環境モデルの導入による頑健な人物追跡の実現

佐藤洋一

情報理工学系研究科電子情報学専攻

1　はじめに

移動する対象, 特に人物を検出し追跡する技術は, コンピュータビジョンの分野で最も重要な課題の 1 つとして，これまでに多くの手法が提案されてきた（例えば，[1, 9, 2, 4, 7, 8, 10] など）. しかし, 複雑な背景を持った環境内において，照明などの変動などに対しても頑健な追跡を実現することは容易ではなく，実際の室内環境などにおいても人物を安定に追跡することを可能とする技術の開発が望まれている．これに対し, 本研究では, ステレオビジョンカメラ及びレンジセンサを用い，多視点画像の統合に加え, レンジセンサから得られる室内形状情報にもとづき室内空間における人物頭部の存在可能性の偏りを考慮することにより,，複雑な屋内環境においても頑健に動作する人物追跡手法を提案する.

提案手法における具体的な処理手順は次のようになる．追跡対象を人物頭部とし, 向きをもった楕円体としてモデル化する. 追跡に用いる時系列フィルタリングとしては, パーティクルフィルタ [3] を用いる. 観測にあたっては, 室内の天井四隅にそれぞれ室内中心方向を向いたステレオビジョンカメラを配置し (図 1) , 色情報と距離情報を取得し, それぞれで人物頭部らしさを評価する．そして, これら多視点画像から得られる情報の統合により, 人物頭部の 3 次元空間位置, 及び, 人物頭部の向きの評価を行う．そのうえで, レンジセンサによって得られた室内形状から, 壁や机の配置による室内空間中の人物頭部の存在可能性を考慮した環境モデルを評価に加えることにより, より安定した追跡を実現する．以上の手法により, 人物の追跡を行い, 実環境下での実験によって本手法の有効性を確認する.

図 1: センサ配置

2　パーティクルフィルタ [3]

ある時刻 t における追跡対象の状態量を $\mathbf{x}_t$, 画像から得られる観測結果を $\mathbf{z}_t$ とおく. また, 時刻 t までに得られる観測結果を $\mathbf{Z}_t = (\mathbf{z}_1, \ldots, \mathbf{z}_t)$ とする．このとき, 時刻 t までの観測結果が得られた時の状態 $\mathbf{x}_t$ の確率密度関数 $P(\mathbf{x}_t \mid \mathbf{Z}_t)$ を推定する.

追跡対象の, 時刻 $t-1$ における確率密度関数 $P(\mathbf{x}_{t-1} \mid \mathbf{Z}_{t-1})$ と, 時刻 $t-1$ から t への動きモデル $P(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ が与えられると, 時刻 t における事前確率 $P(\mathbf{x}_t \mid \mathbf{Z}_{t-1})$ は, マルコフ過程を仮定することにより次のようになる.

$$P(\mathbf{x}_t \mid \mathbf{Z}_{t-1}) = \int P(\mathbf{x}_t \mid \mathbf{x}_{t-1}) P(\mathbf{x}_{t-1} \mid \mathbf{Z}_{t-1}) d\mathbf{x}_{t-1}. \quad (1)$$

このとき, 時刻 t における尤度 $P(\mathbf{z}_t \mid \mathbf{x}_t)$ を画像から推定すると, 時刻 t における確率密度関数

$P(\mathbf{x}_t \mid \mathbf{Z}_t)$ は, ベイズの法則に従い, 次のようになる.

$$P(\mathbf{x}_t \mid \mathbf{Z}_t) \propto P(\mathbf{z}_t \mid \mathbf{x}_t)P(\mathbf{x}_t \mid \mathbf{Z}_{t-1}). \qquad (2)$$

パーティクルフィルタでは, ある時刻 t における確率密度関数 $P(\mathbf{x}_t \mid \mathbf{Z}_t)$ を, 状態 $\mathbf{x}_t$ の仮説群 $\{\mathbf{s}_t^{(1)}, \ldots, \mathbf{s}_t^{(N)}\}$ と各仮説の重み $\{\pi_t^{(1)}, \ldots, \pi_t^{(N)}\}$ の組によって離散的に表現する. ここで, 時刻 t における n 番目の仮説の状態量を $\mathbf{s}_t^{(n)}$ とし, 重みは $\pi_t^{(n)} = P(\mathbf{z}_t \mid \mathbf{x}_t = \mathbf{s}_t^{(n)})$ により評価する.

仮説群に適用されるプロセスは, 次の 3 つの部分から構成される. このプロセスの繰り返しによって追跡が実現される.

1. 時刻 $t-1$ において, 観測 $\mathbf{Z}_{t-1}$ が得られたときの状態量 $\mathbf{x}_{t-1}$ の分布 $P(\mathbf{x}_{t-1} \mid \mathbf{Z}_{t-1})$ が, N 個の重みつき仮説群 $\{(\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)}), n = 1, \ldots, N\}$ で表されているとき, 各仮説の重み $\{\pi_{t-1}^{(1)}, \ldots, \pi_{t-1}^{(N)}\}$ の比に従い, 仮説群 $\{\mathbf{s}_{t-1}^{\prime(1)}, \ldots, \mathbf{s}_{t-1}^{\prime(N)}\}$ を選択する.

2. 選択された仮説群を, 動きモデル $P(\mathbf{x}_t \mid \mathbf{x}_{t-1} = \mathbf{s}_{t-1}^{\prime(n)})$ に従い伝播し, $P(\mathbf{x}_t \mid \mathbf{Z}_{t-1})$ に相当する時刻 t における N 個の仮説群 $\mathbf{s}_t^{(n)}$ を生成する.

3. 重み $\pi_t^{(n)}$ を画像から推定することで, 新しいサンプル $\mathbf{s}_t^{(n)}$ の重み $\pi_t^{(n)} = P(\mathbf{z}_t \mid \mathbf{x}_t = \mathbf{s}_t^{(n)})$ を求める. ただし, 重み $\pi_t^{(n)}$ が $\sum_{n=1}^{N} \pi_t^{(n)} = 1$ となるように正規化を行う. その結果, 時刻 t における $P(\mathbf{x}_t \mid \mathbf{Z}_t)$ の近似表現である $\{(\mathbf{s}_t^{(n)}, \pi_t^{(n)}), n = 1, \ldots, N\}$ を得る. また, 追跡対象の最適な状態量推定として, 仮説群の期待値を用いる.

3 人物頭部の位置及び向きの推定

3.1 人物頭部モデル

人物頭部のモデルとして楕円体を仮定する. まず, 室内空間に 3 次元世界座標 XYZ をとる. このとき, 床面が XY 平面と一致し, 高さ方向が Z 軸となるようにする. 人物頭部の形状は不変とし, 位置を楕円体の中心座標 (x, y, z) で表す. 人物頭部の向きは, 上下方向にはあまり動かないと仮定すると, X 軸を基準とした Z 軸回りの回転 θ のみで表される.

以上により, 人物頭部は 4 次元の状態量 $\mathbf{s} = (x, y, z, \theta)$ で表される. なお, 人物頭部の状態量について, 時刻 t における n 番目の仮説を, $\mathbf{s}_t^{(n)} = (x_t^{(n)}, y_t^{(n)}, z_t^{(n)}, \theta_t^{(n)})$ と表すことにする.

ここで, 時刻 t における n 番目の仮説を表す楕円体 $\Gamma_t^{(n)}$ を i 番目のカメラの画像に投影したときに得られる領域を $\Omega_{i,t}^{(n)}$ とする. 投影関数を F_i とすると, 次のようになる.

$$\Omega_{i,t}^{(n)} = F_i(\Gamma_t^{(n)}). \qquad (3)$$

人物頭部を楕円体でモデル化しているので, $\Omega_{i,t}^{(n)}$ は楕円となる. なお, 領域 $\Omega_{i,t}^{(n)}$ 内の総画素数を $|\Omega_{i,t}^{(n)}|$ で表す.

3.2 色情報の利用による人物頭部らしさの評価

i 番目のカメラのカラー画像から仮説 $\mathbf{s}_t^{(n)}$ の色情報に基づく重み $\pi_{i,t}^{color,(n)}$ を求める. ここで, 本稿での重みとは, 人物頭部らしさをいう.

背景や髪色の個人差の影響を減らすため, 評価には肌色のみを用いる. 色空間としては, HSV 色空間を用いる. これは, 多くの研究でこの色空間が肌色領域の抽出に最も適した色空間のひとつであることが示されているためである [5, 6] . しかし, ここでは, 照明変化の影響を減らすため V を無視し, 肌色領域の判定に H と S からなる 2 次元の色空間を用いる. そのうえで, 人間の肌色の学習データに基づき, あらかじめ肌色領域を定義する. そして, 画素の HS 色空間での色度がその領域に含まれれば, その画素は肌色であると判断する.

重みの評価には, 楕円 $\Omega_{i,t}^{(n)}$ 内部の肌色画素の割合とあらかじめ得ておいた人物頭部の肌色画素の割合との類似度を用いる. 観測によって得られる i 番目のカラー画像において, 楕円 $\Omega_{i,t}^{(n)}$ 内部

の肌色領域に含まれる画素数を $c_{i,t}^{(n)}$, その割合を $\bar{c}_{i,t}^{(n)}$ とすると, 次式が成立する.

$$\bar{c}_{i,t}^{(n)} = \frac{1}{|\Omega_{i,t}^{(n)}|} c_{i,t}^{(n)}. \tag{4}$$

また, 追跡対象の人物に対して, あらかじめ得ておいた人物頭部の肌色画素の割合を $\hat{c}_{i,t}^{(n)}$ とする. 人物頭部がカメラ中心を向いているほど肌色領域は大きく, カメラ中心に対して背を向けているほど肌色領域は小さい. このとき, 重み $\pi_{i,t}^{color,(n)}$ は, これらの差分が少ないほど高く, 次のような評価関数により与えることができる.

$$\pi_{i,t}^{color,(n)} = a_i^{color} - b_i^{color} \left| \bar{c}_{i,t}^{(n)} - \hat{c}_{i,t}^{(n)} \right|. \tag{5}$$

ここで, $a_i^{color}(> 0)$, $b_i^{color}(> 0)$ は定数である. なお, 実験では, これらの値を経験的に与え, 仮説による予測と観測の類似度が高いほど高い重みを与えた.

3.3 距離情報の利用による人物頭部らしさの評価

i 番目のカメラの距離画像から仮説 $\mathbf{s}_t^{(n)}$ の距離情報に基づく重み $\pi_{i,t}^{depth,(n)}$ を求める.

観測によって得られる, i 番目のカメラの距離画像に対して, 楕円 $\Omega_{i,t}^{(n)}$ 内部の画素 p のカメラ座標を $\tilde{\mathbf{w}}_{i,t}^{(n)}(p)$, 楕円体の中心 $\mathbf{v}_t^{(n)} = (x_t^{(n)}, y_t^{(n)}, z_t^{(n)})$ を i 番目のカメラのカメラ座標系へ投影した結果の座標を $\tilde{\mathbf{v}}_{i,t}^{(n)}$, $\tilde{\mathbf{w}}_{i,t}^{(n)}(p)$ と $\tilde{\mathbf{v}}_{i,t}^{(n)}$ との距離を $d_{i,t}^{(n)}(p)$ とすると, 次式が成立する.

$$d_{i,t}^{(n)}(p) = \left| \tilde{\mathbf{w}}_{i,t}^{(n)}(p) - \tilde{\mathbf{v}}_{i,t}^{(n)} \right|. \tag{6}$$

画素 p に対応する人物頭部楕円体上の点の楕円体中心からの距離は既知であるので, これを $\hat{d}_{i,t}^{(n)}(p)$ とする. このとき, 重み $\pi_{i,t}^{depth,(n)}$ は, 各画素におけるこれらの差分の総和が少ないほど高く, 次のような評価関数により与えることができる.

$$\pi_{i,t}^{depth,(n)} = a_i^{depth} - b_i^{depth} \left(\frac{1}{|\Omega_{i,t}^{(n)}|} \sum_{p \in \Omega_{i,t}^{(n)}} \left| d_{i,t}^{(n)}(p) - \hat{d}_{i,t}^{(n)}(p) \right| \right) \tag{7}$$

ここで, $a_i^{depth}(> 0)$, $b_i^{depth}(> 0)$ は定数である. なお, 実験では, 色情報の場合と同様, これらの値を経験的に与え, 仮説による予測と観測の類似度が高いほど高い重みを与えた.

3.4 環境モデルの導入

本稿ではさらに, 環境モデルの導入により追跡の安定化を図る.

レンジセンサからあらかじめ得ておいた室内の 3 次元形状に基づき, 壁や机の配置による室内空間における人物頭部の存在可能性を考える. こうして求められた空間中の人物頭部の存在可能性を環境モデルと呼ぶことにする. 環境モデルを考慮することにより, パーティクルフィルタの仮説群の評価において, 背景中に存在する物体内部のような人物頭部が存在し得ない領域, あるいは机や棚などの上部といった人物頭部の存在可能性の低い領域に生成される仮説の重みを抑えることが可能となる. その結果, 人物頭部の存在可能性の低い領域からの仮説の生成が抑えられ, より安定した追跡が期待される.

環境モデルとして, 次のような 3 つの領域からなるモデルを考える.

- 領域 A： 壁の外側及び机や棚など静的物体の内部. 人物頭部は存在しない.
- 領域 B： 机や棚など静的物体の垂直方向. その他の空間に比べて人物頭部の存在可能性は低い.
- 領域 C： 低い領域, あるいは身長より上の領域. その他の空間に比べて人物頭部の存在可能性は低い.

これに基づき, 状態量 $\mathbf{s}_t^{(n)}$ における人物頭部の存在可能性に基づく重み $e_t^{(n)}$ を求める. 各領域における人物頭部の存在可能性をそれぞれ $e_t^{A,(n)}, e_t^{B,(n)}, e_t^{C,(n)}$, どの領域にも属さない場合の人物頭部の存在可能性を α とし, 領域が重なった場合はその最小値をとる. すなわち,

$$e_t^{(n)} = \min(e_t^{A,(n)}, e_t^{B,(n)}, e_t^{C,(n)}, \alpha). \tag{8}$$

なお, 実験では, $e_t^{A,(n)} = 0$, $e_t^{B,(n)} = 0.5$, $e_t^{C,(n)} = 0.2$, $\alpha = 1.0$ とした.

3.5 情報の統合

複数カメラからの色情報及び距離情報と環境モデルとの情報の統合を行う.

ここで, 色情報は, 人物頭部の対称性により単一のカメラでは左右対称な状態を判別できず, 距離情報からは向きの判別はできないが追跡対象と背景を区別することができる. また, 環境モデルを導入することにより, 室内空間中の人物頭部の存在可能性の高い仮説に高い重みを与えることができる. 従って, 3 次元での追跡を実現し, かつ, 安定した追跡を実現するためには, これらの情報の統合が必要となる.

統合手法としては, 実時間性を損なわないよう単純化し, 仮説 $\mathbf{s}_t^{(n)}$ における各カメラの色情報に基づく重み $\pi_{i,t}^{color,(n)}$ 及び距離情報に基づく重み $\pi_{i,t}^{depth,(n)}$ と環境モデルによって定義される人物頭部の存在可能性 $e_t^{(n)}$ との積をとる. すなわち,

$$\pi_t^{(n)} = e_t^{(n)} \prod_i \pi_{i,t}^{color,(n)} \prod_i \pi_{i,t}^{depth,(n)}. \qquad (9)$$

このようにして, 環境モデルを考慮した仮説の重みが得られる.

4 実験結果

4.1 実験システムの概要

システムは 1 台のサーバ PC と 4 台のクライアント PC (CPU Intel Pentium4 2.0GHz, Memory 1.0GByte) からなり, これらを通信速度が 1Gbps の Gigabit Ethernet で接続した. また, ステレオビジョンカメラとして Point Grey 製の Digiclops を, レンジセンサとして SICK 製の LMS200 を用いた. これらのセンサはあらかじめキャリブレーションしておいた. これにより, 各センサの観測をあらかじめ設定しておいた世界座標を介して連携することが可能である. クライアント PC にはそれぞれ 1 つの Digiclops カメラが取り付けられており, 色情報及び距離情報を取得できる. また, あらかじめレンジセンサから得られた室内形状から環境モデルを作成しておく.

4.2 追跡結果

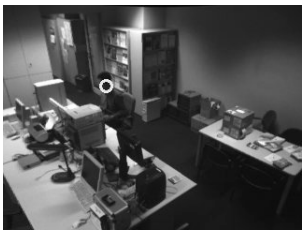
以上の準備の下, 本稿で提案した手法を用いて, 実環境下で人物頭部追跡実験を行った. 追跡時のフレームレートは約 4fps であった.

座っている状態から立ち上がり, 室内を頭部の向きを変化させて歩いた場合について追跡実験を行った. 実験で得られた画像の一部を図 4.2 に示す. 各フレームには, 情報統合後の人物頭部の状態量の期待値による推定結果を, カメラの方向を向いているほど輝度が高くなるような円として, カラー画像に重ねて表示した. なお, 仮説の重みの評価には, 色情報, 距離情報に加え環境モデルを用いた.

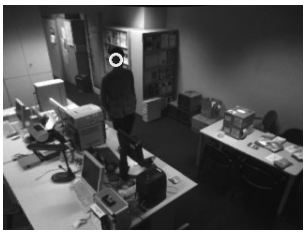
図 4.2 から, 仮説群の期待値によって, ほぼ人物頭部中心を推定できていることがわかる. これは, 情報統合後の重みが高い仮説が, 全体として人物頭部付近に集まっていることによるためであると考えられる. また, 色情報による評価において, カメラの向きに応じた人物頭部の色モデルが有効であり, 複数カメラでの観測結果の情報統合により, 人物頭部の位置及び向きの推定が有効に働いていると考えられる.

また, 提案手法による追跡の精度を定量的に調べるために, 画像中の人物頭部を手動で求め, 複数画像から逆投影して求めた 3 次元座標値を真の位置と見なし, 推定結果と比較した. 図 4.2 に, 推定結果とそれに対応する人物頭部の真の位置の 3 次元空間上及び XY 平面上での軌跡を示す. また, Z 軸方向及び XY 平面上での誤差の平均及び標準偏差を表 1 に示す.

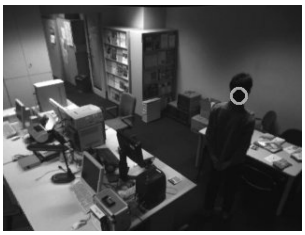
ここで, Z 軸方向の検出誤差は 6cm 程度であり, 下方に検出された. これは, 色情報による評価が首も検出していることと, 距離情報による評価が服も検出しているためと推定される. 一方, XY 平面上での検出誤差は 4cm 程度であり, 十分正確



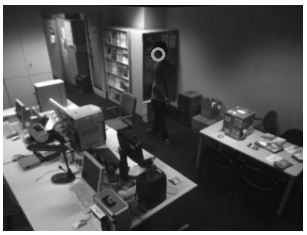
(a) #30



(b) #50



(c) #70



(d) #90

図 2: 人物頭部の追跡結果

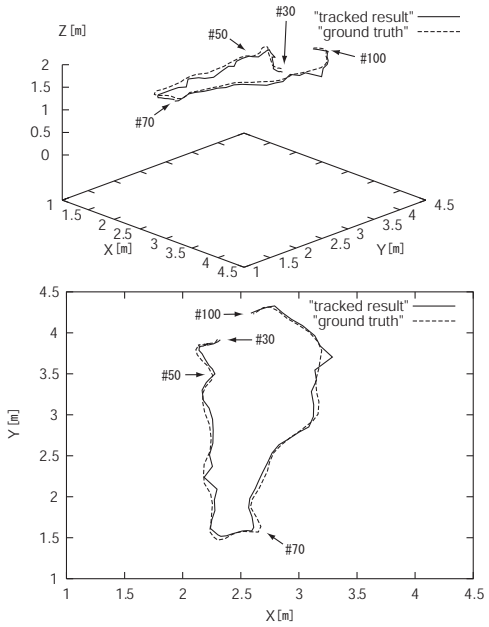


図 3: 人物頭部の追跡結果の軌跡

表 1: 人物頭部の追跡における検出誤差

	平均 [cm]	標準偏差 [cm]
Z 軸方向	6.02	2.62
XY 平面	4.34	2.70

に追跡できているといえる.

4.3 環境モデルの有効性

同じ画像シーケンスに対して, 環境モデルの導入の有効性を確認するために, 環境モデルを利用しない場合と利用した場合についての比較を行った.

環境モデルを利用しなかった場合, (#60) の時点で, 追跡対象が 4 つのカメラ全ての視野範囲内にあるにもかかわらず仮説が複数のカメラの視野範囲外に発散し, 状態量推定に失敗した. そこで, この直前のフレームにおけるパーティクルフィルタの状態を色情報の評価結果と共に図 4.3 に表示した. また, 比較のため, 環境モデルを利用した場合の結果を図 4.3 に表示した.

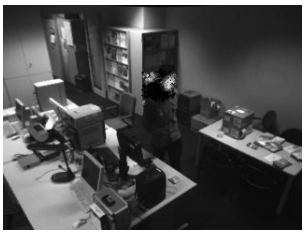
環境モデルを利用しなかった場合, 図 4.3 から, カメラ 1 (#58) において, 背後の肌色に近い物体について色評価による誤検出が発生しており, この結果, 仮説が人物頭部以外の領域に引き寄せられると共に, 仮説の向きについても誤検出が起こっていることがわかる. この様子を別のカメラ 2 (#58) から見ると, 人物頭部には, 仮説群がほとんど発生していないことがわかる. そのため, 環境モデルを利用しない場合では追跡に失敗しているといえる.

一方, 環境モデルを利用した場合, 図 4.3 から, カメラ 2 (#58) において, 人物頭部にも仮説が発生していることがわかる. 机などの物体や高さの考慮という環境モデルの効果により, 誤検出による仮説の生成が抑制された結果, 人物頭部の存在可能性の高い部分に効率的に仮説が発生し, 追跡を続けられたといえる.

以上のように, 環境モデルの導入により, 人物追跡の安定性が向上したといえる.

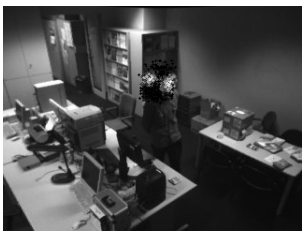
5 おわりに

本稿では, 複数センサからの入力情報の統合による, 人物頭部の向きを考慮した 3 次元空間実時




(a) #58 (カメラ 1)　　(b) #58 (カメラ 2)

図 4: 環境モデルを利用しない場合の色情報の観測




(a) #58 (カメラ 1)　　(b) #58 (カメラ 2)

図 5: 環境モデルを利用した場合の色情報の観測

間追跡手法を提案した.

パーティクルフィルタを用いた人物頭部の追跡において, 複数のステレオビジョンカメラから得られる色情報及び距離情報による人物頭部らしさの評価に加え, レンジセンサから得られる環境モデルを統合した. これにより, 室内空間における人物頭部の存在可能性の偏りを考慮したより安定した人物頭部の位置及び向きの推定を実現することができた.

今後は, オクルージョンによって追跡対象が遮蔽された場合を区別するなど情報の信頼性を考慮した情報統合, また, 長期の観測から人物の行動履歴を取得し, 人物の通りやすい経路や通ることが少ない経路などを考慮し, 環境モデルに反映させる手法を検討していく予定である.

参考文献

[1] S. Birchfield: Elliptical Head Tracking Using Intensity Gradients and Color Histograms, Proc. IEEE CVPR '98, pp.232-237, 1998.

[2] I. Haritaoglu, D. Harwood and L. S. Davis: A Real-Time System for Detecting and Tracking People in 2 1/2 D, Proc. ECCV '98, pp.877-892, 1998.

[3] M. Isard and A. Blake: Condensation - Conditional Density Propagation for Visual Tracking, Int. J. Computer Vision, vol.29, no.1, pp.5-28, 1998.

[4] G. Loy, L. Fletcher, N. Apostoloff and A. Zelinsky: An Adaptive Fusion Architecture for Target Tracking, Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FGR '02), pp.261-265, 2002.

[5] J. Terrillon, A. Pilpre, Y. Niwa and K. Yamamoto: Analysis of Human Skin Color Images for a Large Set of Color Space and for Different Camera Systems, Proc. IAPR Workshop on Machine Vision Applications (MVA '02) pp.20-25, 2002.

[6] B. D. Zarit, D. J. Super and F. K. H. Quek: Comparison of Five Color Models in Skin Pixel Classification, Proc. Int. Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pp.58-63, 1999.

[7] 浮田宗伯，松山隆司: 能動視覚エージェント群による複数対象の実時間協調追跡, 情報処理学会 CVIM 研究会論文誌, vol.43，no.SIG11，pp.64-79，December 2002.

[8] 大塚和弘, 武川直樹: 多視点観測に基づく複数物体の相互オクルージョン解析と逐次状態推定, 情報処理学会 CVIM 研究会論文誌. vol.44, no.SIG17, 2003.

[9] 杉本晃宏, 谷内清剛, 松山隆司: 確信度付き仮説群の相互作用に基づく複数対象追跡, 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研究会論文誌, vol.43 no.SIG04, June 2002.

[10] 中島平，浜崎浩二，岡谷貴之，出口光一郎: CONDENSATION を用いた多視点画像の融合による複数人物の追跡, MIRU 2002, vol.2, pp.317-322, 2002.