を従来の Tree-based filtering と同様に計算することができる。4. 節で述べた隠蔽を考慮した運動モデルと前フレームの事後確率から計算される事前確率、および 5. 節で述べた尤度を用いて各ノードの事後確率を計算し、下位の階層では上位階層の事後確率が閾値より高いノードに対してのみ事後確率を計算する。

提案手法では、木構造の最下層において最も高い解像度で状態空間の分割が行われていても、状態空間の分割を画像特徴距離に基づいて行っているため、腕が胴体に隠蔽されて見えてい

ない姿勢等、大きく異なる姿勢が同じノードに登録されている場合がある。

そこで、最下層  $L$  のノードに関する事後確率が閾値より大きいノードについて、ノードに含まれる各姿勢サンプルの事後確率を計算し、事後確率が最も高い姿勢サンプルを現在のフレームの推定結果とする。姿勢の木構造は、画像特徴距離に基づいて構成されているため、最下層の同じノード内の姿勢サンプルの尤度に関しては一定とみなし、ノードに関する尤度  $p(z_t | \theta_t^{jL})$  を用いる。

$$p(x_t^{njL} | z_{1:t}) = c_t p(z_t | x_t^{njL}) p(x_t^{njL} | z_{1:t-1}) \quad (12)$$

$$\sim c_t p(z_t | \theta_t^{jL}) p(x_t^{njL} | z_{1:t-1}) \quad (13)$$

ここで、 $x_t^{njL}$  は、最下層  $L$  のノード  $j$  に属する  $n$  番目の姿勢サンプル、 $p(x_t^{njL} | z_{1:t-1})$  は姿勢サンプル  $x_t^{njL}$  に関する事前確率で、次式のように計算できる。

$$p(x_t^{njL} | z_{1:t-1}) = \sum_{i=1}^{N_R} p(x_t^{njL} | x_{t-1}^i) p(x_{t-1}^i | z_{1:t-1}) \quad (14)$$

ここで、 $x_{t-1}^i$  は状態空間を構成する全ての姿勢サンプルで、 $N_R$  はその個数である。

このように、最も解像度の高い階層においても一意に定まらない姿勢は、画像特徴からは判断できないため、運動モデルによる姿勢の時間的連続性に基づいて現在の姿勢を定めることは妥当である。つまり、最下層の同一ノード内では、式 (4) の運動モデルにより事前確率分布のみ変化して、各姿勢サンプルについての事後確率分布が計算される。ただし、初期確率分布  $p(x_0) = p(x_0 | z_{1:0})$  は既知とする。初期確率分布に一樣分布を仮定するなど、一意な初期姿勢が与えられていない場合、初期フレームにおいて画像による観測から姿勢を一意に決められなければならない、あいまい性が残ったままとなる。本論文では、初期姿勢は正面向きの直立姿勢と仮定し、そのような姿勢を平均値とする正規分布を初期確率分布とする。

## 7. 実験

実画像を用いた追跡実験によって、本手法が有効に働くことを確かめる。以下の実験では、640×480 画素の解像度を持つ単眼の固定カラーカメラを使用し、画像入力時に 320×240 画素に縮小して姿勢推定処理を行う。姿勢サンプルの収集は、市販のジャイロ式モーションキャプチャシステムを用いて行い、主に立った状態で様々な動きを取得した。取得した姿勢の任意の 2 つの姿勢間で、関節角度の差が 5 度以下となるような姿勢がないように、類似姿勢を削除する。姿勢の木構造は 3 階層で、計算量を削減するため、上位階層から順に、80×60 画素、160×120 画素、320×240 画素の多重解像度を用いてシルエットの間の距離を計算している。本手法は、シルエット間の距離に基づいて姿勢推定を行うため、人体の三次元形状モデル (図 1(a) 参照) と、実際に追跡を行う人物の体型の違いによって追跡の精度が低下する。そこで実験では、洋服の JIS 規格 (JIS L4004 および JIS L4005) に基づいて、男性 10 体、女性 14 体の典型的な体型の標準人体モデルを使用し、それぞれについて



(a) Original image (b) Examples of estimated postures

Method	Fail	Error	Success
wXOR+DM1	24	0	0
wXOR+DM2	24	0	0
nwXOR+DM1	7	7	10
nwXOR+DM2	0	10	14

(c) Results of posture estimation

図 4 隠蔽される右腕の追跡結果

姿勢の木構造やシルエットを計算して姿勢辞書を構成しておく。追跡時には、追跡対象人物の身長と胸囲を用いて、最も体型の近い標準体型モデルを選択する。また、シルエットを安定に抽出するため、青い一様な背景を用いている。

まず、提案手法の有効性を検証するための実験を行う。図 4 のように、横を向いた状態で右腕を前後に振る動作を推定する。この実験に用いた姿勢サンプルは、主に図 4(a) のような横向きの姿勢で、右腕を様々に動かして取得した。姿勢サンプル数は 3,417、木構造ノード数は上位層から順に、693、3,013、3,318 ノードとなった。図 4(b)、4(c) に、24 往復右腕を振って往復追跡に成功 (Success) した回数、成功したが腕の角度が不正確 (Error) であった回数、失敗 (Fail) した回数を目視でカウントした結果を示す。従来の core-weighted XOR (wXOR) では一度も腕の往復動作の追跡に成功しなかったが、正規化した wXOR (nwXOR) により、胴体など他の部位に比べて細い部位である右腕の追跡性能が改善されていることが分かる。また、隠蔽を考慮しない運動モデル (DM1) に比べて隠蔽を考慮した運動モデル (DM2) を用いたほうが結果が良くなっており、隠蔽を考慮した運動モデルの有効性が分かる。これは、胴体に隠蔽されている間の右腕の運動 (ほぼ等速運動) が、4. 節で述べた運動モデル (前フレームの姿勢を平均値とする正規分布) に合致していないためである。

図 5 は、様々な動作を含む画像列に対して追跡を行った結果から、3 種類の動作に関する結果を切り出して表示したものである。この実験に用いた画像列の長さは 2 分間でフレームレートは 15 fps である。姿勢サンプル数は 57,136 姿勢で、シルエット距離に基づいて木構造を生成した結果、図 2 のようにノード数は上位層から順に、7,828、26,805、44,108 ノードとなった。図 5(a) は、カメラの光軸に垂直な軸周りの回転運動で、腕が隠蔽されたり、正面と背面の区別が難しい例であるが、正しく姿勢推定が行われている。図 5(b) は、ゴルフスイングをカメラに向かって行っている。カメラから遠い方の右腕が隠蔽され、また、カメラの光軸方向への運動が多く含まれるが、正しく推定できている。最後に図 5(c) は、他の部位に比べて細い右手だけを動かした例であるが、5.2 節で述べたシルエット距離により、



(a) Turn



(b) Golf swing



(b) Pointing

図 5 追跡結果 (上段は原画像に姿勢推定結果の人体モデルを重ね書きした画像, 下段は原画像.)

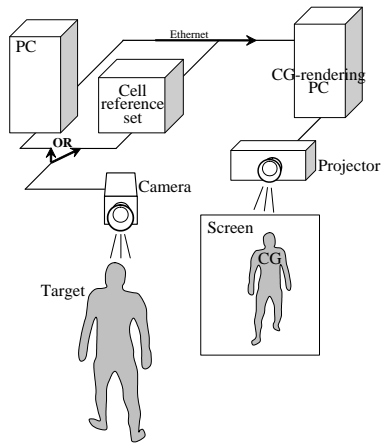
このような部位の動きも安定に推定できていることが分かる。

## 8. オンラインモーションキャプチャシステム

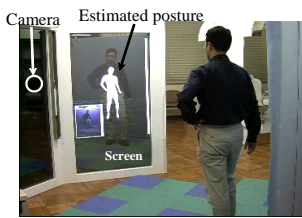
提案手法を用いて、図 6(a) のようなオンライン姿勢推定システムを構成した。カメラで取得した画像を、Opteron<sup>TM</sup> 280 を 2 基搭載したハイエンド PC、または Cell ブロードバンドエンジン (Cell) を用いたシステム (図 6(c) 参照) に入力して姿勢推定を行い、CG レンダリング用の PC に推定結果の関節角度データを送信して CG の人体モデルを描画し、大型ディスプレイに表示する。図 6(b) は、このシステムを用いてオンライン姿勢推定を行っている様子である。

ハイエンド PC には、Opteron<sup>TM</sup> 280 CPU が 2 基搭載されており、それぞれの CPU は 2 つの CPU コアを持ち 2.4 GHz で動作する。この 4 つの CPU コアを用いて、シルエット距離の計算をスレッド化して 4 並列で処理することにより高速化を行っている。その結果、1 フレーム平均 128 ms 程度の処理時間を達成している (図 6(d) 参照)。

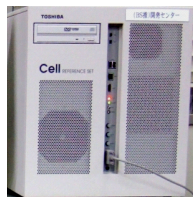
一方、本実験で用いた Cell は、3.2GHz で動作し管理用の 1 つの Power Processor Element (PPE) と高い演算性能を持つ 7 つの Synergistic Processor Element (SPE) からなる Many



(a) System organization



(b) System overview



(c) Cell reference set

	High-end-class PC			Cell processor		
	Min.	Max.	Ave.	Min.	Max.	Ave.
Capture	21	29	22	20	53	23
Estimation	21	236	104	17	169	62
Total	43	261	127	39	202	86

(d) Processing time [ms/frame]

図6 オンライン姿勢推定

core プロセッサである。この Cell を用いたシステム (図 6(c) 参照) では、背景差分およびシルエット距離計算の各処理を 7 つの SPE で並列に行うことにより高速化を行い、1 フレームあたりの平均処理時間は、86 ms を達成している。

## 9. まとめ

単眼カメラを用いてマーカー等を必要としない人体全身の姿勢推定を行うため、画像特徴であるシルエットの差に基づいて生成した姿勢の木構造を用いて、効率的に姿勢推定を行う手法を提案した。腕などの部位の隠蔽を考慮した運動モデルを用いることにより、隠蔽部位の追跡を安定化した。また、部位の太さに対して均等な重みを用いた重み付排他論理和によるシルエット距離を提案した。これにより、細い部位でも姿勢推定が安定に行えるようになった。実験によって提案手法の有効性を確かめ、ハイエンドクラスの PC および Cell プロセッサを用いてオンラインモーションキャプチャシステムを実現し、それぞれ 1 フレームあたりの平均処理時間 127 ms, 86 ms 程度の速度で処理できることを確かめた。

今後の課題としては、エッジ、速度といったその他の画像特徴を使用して認識精度を高めることなどがある。

謝辞 Cell プロセッサへの移植に協力頂いた、(株) 東芝 研究

開発センターの横井謙太郎主事、同 Semiconductor 社の福田悦生参事、近藤伸宏主事、檜田和浩主事、(株) フィックスターズの飯塚博通氏、深野佑公氏に深く感謝いたします。

## 文 献

- [1] D. Gavrila and L. Davis: "3D model-based tracking of humans in action: A multi-view approach", Proc. of CVPR, pp. 73-80 (1996).
- [2] R. Plänkers and P. Fua: "Tracking and modeling people in video sequences", Computer Vision and Image Understanding, **81**, (2001).
- [3] A. Agarwal and B. Triggs: "3D human pose from silhouettes by relevance vector regression", Proc. of CVPR, Vol. 2, pp. 882-888 (2004).
- [4] Q. Delamarre and O. Faugeras: "3D articulated models and multi-view tracking with silhouettes", Proc. of ICCV, Vol. 2, pp. 716-721 (1999).
- [5] M. Brand: "Shadow puppetry", Proc. of ICCV, pp. 1237-1244 (1999).
- [6] A. Senior: "Real-time articulated human body tracking using silhouette information", Proc. of IEEE Workshop on Visual Surveillance/PETS, pp. 30-37 (2003).
- [7] M. Yamamoto, Y. Ohta, T. Yamagiwa, K. Yagishita, H. Yamana and N. Ohkubo: "Human action tracking guided by key-frames", Proc. of FG, pp. 354-361 (2000).
- [8] J. Deutscher, A. Blake and I. Reid: "Articulated body motion capture by annealed particle filtering", Proc. of CVPR, Vol. 2, pp. 1144-1149 (2000).
- [9] C. Sminchisescu and B. Triggs: "Estimating articulated human motion with covariance scaled sampling", IJRR, **22**, 6, pp. 371-391 (2003).
- [10] P. Felzenszwalb and D. Huttenlocher: "Efficient matching of pictorial structures", Proc. of CVPR, Vol. 2, pp. 66-73 (2000).
- [11] N. Date, et al.: "Real-time human motion sensing based on vision-based inverse kinematics for interactive applications", Proc. of ICPR, Vol. 3, pp. 318-321 (2004).
- [12] A. Thayananthan, R. Navaratnam, B. Stenger, P. H. S. Torr and R. Cipolla: "Multivariate relevance vector machines for tracking", Proc. of ECCV, Vol. 3, Graz, Austria, pp. 124-138 (2006).
- [13] B. Stenger, A. Thayananthan, P. H. S. Torr and R. Cipolla: "Filtering using a tree-based estimator", Proc. of ICCV, Vol. 2, pp. 1063-1070 (2003).
- [14] D. M. Gavrila: "Pedestrian detection from a moving vehicle", Proc. of ECCV, Vol. 2, pp. 37-49 (2000).
- [15] C. F. Olson and D. P. Huttenlocher: "Automatic target recognition by matching oriented edge pixels", Transactions on Image Processing, **6**, 1, pp. 103-113 (1997).
- [16] 島田, 白井, 久野: "確率に基づく探索と照合を用いた画像からの手指の三次元姿勢推定", 信学論, **J79-D-II**, 7, pp. 1210-1217 (1996).
- [17] H. Sidenbladh, M. Black and D. Fleet: "Stochastic tracking of 3D human figures using 2d image motion", Proc. of ECCV, pp. 702-718 (2000).
- [18] Y. Chen, J. Lee, R. Parent and R. Machiraju: "Markerless monocular motion capture using image features and physical constraints", Proc. of Computer Graphics International, pp. 36-43 (2005).
- [19] A. Thayananthan, B. Stenger, P. H. S. Torr and R. Cipolla: "Learning a kinematic prior for tree-based filtering", Proc. of BMVC, Vol. 2, pp. 589-598 (2003).