コンピュータビジョンと
イメージメディア 114－6
（1999．1．21）

多視点動画像を用いた腕の動きのモデリングに関する研究

相馬 俊一　　長橋 宏
東京工業大学工学部像情報工学研究施設

Abstract

近年，人間動作の動画像追跡は様々な分野で要求が高まりつつある．従来の動画像からの人体動作の追跡では人物モデルを画像に投影する手法が主流であったが，オクルージョンによる追跡精度の低下や，処理が繁雑になるといった問題点があげられている．そこで，本研究では人物画像を3次元空間へ逆投影し，人物モデルとのマッチングをとるという手法を選択することで，処理の単純化をはかる．さらに，カルマンフィルタを効率良く用いることでより精度の高い人体の運動推定を行う手法を提案する．本研究では人体の運動追跡の一環としてマンマシンインタフェイスとしても応用範囲が広いと考えられる腕の運動追跡を扱う．シミュレーション実験により提案手法が人体の運動追跡の精度向上に有効であることを示す．

Estimation of Human Arm Movement using Multiple Image Sequences

Syunichi SOMA　　Hiroshi NAGAHASHI
Imaging Science and Engineering Labratory,
Faculty of Engineering, Tokyo Institute of Technology.

Abstract

The tracking of human motion from image sequence is important for man-machine interface technique or gesture recognition . In previous studies, the tracking of human motion is performed by projecting 3D human-model to images. However, these methods have some problems, for example, tracking error by occlusion or troublesome matching processes. In this paper, to overcome these problems, we construct human motion volume data from multiple viewpoint images, and match 3D human-model with the volume data frame by frame. This process enables simple matching. Then, to improve our tracking results we incorporate our system within a Kalman filter framework. We treat a human arm motion for application of man-machine interface. We presented experimental results of the 3D motion tracking.

1　はじめに

近年，ＰＣに代表されるようにハードウェアの高性能化もあいまって，テレビ，放送，教育，シミュレーションなど，人間活動のさまざまな分野で人物像がコンピュータグラフィックスによって表現されてきている．ＣＧでアニメーションを作成する際，人体の動きを何らかの形で運動データとして獲得し，コンピュータ内部に取り込む必要が生じる．動作関数を定義して運動を記述する試みも見られるが，実際の人物の動きをアニメーションで表現できる方がより幅広い応用が期待できると考えられる．現在まで，さまざまな方法で人体の動きを推定する方法が提案されてきている．[1][4][7][8][10][11] 従来の人体の動きのモデリング，または人体の動きの認識に関する研究では，データグローブやデータスーツを扱うものがあったが，被験者に負担がかかり，自然な動きを計測することは容易ではない．そのため，非接触で計測する方法として，画像を用いて計測する方法がある．この方法は被験者に対する負担も軽く有効であると考えられる．近年，3台以上のカメラを用いた多眼視による方法が提案されているが，[7]cite マルチカメラ [9] 撮影画像または撮影画像に画像処理を施した画像と人体モデルとを，2次元空間上でマッチングをとることによって人体の動きのモデリングをしている．しかし，3次元の情報を持つ人体

モデルを2次元空間上に投影し，2次元空間内でマッチングをとることは，もともとある情報を欠落させることとなる．そこで本研究では，3次元の情報を持つ人体モデルと3次元データを，3次元空間上でマッチングする手法を提案する．3次元空間上でマッチングをとることによって，通常の2次元画像にモデルを投影する手法に比べてモデルとのマッチングを容易にとることが可能となる．本稿ではこの手法の応用例として，マンマシンインタフェイスとしても応用範囲が広いと考えられる腕の運動追跡実験を行った．また，追跡の誤差についても検討を行い，誤差の分散をあらかじめ計測しカルマンフィルタを用いることで追跡制度の向上が行われることを示す．

2 ボクセルデータ列の獲得

2.1 カメラキャリブレーション

ワールド座標系とスクリーン座標系との対応に透視変換モデルを適応すると，ワールド座標系の同次座標系表現の点 $(X,Y,Z,1)$ とスクリーン座標系の点 (X_c,Y_c) の対応は，式(1)で表せる．式中の 3×4 の行列 C を，カメラパラメータと呼び，カメラパラメータを求めることをカメラキャリブレーションとよぶ．以下，カメラキャリブレーションの手法について述べる．[3]

$$\begin{bmatrix} H_cX_c \\ H_cY_c \\ H_c \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} \\ C_{21} & C_{22} & C_{23} & C_{24} \\ C_{31} & C_{32} & C_{33} & C_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{1}$$

まず，式(1)を展開することで式(2)を得ることができる．

$$\begin{cases} C_{11}X + C_{12}Y + C_{13}Z + C_{14} \\ -C_{31}XX_c - C_{32}YX_c - C_{33}ZX_c - C_{34}X_c = 0 \\ C_{21}X + C_{22}Y + C_{23}Z + C_{24} \\ -C_{31}XY_c - C_{32}YY_c - C_{33}ZY_c - C_{34}Y_c = 0 \end{cases} \tag{2}$$

式(2)において，$C_{34}=1$ とし，行列表現に書き換えることによって式(3)を得ることができる．

$$AC = R \tag{3}$$

ただし

$$A = \begin{bmatrix} X\ Y\ Z\ 1 & 0\ 0\ 0\ 0 & -XX_c\ -YX_c\ -ZX_c \\ 0\ 0\ 0\ 0 & X\ Y\ Z\ 1 & -XY_c\ -YY_c\ -ZY_c \end{bmatrix}$$

$$C = \begin{bmatrix} C_{11} \\ C_{12} \\ \vdots \\ C_{33} \end{bmatrix}$$

$$R = \begin{bmatrix} X_c \\ Y_c \end{bmatrix}$$

式(2)は，ワールド座標系の点 (X,Y,Z) とスクリーン座標系の点 (X_c,Y_c) の対応関係が1組与えられると，カメラパラメータを算出するのに必要な11元方程式が2本得られることを意味している．この場合 C の解を得るためには同一平面上にない6組以上の対応関係を与えることが必要となる．$n(n\geq 6)$ 点の対応関係が与えられると，式(4)が得られる．

$$A_mC = R_m \tag{4}$$

ただし、

$$A_m = \begin{bmatrix} X_1\ Y_1\ Z_1\ 1\ 0\ 0\ 0\ 0\ -X_1X_{c1}\ -Y_1X_{c1}\ -Z_1X_{c1} \\ 0\ 0\ 0\ 0\ X_1\ Y_1\ Z_1\ 1\ -X_1Y_{c1}\ -Y_1Y_{c1}\ -Z_1Y_{c1} \\ \cdots \\ X_n\ Y_n\ Z_n\ 1\ 0\ 0\ 0\ 0\ -X_nX_{cn}\ -Y_nX_{cn}\ -Z_nX_{cn} \\ 0\ 0\ 0\ 0\ X_n\ Y_n\ Z_n\ 1\ -X_nY_{cn}\ -Y_nY_{cn}\ -Z_nY_{cn} \end{bmatrix}$$

$$R_m = \begin{bmatrix} X_{c1} \\ Y_{c1} \\ \vdots \\ X_{cn} \\ Y_{cn} \end{bmatrix}$$

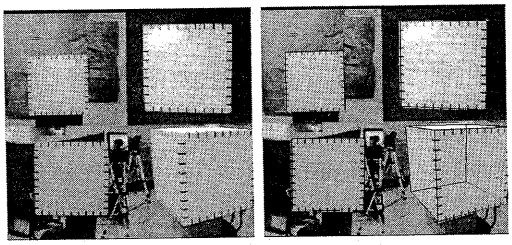
式(4)からカメラパラメータは次のように求められる．

$$C = (A_m^tA_m)^{-1}A_m^tR_m \tag{5}$$

図1に，カメラキャリブレーションに用いた画像を示す．図1(a)のように基準立方体を撮像し，画像上の点と基準立方体上のワールド座表系の点の対応をとることで各カメラに対するカメラパラメータが求められる．求められたカメラパラメータを用いて一辺50cmの立方体を画像上に投影した様子を図1(b)に示す．

2.2 ボクセルデータの獲得

本稿では，透視変換を用いているため，人物領域をボクセル空間中に逆投影することによって，人物領域のクサビ型の3次元的な領域をボクセル空間に得ることができる．しかし，カメラ1台だけでは3次元的な人物領域を特定できない．複数台のカメラ



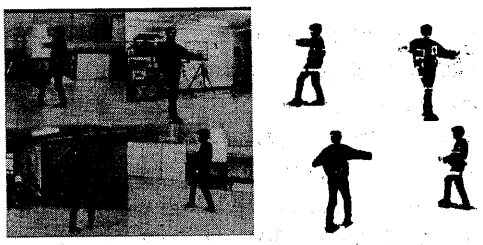
(a) 基準立方体　(b) 立方体を投影した様子

図 1: カメラキャリブレーション

を用いて，全てのカメラの人物領域をボクセル空間中に逆投影し得られる領域の共通部分を抽出することにより，人物領域が絞られてくる．複雑に入り組んだ人物を完全に再現することは不可能であるが，カメラの台数が多くなるほど人物の存在領域が絞られてくることになり，より実際の人物に近いボクセルデータが得られる．

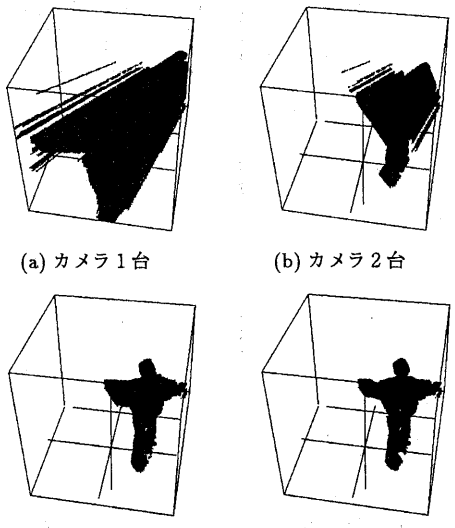
人物領域の抽出は，背景画像との差分を RGB 空間でとり，二値化することによってシルエット画像として得る．全てのカメラのシルエット領域をボクセル空間中に逆投影し，得られる領域の共通部分を抽出することによりボクセルデータを獲得する．

図 2 (a) に，入力画像の例を示す．図 2 (a) の撮影の際には部屋の蛍光灯の他には特別に照明を使用していない．図 3 (a)，(b)，(c)，(d) は，それぞれ，カメラを 1 台，2 台，3 台，4 台使ってボクセルデータが生成される様子をワールド座標系の座標軸とボクセル空間の範囲とともに示したものである．なお，ボクセルの大きさは，実世界との対応もとりやすいという点から，一辺 1cm の立方体とした．



(a) 対象とする人物画像　(b) シルエット画像

図 2: 人物領域の切り出し



(a) カメラ 1 台　(b) カメラ 2 台

(c) カメラ 3 台　(d) カメラ 4 台

図 3: ボクセルデータの生成される様子

3 腕の姿勢推定

3.1 腕の姿勢推定の概要

本稿ではマンマシンインタフェイスとしても応用範囲が広いと考えられる腕の姿勢推定を行う手法を提案する．腕のボクセルデータと腕モデルとのマッチングは，既知腕モデルとの共通部分の体積を計測し，共通部分の割合ががある一定のしきい値を超える場合に真の姿勢に近いとして腕の姿勢を推定する．3 次元データと 3 次元モデルとのマッチングをとることで，図 4 に示されるように従来の 2 次元的な画像上でのマッチングで必要であった「吟味」といった知的な作業を計算処理で置き換えることができ，よりマッチングが容易になると考えられる．

3.2 腕モデル

本研究で扱う腕モデルは，複数の剛体パーツの木構造で表現されるものを扱う．実験では，図 5 に示すように 4 個のパーツから構成され，胴体パーツを根とする人体モデルを使用した．それぞれのパーツは，パーツ毎に形状と体積の情報を持つボクセルデータである．本研究で提案する手法は，それぞれのパーツの形状に実際に 3 次元計測したデータを用いることができるが，実験では人体形状に比較的近

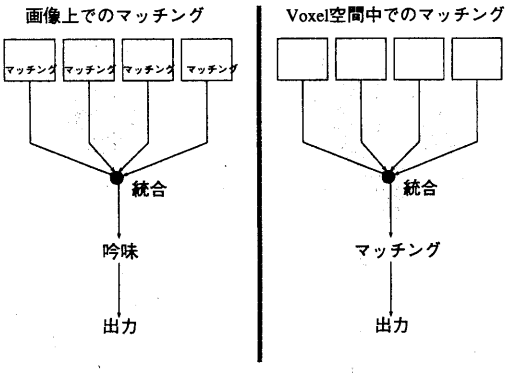


図 4: 画像上でマッチングをとる場合とボクセル空間中でマッチングをとる場合の違い

いと思われる楕円柱を用いた．人体モデルの運動の自由度は，左右上腕に各 3，左右下腕に各 1 の合計 8 自由度とした．この 8 個のパラメータから構成されるパラメータセットを与えることにより，各パーツの位置は一意に定まる．

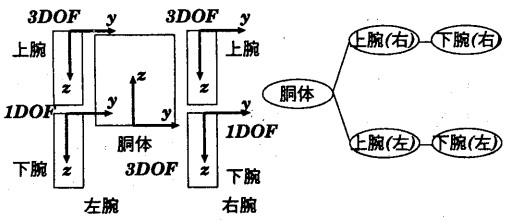


図 5: 腕モデルの自由度の配置

3.3 評価基準

モデルの内部に含まれるボクセルの数が多い程もっともらしい姿勢に近くなるが，各パーツの大きさが異なるため，単に数の多さだけでは判断できない．そこでマッチングには，各パーツの体積に占めるボクセルデータ体積の割合，つまりはパーツのボクセルの数に対するパーツ内部に含まれるボクセルデータ数の割合が高いものが，よりもっともらしい姿勢を表すという評価基準を設定した．

$$\text{評価基準} = \frac{\text{各パーツに含まれるボクセルデータの体積}}{\text{各パーツの体積}} \tag{6}$$

4 人体の姿勢推定の精度向上に関する検討

人体の姿勢を精度良く推定することは様々な応用範囲を考えた場合に非常に重要な点となりうる．従来の姿勢推定法ではマーカやデータスーツ，データグローブ等を用いたものがあるが，非接触でこのレベルの精度を実現することはなされていない．そこで，本研究では真値との誤差をあらかじめ計測し，カルマンフィルタにより最小化を行うことで追跡精度の向上を試みた．

4.1 誤差の検討

ボクセルデータと人体モデルとのマッチングをとる場合に生じる誤差の種類について検討した．この場合生じる誤差の種類には以下のものが考えられる．

(1) ボクセルデータにのるノイズ

(2) 関節角の量子化誤差

(3) モデルの形状誤差

これらのノイズの多くは撮像された画像の解像度，人物領域切り出しの精度などに起因し白色性であると仮定することができる．

人体の運動追跡をボクセルデータと人物モデルとのマッチングという形で実現する場合，その誤差はモデルとボクセルデータとのマッチング率という形で計測され，白色性であると考えられる．本研究ではこの誤差を AR モデルとカルマンフィルタを用いることで軽減することを考える．

4.2 カルマンフィルタの適応 [4]

マッチング率の低いフレームが，ある画像シーケンスの途中で起こったとすれば，それまで観測した運動パラメータを使って，そのフレームの運動を推定することができる．推定する運動パラメータを過去のパラメータの値で表現できるとし，AR モデル (Autoregressive Model) でモデル化する．即ち，時刻 j での，ある一つの運動パラメータの値 y_j は，過去 h フレームのパラメータ値の線形和で表されるとする．

$$y_j = \sum_{i=1}^{h} a_{ji} y_{j-1} + v_j \tag{7}$$

ここで，a_iは自己回帰係数である．また，v_jは平均0分散σ_jの正規分布に従う白色雑音とする．

カルマンフィルタを用いるために，運動パラメータの時系列を状態空間モデルに当てはめて考える必要がある．状態空間モデルは次のように定める．

$$\boldsymbol{s}_j = \boldsymbol{F}_j \boldsymbol{s}_{j-1} + \boldsymbol{G} v_j \qquad (j = 0, 1, 2, \cdots) \quad (8)$$

$$\boldsymbol{y}_j = \boldsymbol{H} \boldsymbol{s}_j + w \quad (9)$$

ここで，状態ベクトル$\boldsymbol{s}_j$，および行列F，G，Hは次のように定義される．

$$\boldsymbol{s}_j = (y_j, y_{j-1}, \cdots, y_{j-h+1})^T \quad (10)$$

$$g(t_k) = \begin{pmatrix} a_{j1} & a_{j2} & \cdots & a_{jh} \\ 1 & & & 0 \\ & \ddots & & \vdots \\ & & 1 & 0 \end{pmatrix} \quad (11)$$

$$\boldsymbol{G} = (1, 0, \cdots, 0)^T \quad (12)$$

$$\boldsymbol{H} = (1, 0, \cdots, 0)^T \quad (13)$$

w_jは平均0，分散σ_wの正規性白色雑音であり、あらかじめ予備実験により求めてあるものとする．

このような状態空間モデルを各運動パラメータについて当てはめる．状態空間モデルとして表現したパラメータは，カルマンフィルタにより予測値を推定することができる．このようにして，マッチング率の低いフレームに対してモデルの位置と姿勢を推定する．全体のアルゴリズムを図6に示す．

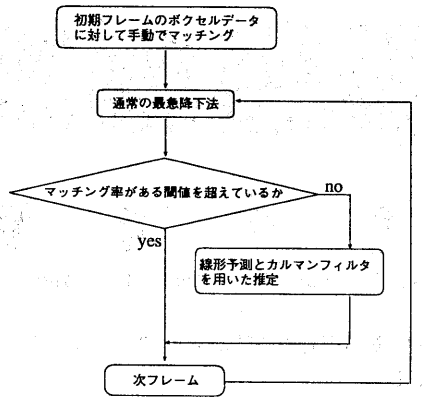


図 6: 運動推定処理のアルゴリズム

5 実験

5.1 システム構成

本研究で用いるシステムの構成を図7に示す．4台のCCDカメラで撮影を行い，Quad Switcherを用いて4台のカメラから得られる画像を，1枚の画像を4分割した形の画像にまとめ，ビデオディスクに取り込む．ＰＣを通してビデオディスクを制御し，ビデオディスクに記録した画像を，AD/DA変換器を通してＰＣに取り込み，ワークステーションまたはＰＣで処理を行う．

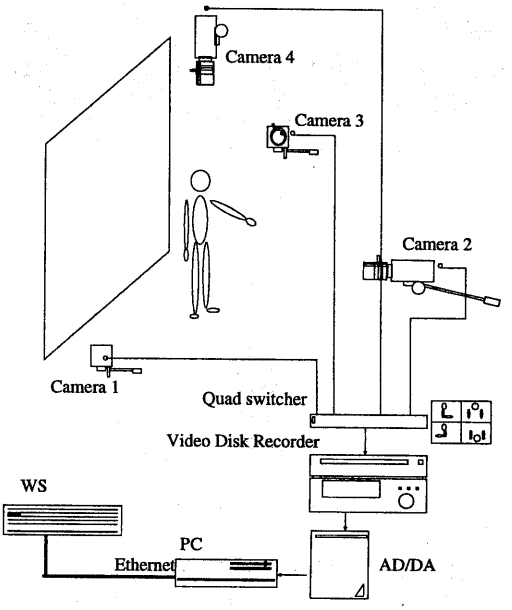


図 7: システム構成

5.2 実験結果

実験ではマンマシンインタフェイスとしても応用範囲が広いと考えられる，人物の腕の運動の推定を行った．カメラは人物の腕の動きが撮像しやすいように，人物に対し正面，真上，右斜め前方，左斜め前方の4箇所に配置しそれぞれのカメラに対してキャリブレーションを行った．また，背景は肌色の領域が抽出しやすいようにブルーの背景を使用した．

実験に用いた画像を図8 (a) に示す．まず，図8 (a) の様に4台のカメラから獲られた画像を1枚にまとめ，図8 (b) に示されるように，肌色の領域を

切り出す．それをボクセル空間へ逆投影し，肌色領域のボクセルデータを作成する．作成された腕領域のボクセルデータを図 9 に示す．ボクセル空間は座標系の原点を人物の腰の辺りにとり，人物の正面方向に X 軸，右手方向に Y 軸，真上方向に Z 軸を配置した．また，単位ボクセルを一辺 1cm の立方体とし，空間の広さは 150(X 方向)×150(Y 方向)×100(Z 方向) とした．

(a) 対象とする人物画像　　(b) シルエット画像

図 8: 腕領域の切り出し

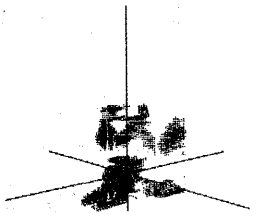


図 9: 実験に用いたボクセルデータ

初期フレームに対し手動でモデルをマッチングさせ，その後のフレームを図 6 に示すフローチャートに従い追跡を行った．

実験結果を図 11 に示す．図 10 と比較してわかるように比較的粗い近似モデルにおいてもカルマンフィルタとＡＲモデルを併用することでよりロバストな腕の運動追跡がなされていることがわかる．

6　本研究のまとめ

腕のボクセルデータと腕モデルとを，3 次元空間内でマッチングをとり，運動追跡を行う手法について述べた．

今後の課題として，ボクセルデータ獲得の高速化の手法の検討があげられる．ボクセルデータからの姿勢推定に関する事としては，ボクセルデータとモデルに色情報を付加することによる表面情報の利用，

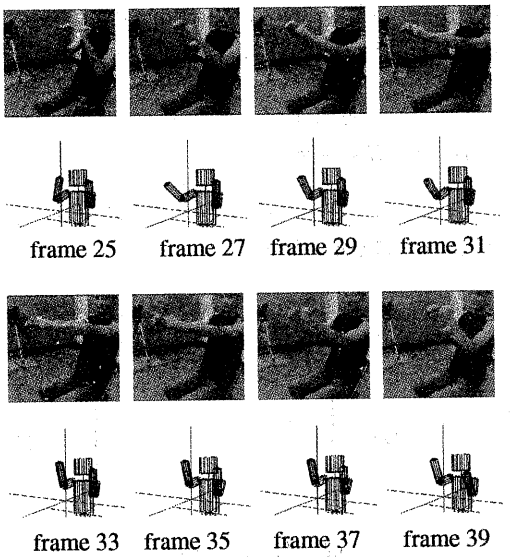


図 10: 腕の運動の追跡結果 (カルマンフィルタを用いない場合)

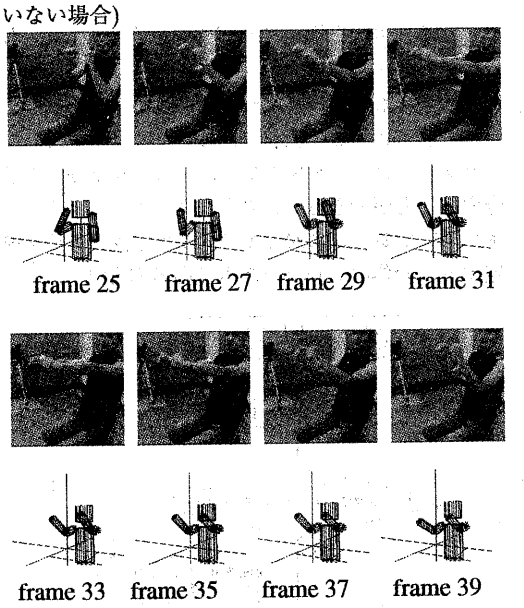


図 11: 腕の運動の追跡結果 (カルマンフィルタを用いた場合)

切り出されたボクセルデータの境界部分に近い程重みを付加することによる外側領域への収束，慣性や運動学を導入することによる初期パラメータセットの改良などがあげられる．

参考文献

[1] 井上 英輝，長橋 宏，“多視点動画像を用いた人体の動きの認識に関する研究”，映像情報メディア学会冬季大会講演予稿集，6-5，pp.105，1997年12月．

[2] 中島 正之 “3次元CG”，テレビジョン学会編，オーム社，1994年2月．

[3] 井口 征士，佐藤 宏介 “三次元画像計測”，昭晃堂，1990年11月．

[4] 大田 佳人，山際 貴志，山本 正信 “キーフレーム拘束を利用した単眼動画像からの人間動作の追跡”，電子情報通信学会学会誌，Vol.J81，No.9，pp.2008-2018，1998年9月．

[5] 青木 由直，棚橋 真 “衛星通信を利用した知的通信方式による手話画像伝送の研究”，電子情報通信学会学会誌，Vol.J80，No.5，pp.441-442，1997年5月．

[6] 川田 聡，近藤 拓也，山本 正信 “ロボットアームモデルに基づく人間動作の3次元動画像追跡”，電子情報通信学会春季大会講演論文集，SD-8-4，7-383，1994年3月．

[7] 山本 正信，川田 聡，近藤 拓也，越川 和忠 “ロボットモデルに基づく人間動作の3次元動画像追跡”，電子情報通信学会論文誌D-II，Vol.J79-D-II，No.1，pp.71-83，1996年1月．

[8] 川田聡，佐藤明知，大崎喜彦，山本正信 “人間動作のマルチカメラによる追跡と仮想世界での再現”，電子情報通信学会技術研究報告，PRU 95-96，pp.103-108，1995年7月．

[9] 佐藤 明知，川田 聡，大崎 喜彦，山本 正信 “多視点動画像からの人間動作の追跡と再構成”，電子情報通信学会論文誌 D-II，Vol.J80-D-II，No.6，pp.1581-1589，1997年6月．

[10] 亀田 能成，美濃 導彦，池田 克夫 “シルエット画像を用いた人体の動作推定”，電子情報通信学会秋季大会講演論文集，D-355，7-363，1994年9月．

[11] 亀田 能成，美濃 導彦，池田 克夫 “モデルを用いた人体の動作推定法—慣性の導入—”，電子情報通信学会春季大会講演論文集，D-658，p.384，1995年3月．

[12] 亀田 能成，美濃 導彦，池田 克夫 “差分画像を利用した人体の動作認識”，電子情報通信学会技術研究報告，PRU 95-98，pp.115-120，1995年7月．

[13] D.M.Gavrila and L.S.Davis “3-D model-based tracking of humans in action a multi-view approach”，Proceedings 1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition，pp.73-80，1996年6月．