社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
PRMU2000-230 (2001-03)

多視点カメラを用いた色領域追跡に基づく3次元人体動作の実時間推定

米元 聡　　星野 竜也　　有田 大作　　谷口 倫一郎

九州大学大学院システム情報科学研究院
〒816-8580 福岡県春日市春日公園 6–1
TEL: 092-583-7618
{yonemoto,hoshino,arita,rin}@limu.is.kyushu-u.ac.jp

あらまし　実生活環境において誰しもが利用可能な仮想環境インタインタフェースを構築するためには, 機器の装着を強いられることなく, しかも実時間で人間の3次元の動作を計算機へ入力する手段が必要となる. 本研究では, センサとしてカメラを用いて人間の頭部や手の3次元位置を追跡する手法について提案する. 色情報を用いて人体部位の領域の追跡を行うが, その際, 変形を許容した楕円形状モデルによって領域追跡の精度の向上を試みている. また, 複数のカメラを用いた計測範囲の拡張についても示す.

キーワード　実時間多視点動画像処理　動作推定　ヒューマンインタフェース　逆運動学

A Real-time Human Motion Traking System Using Multiple Cameras

Satoshi YONEMOTO, Ryuya HOSHINO, Daisaku ARITA and Rin-ichiro TANIGUCHI

Graduate School of Information Science and Electrical Engineering, Kyushu University
6-1, Kasuga-koen, Kasuga, Fukuoka 816-8580 JAPAN
+81-92-583-7618
{yonemoto,hoshino,arita,rin}@limu.is.kyushu-u.ac.jp

Abstract　*This paper presents a real-time human motion tracking using skin-color-based Gaussian blobs and human motion synthesis based on real-time inverse kinematics. Our purpose is to do seamless mapping of human motion in the real world into virtual environments. In general, virtual environment applications such as 'smart' interface require real-time human motion tracking system without special devices or markers. However, since such vision-based human motion tracking system is essentially unstable and can only acquire partially observable information, visual features must be estimated robustly with occlusion and with adjacent influence. In this paper, we demonstrate a real-time and on-line real-virtual interaction system which realizes human figure motion synthesis from limited perceptual cues.*

key words　real-time video processing, human interface and inverse kinematics

1 はじめに

近年は"スマートルーム"[1] など, 接触型の器具を用いることなく, ビデオカメラを用いて人間の行為を知覚するような知的なサイバースペース生成の研究が進んでいる. これは非接触な方式を実現することが, コンピュータビジョンの達成すべき研究課題であるばかりでなく, 仮想空間への入口となる計測空間と日常の生活空間とをシームレスに繋げるために有効と考えられているからであり, 本研究もこのようなセンサを装着することによる制約を打破する非接触型の知覚系の構築を目指すものである.

1.1 関連研究

ビデオカメラで撮影した映像より人間の動作を推定する非接触型の方法は, 従来より盛んに研究が進められてきた. 特に, 人体に関する詳細な形状・動き情報を推定する方法としては, オフラインで獲得された画像系列に対しモデルベースで解析する方法 [2][3] がある. しかし処理時間が膨大であることや, モデル像と画像のマッチングを行なうに必要な精度である初期情報は通常オフラインでしか得ることができないという制約のため, 実時間, オンラインでの適用には課題が残っていた. 一方, 近年の計算機能力の向上により, オンラインでの画像獲得およびその解析がようやく可能となりつつあるため, 今日では画像を詳細に解析するアプローチとは異なり, 比較的単純かつ高速な画像解析手法を用い, 実時間性を重視したものが開発され始めている. ほぼ実時間で動作する人体動作解析システムには大別して色情報を用いるもの, 輪郭情報を用いるものの 2 つのアプローチがあり, 具体的には以下のようなものが提案されている.

色情報を用いる手法 *Pfinder*[1] は, 一般的な領域クラスタリング手法を用いて 2 次元の blob 特徴, すなわち人体を推定するための画像特徴としての領域を検出する. blob 特徴とは, 均一な色領域の重心およびその形状のモーメントで定義される特徴である. しかし, 人体の部位を領域に対応付けるためのアルゴリズムや, シーンの制約条件が不明であり, 人体を自動に推定するためには何らかの事前情報を必要とする. 実際のアプリケーションでは, 抽出した 2 次元の *blob* 特徴をもとに, 2 視点からのステレオ計測により 3 次元情報の推定を行なっている. 比較的安価なシステム構成でありながら実時間に近い速さで動作可能であり, 非接触方式による比較的精度の低い抽出結果でも多くのアプリケーションが十分に機能することを示した.

最近 *Pfinder* をベースに, システムに物理法則に基づいた制御機構を導入することで, 抽出結果の不安定さを回避し, リアリティの高い上半身動作の再現を達成することが可能な方法が開発された [4]. 予め想定した動作に対しては良好な再現結果を得ることができるが, 全身動作のように多くの動作表現を行うためには工夫が必要である.

シルエット輪郭情報を用いる手法 シルエット輪郭から人体情報を解析し, その行動, 人数などを認識する手法 [5]. このシステムは人間のモニタリングシステムであり, 人体モデルの再現を目的としていないため, 連続的な人体動作の再現には適していない.

1.2 人体動作推定の問題

上記の関連研究より, なるべく制約のない, リアリティの高い人体動作の再現を行うシステムの開発が課題であると言え, 本研究では人体動作の再現を行うために, 以下の課題を実現するシステムを提案する.

- 少ない画像特徴からリアリティの高い人体動作を再現する.
- 実時間で動作可能であり, ユーザ介在の後処理を必要としないなどオンライン性がある.
- 計測範囲を拡張するため, 多視点情報を利用する.

また, 本システムでは前述の *Pfinder* システムと同様に, 画像特徴として色領域 blob を用いる. その理由は, 画像上で手先や足先の追跡が他の画像特徴に比べ安定で, 抽出精度の人体姿勢依存性が低いと考えるからである.

2 Blob 特徴の追跡

以下では, 手や顔に相当する色領域 blob を追跡し, その 3 次元位置を求める方法について述べる.

2.1 領域の検出

背景差分と外接矩形の検出 前処理として, 背景差分を施し, 人体領域を囲む外接矩形を検出する.

色識別 本手法では, 入力画像中に観測される肌色領域は手, 顔のいづれかの領域であると解釈される. 色識別の方法に, 照明変化に比較的ロバストな方法として, パラメトリックな色モデルによる識別手法を用いている [6]. 各画素の色特徴 (r,g,b) が, 次のような濃淡値 i をパラメータとした 2 次形式で表せると仮定する.

$$R = R_2 i^2 + R_1 i, G = G_2 i^2 + G_1 i, B = B_2 i^2 + B_1 i \quad (1)$$

ここで, 6 つのモデルパラメータ係数 $R_1, \cdots, B_2$ は, 予め学習用実画像から推定したものである.

色識別においては, 以下の式により, 各画素について観測される色特徴 (r,g,b) とモデル色特徴間のマッチングを行い, しきい値処理を行う.

$$error = (\hat{r}-\hat{R})^2 + (\hat{g}-\hat{G})^2 \tag{2}$$

ここで, $\hat{r} = r/(r+g+b)$, $\hat{g} = g/(r+g+b)$, $\hat{R} = R/(R+G+B)$, $\hat{G} = G/(R+G+B)$.

2.2 Blob の定義

領域特徴の 1 次モーメントまで考慮したガウシアン blob 特徴は, 楕円形状モデル $\mathbf{e} = (\mu, \mathbf{\Sigma})$ である (平均 μ, 共分散 $\mathbf{\Sigma}$). しかし, 肌色領域のように, 同じ色特性の領域を複数個追跡する場合, このモデルでは画素単位での識別に誤りが生じる恐れがある. 領域追跡を正確に行うためには, より正確な領域形状, 領域境界を把握する必要がある.

そこで, 本研究では, blob 特徴に正確な領域境界を表現可能なように局所的な変形 $\mathbf{d}$ を追加する.

$$\mathbf{p}(s) = \mathbf{e}(s) + \mathbf{d}(s) \tag{3}$$

ここで, s は楕円周のパラメータを表す. 以下, 「blob」を, 式 (3) の $\mathbf{p}$ と定義する.

2.3 画素の識別

初期ラベリング まず, 外接矩形内の領域を各ガウシアン blob 領域へとクラスタリングする. どの部位に相当するか決定する尺度としては, 前フレームまでに推定された 2 次元重心位置 (あるいは予測位置) との近接性のみを用いる. これにより, 初期の推定位置 (μ', $\mathbf{\Sigma}$') が求まる.

変形楕円による輪郭境界の探索 次のステップは, 初期推定位置 (μ',$\mathbf{\Sigma}$') を用いて領域境界を決定する, すなわちどの画素が各 blob に属するかを厳密に決定することである. 境界探索のための探索領域を, 初期楕円の周上に配置する (図 1(上) 参照). 各探索点における局所窓 W から, 尤もらしい領域境界を決定する. 通常, 境界を決定する尺度としてエッジ勾配などが用いられるが, 本手法では, 衣服のエッジの影響も考えられるため, 以下のような色の分離境界を定める分離度 $J(\mathbf{p})$ をその尺度としている.

$$J(\mathbf{p}) = \left| N_W/2 - \sum_{(x_i,y_i)\in W} p(x_i,y_i|skin) + \varepsilon \right|^{-1} \quad if \sum_{(x_i,y_i)\in W} n(x_i,y_i) > 0 \tag{4}$$

ここで, $p(x_i,y_i|skin)$ は 肌色の尤度, N_W は W 内の画素数, $n(x_i,y_i) = \mathbf{p}_i\vec{\mathbf{p}} \cdot \mu\vec{\mathbf{p}}$(図 1(下) 参照). したがって, $\mathbf{d}(\mathbf{s}) = \arg\max J(\cdot)$.

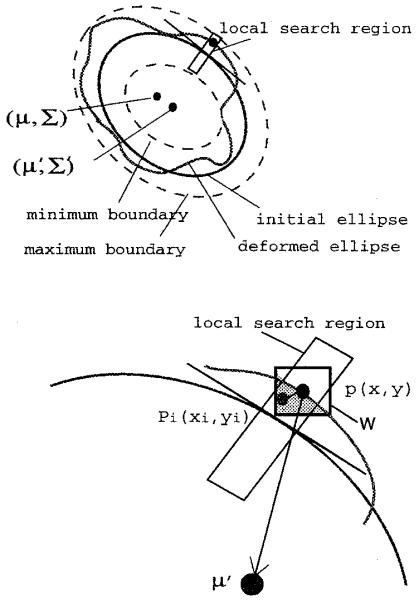

図 1: (上) 変形楕円モデル. (下) 境界探索の領域.

推定した変形楕円内で対応画素の重心 μ を計算することにより, blob の 2 次元位置を求める.

2.4 3 次元位置の復元

2 視点の画像から blob 特徴が観測される時, その 3 次元位置をステレオ視により計算することが可能である. 予め得られるキャリブレーション情報により, カメラ座標原点と blob 中心位置を結ぶ視線を各視点について計算し, それら視線の交差位置を求める.

視線 1 が, $\mathbf{T}_1 = \mathbf{o}_1 + t_1\mathbf{d}_1$ と表現され (t_1 は媒介変数, $\mathbf{o}_1$ はその視点のカメラ座標原点, $\mathbf{d}_1$ はその単位視線ベクトル), 視線 2 が, $\mathbf{T}_j = \mathbf{o}_2 + t_2\mathbf{d}_2$($t_2$ は媒介変数, $\mathbf{o}_2$ はカメラ座標原点, $\mathbf{d}_2$ は単位視線ベクトル) と表現される時, 視線 1 上にあり, 視線 2 から最小距離である点 $\mathbf{T}$ は, t_1 をパラメータとして求めると以下のようになる.

$$\mathbf{T} = \mathbf{o}_1 - \frac{(\mathbf{d}_1 \times \mathbf{m}_2, \mathbf{o}_1 \times \mathbf{m}_2 - \mathbf{n}_2)}{||\mathbf{d}_1 \times \mathbf{m}_2||^2}\mathbf{d}_1 \tag{5}$$

ただし

$$\mathbf{m}_j = \frac{\mathbf{d}_j}{\sqrt{1+||\mathbf{o}_j \times \mathbf{d}_j||^2}} \quad , \quad \mathbf{n}_j = \frac{\mathbf{o}_j \times \mathbf{d}_j}{\sqrt{1+||\mathbf{o}_j \times \mathbf{d}_j||^2}}.$$

この点 $\mathbf{T}$ が 3 次元 blob 位置である.

2.5 隠れの検出

人体領域の追跡においては, 顔領域が一定位置にあり, 手領域が顔領域を隠す状況や, 片方の手領域がもう一方の

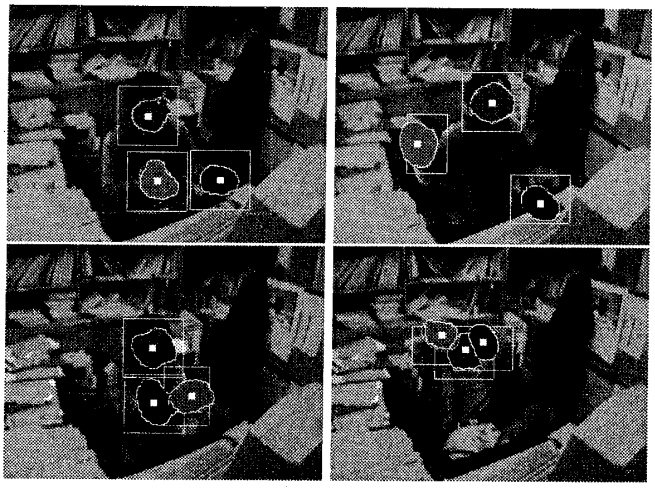

図 2: 推定した変形楕円領域の例.

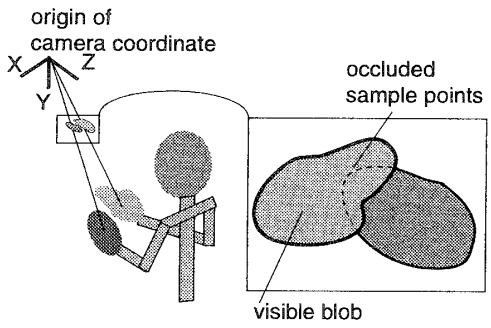

図 3: 隠れの検出.

手領域を一時的に隠し, 横切るというような状況が起こり得る. この場合, blob 領域の可視性が重要である. 推定したモデルの 3 次元予測位置を再投影することにより, どの blob が可視であるかを判定することが可能である. このモデルベースの予測を「画像生成による解析」[2] と呼ぶ. 具体的には, まず, カメラ座標原点からの距離を Z バッファ法により計算し, 次フレームでの blob の可視性を求める. それをもとに, 隠れが起こっている blob についてはその影響を最小限にするため, 画素識別の際に隠れた境界の推定を行わず, 初期楕円弧に境界を固定するように制限する (図 3参照).

3 人体モデルの動作生成

3.1 逆運動学

上記追跡処理により得られるのは頭部・両手などの 3D blob 位置のみである. 限られた少数の知覚データを用いて人体モデルの姿勢を生成するため, 以下の逆運動学解法 (インバースキネマティクス) を用いてこの問題を解決する [6]. なお人体モデルの表現としては, 知覚データより求まる位置に両膝・両肘を加えた部位数 14, 自由度 23(詳細は後述) を考えた多関節構造モデルとする (図 4).

3.2 モデル当てはめ

まず逆運動学におけるゴールとして様々なものが考えられるが, 我々のような目的の場合, 入力される 3-D blob 位置が正確でないことが起こりうる. 特に, ゴール位置が解の存在範囲を越えて設定された場合, 連続性が満たされなくなる. よって, 実際のゴールを方向およびその距離 (絶対値) と定義することで対処する. つまり, ゴールの方向は必ず満たし, その距離は設定ゴール位置が解の存在範囲内である時のみ 3-D blob 位置に一致させる (図 4 左上).

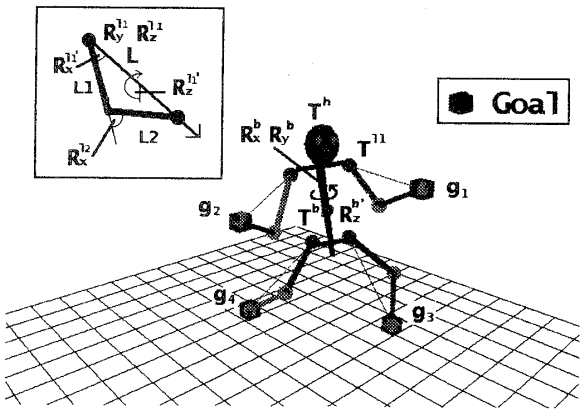

図 4: 人体モデルの構造およびゴール方向の定義

実際には以下のような, 2 つの可変リンクに関する運動学方程式を解析的に解く.

$$\begin{aligned}\mathbf{g}_i = \quad & \mathbf{T}^b\mathbf{R}^b(0, {R_y}^b, {R_x}^b)\mathbf{R}^{b'}({R_z}^{b'}, 0, 0) \\ & \mathbf{T}^{l_1}\mathbf{R}^{l_1}({R_z}^{l_1}, {R_y}^{l_1}, 0)\mathbf{R}^{l'_1}({R_z}^{l'_1}, 0, {R_x}^{l'_1}) \\ & \mathbf{T}^{l_2}\mathbf{R}^{l_2}(0, 0, {R_x}^{l_2})\ \mathbf{t}^e\end{aligned} \quad (6)$$

ここで, 左辺はゴール $i(1, \cdots, 4)$ を表す位置ベクトル $\mathbf{g}_i = (g_x, g_y, g_z, 1)^T$, $\mathbf{T}^b$ および $\mathbf{R}^b$, $\mathbf{R}^{b'}$ は胴体に関する相対並進, 回転行列, $\mathbf{T}^{l_1}$ および $\mathbf{R}^{l_1}$, $\mathbf{R}^{l'_1}$, $\mathbf{T}^{l_2}$ および $\mathbf{R}^{l_2}$ はリンク 1, 2 に関する相対並進, 回転行列である ($'$ のものは解析的に解くために分けられた回転行列に相当する). $\mathbf{t}^e$ はリンク 2 の末端のゴール位置への局所座標系での並進ベクトルである. また $\mathbf{R}(R_z, R_y, R_x)$ の回転自由度 R_z, R_y, R_x はそれぞれロール・ピッチ・ヨーを表す (解析解の詳細は文献 [6] 参照のこと).

4 実験結果

4.1 システム構成

実時間・オンラインでの動作を実現するため, 本手法を PC クラスタ上に実装した. 本実験においては, PC クラスタを, 高速ネットワーク Myrinet で接続された最大 8 台の PC(700MHz × 2 の CPU) と 6 台の IEEE1394 ベースのデジタルカラービデオカメラ SONY DFW-V500 により構成している.

以下の実験では, 2〜6 台の PC が画像獲得・2 次元画像処理を行い, 1 台の PC が 2 次元画像処理を行った PC からの情報を受信し, 3 次元の復元処理, 人体モデルの動作生成, 人体モデルの描画を行う. さらに 1 台の PC が 2 次元画像処理を行う PC に同期信号を送信し, PC 間の同期を管理する (図 5).

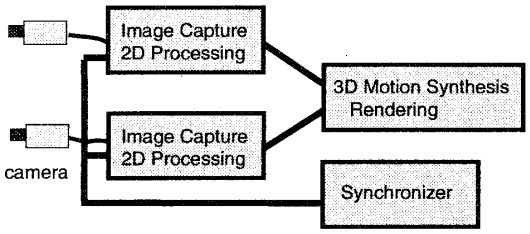

図 5: 多視点動画像処理の PC クラスタへの実装.

4.2 デスクトップインタフェースへの適用

肌色情報のみを用いた人体動作再構成システムとして, 上半身の動作の再現を行うデスクトップ型の 3 次元操作用インタフェースを構築した (図 6). このようなアプリケーションでの利用においては, 手と顔の重なりや交差が問題となるが, 本手法により, ある程度誤対応を防ぐことができた.

4.3 多視点カメラの利用による計測範囲の拡張

本手法においては, 人体動作の計測は正面像の対に対してのみ適用可能である. したがって, 多視点カメラを用いて正面像の選択を行うことができれば, 計測範囲の拡張を行うことが可能となる. しかしながら, 高精度に人体の向き (パン角) を求めるには胴体を正確に測定する必要があり, 現在のアプローチでは困難である. そこで本実験では簡単に, 両足先のなす角度により, 正面像の視点対を選択することを考える. その詳細を以下に述べる: 鉛直上方からの仮想カメラ座標を想定し, 左右足先の投影点 $\mathbf{p}_l$, $\mathbf{p}_r$ を結ぶ 2 次元ベクトルと, 各視点のカメラ座標原点の投影点 $\mathbf{p}_c$, ワールド座標原点の投影点 $\mathbf{p}_w$ を結ぶ 2 次元ベクトルを求める. それらを正規化し, ベクトルのなす角度を

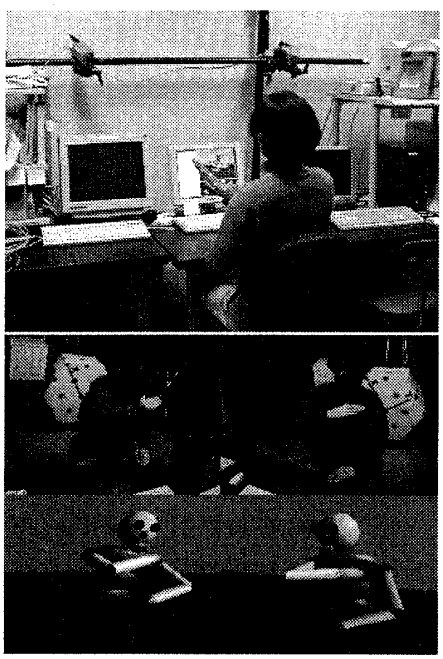

図 6: デスクトップインタフェースへの適用: (上) システム構成. (下) 入力画像と同一視点からの推定モデル像.

計算することにより, 最適なすなわち直角に近い角度の隣り合う視点対を選択する.

本実験では, 6 つのカメラを同等の視野を保ちつつ 60 度の均等間隔で円周上に配置し (図 7), 直立状態の動作に関しては, どの向きでも正面像を観測可能な観測環境を想定した. 図 8に, その入力画像および人体モデルの再構成像を示す[1]. また, 選択された視点対を太枠で示している. 再投影による blob の予測位置を用いているので, 滑らかな視点の切替えを行うことができ, 推定位置の変動の影響は少ないことが確認できた. 6 つのカメラ程度でも, あらゆる人体の向きについての安定な正面像を選択可能であった.

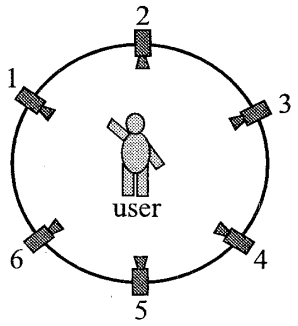

図 7: 6 視点カメラの均等配置.

[1]実時間処理中に撮影したショットである.

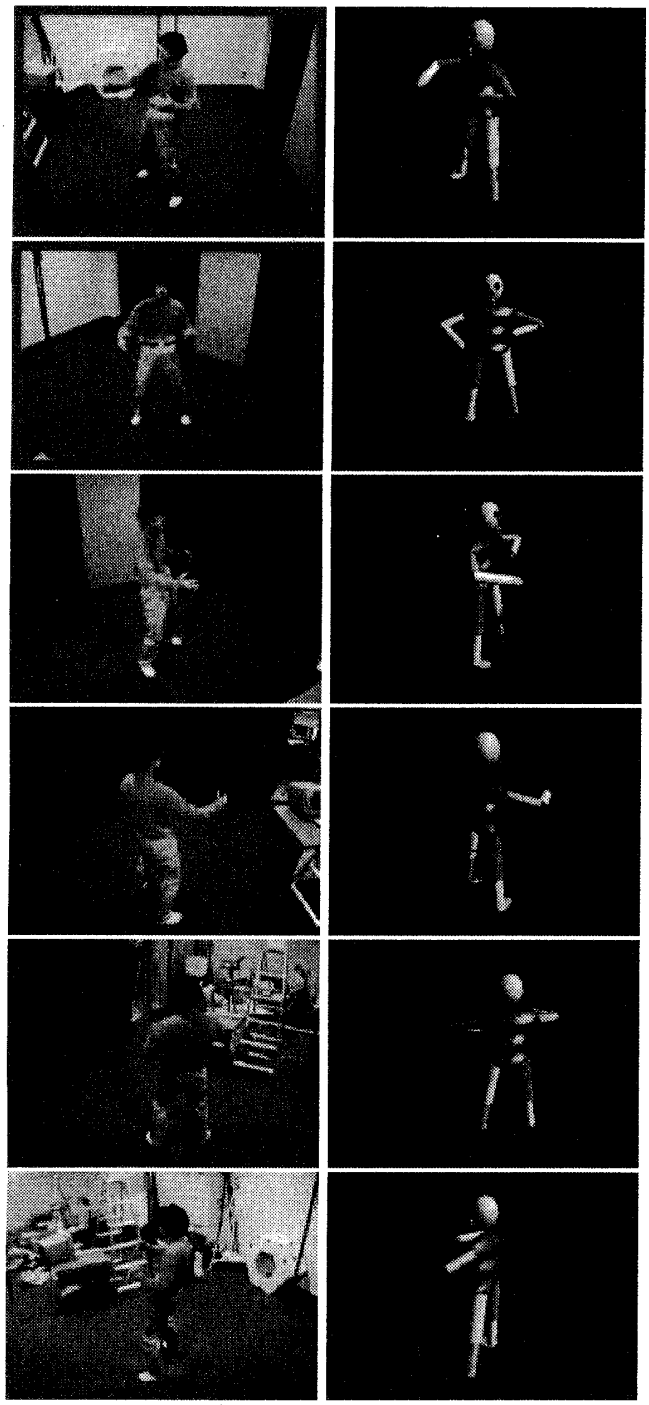

図 8: オンライン視点選択による全身動作の再構成の例 (左: 入力画像 (視点 1-6). 右: 同一視点からの人体モデル再構成像).

5 おわりに

本論文では多視点動画像処理による非接触型の実時間動作推定システムを提案し, 複数 blob 存在下での安定な追跡方法, 小数の知覚データに対する実時間逆運動学の導入による人体動作の再生方法について述べた. 本システムはオンラインかつ実時間で動作するため, 仮想空間とのインタラクションなど様々なアプリケーションへの適用が可能である. また, 特殊なマーカなど装着不要な高度なモーションキャプチャシステムであるため, その応用範囲は広いと考えられる. さらに全周に配置した多視点カメラを用いることによって, 計測範囲の拡張を行い, 正面像を対象とする手法の適用制限を緩和することができた. 本システムに力学的あるいは感性的な動作フィルタリングを施すことによって, より人間らしい動作の生成や, 知覚データに大きく影響されない安定な動作の再現が可能となると考えられるが, これは今後の課題である.

参考文献

[1] C.Wren, A.Azarbayejani, T.Darrell, A.Pentland, "Pfinder: Real-Time Tracking of the Human Body", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.19, No.7, pp.780-785, 1997.

[2] Satoshi Yonemoto, Naoyuki Tsuruta, Rin-ichiro Taniguchi, Tracking of 3D Multi-Part Objects Using Multiple Viewpoint Time-varying Sequences, in Proceedings of 14th International Conference on Pattern Recognition(ICPR'98), pp.490–494, 1998.

[3] C.Bregler, J.Malik, "Tracking People with Twists and Exponential Maps", in *Proc. of Computer Vision and Pattern Recognition*, pp.8–15, 1998.

[4] C.Wren, A.Pentland, "Understanding Purposeful Human Motion", *in Fourth IEEE International conference on Automatic Face and Gesture Recognition*, 2000.

[5] Haritaoglu, I. and Harwood, D. and Davis, L.S., *W4S: A real-time system for detecting and tracking people*, in Proceedings of International Conference on Computer Vision and Pattern Recognition(CVPR'98),pp.962, 1998.

[6] Satoshi Yonemoto, Daisaku Arita and Rin-ichiro Taniguchi, Real-Time Human Motion Analysis and IK-based Human Figure Control, in Proceedings of Workshop on Human Motion(HUMO2000), pp.149-154, 2000.