コンピュータビジョンと
イメージメディア 109－3
（1998. 1. 22）

複数の能動的カメラを用いた人物の抽出と追跡

藪内 勉　　岩井 儀雄　　谷内田 正彦

大阪大学大学院基礎工学研究科システム人間系専攻

近年、ジェスチャー認識について多くの研究がなされている。これらの研究は多くの場合、人物は大きく動かず固定されたカメラの視界内に入っていることを仮定している。しかし、カメラの視界範囲は限られているので人物が大きく動くと画像にとらえることができなくなる、そこで、本研究ではアクティブカメラを用いてカメラの視界を補い、移動する人物を追跡する。また、多様な視点からの観測と観測できる視界を広げるために複数のカメラを用いる。

複数のカメラを用いる従来の研究は、カメラと追跡対象の関係はあらかじめ決められていたが、固定的な役割分担では移動する対象を効率よく追跡できない。そこで、本研究では役割分担を動的に変更する手法を用いて、複数のカメラで複数対象を効率的に追跡する。

Detecting and Tracking Human with Multiple Active Camera

Tsutomu Yabuuchi, Yoshio Iwai, and Masahiko Yachida

Department of Systems and Human Science

Graduate School of Engineering Science, Osaka University Japan

Recently many studies have been made on the gesture recognition. Most of these studies were made on the assumption that the person doesn't move and the camera takes a close-up image of the targets. The limit of observation range restricts a movement of human. We use active cameras in order to pursue the face and hands and to take a close-up image of the objects. We use multiple cameras because multiple viewpoint can improve accuracy of the gesture recognition. Previous studies have been made on a static assignment between multiple cameras and multiple objects. The moving object could not be taken appropriately with the static assignment method. Therefore we propose a method which assigns multiple cameras to multiple objects.

1. 序論

近年の技術の急速な発展によって、コンピューターが使われる範囲は広がっている。かつては、専門家が使うことが前提であったコンピューターは、現在ではさまざまな人にさまざまな用途で使われている。しかし、キーボード・マウスなどの従来のマンマシンインターフェースは、人間が機械に歩み寄る形で実現されており、人間にとって使いやすいとはいえい。しかし、人間同士のコミュニケーションで使われているジェスチャーなどの情報をコンピューターが理解できれば、人間にとってより自然なインターフェースが実現できる。そこで、ジェスチャーをコンピューターへの入力として使う方法が研究されている。

データグローブなどの接触センサを使う方法は被験者にとって負担である。人間が使いやすいインターフェースを実現するためにはカメラのようなリモートセンサーを使うことが望ましい。それゆえ、カメラからの画像を用いてジェスチャーを認識する方法は、広く研究されてきた。しかし、これらの研究の多くは、カメラの視界内に対象を捉えていることを前提としている。たとえば、表情認識[1]や個人識別[2]には顔の画像が必要であり、手指動作の認識[3-6]には手のズームアップが必要である。

また、人物が移動することを制限するようでは、使いやすいインターフェースとはいえない。そこで、能動的カメラを用いて人物を追跡することが必要である。ジェスチャー認識では特定の方向から見た画像が必要な場合がある。例えば、表情認識を行うためには顔の正面画像が必要であることが多い。また、2 つ以上の視点から対象を観測することで、ステレオ視による三次元位置の計測が可能になる。そこで、複数のカメラを用いて、多様な視点から観測した画像を用いる必要がある。

一方、人間のジェスチャーには多くの種類があり、ジェスチャー認識のためには、顔、手、体全体など複数の対象を複数のカメラで追跡しなければならない[7]。つまり、どのカメラがどの対象の追跡を担当するかを決定しなければならない[8]。しかし、静的な割当では移動する対象を効率よく追跡することはできない。

本研究では、複数の能動的カメラを用いて人物を追跡し、状況に応じてカメラが追跡を担当する対象を動的に決定する方法についてのべる。

2. システムの構成

2.1. 全体の構成

図 1に全体の構成を示す。アクティブカメラ・固定カメラで人物を撮影する。画像を制御用コンピューターに入力し、カメラの行動決定に基づいてカメラを制御する。

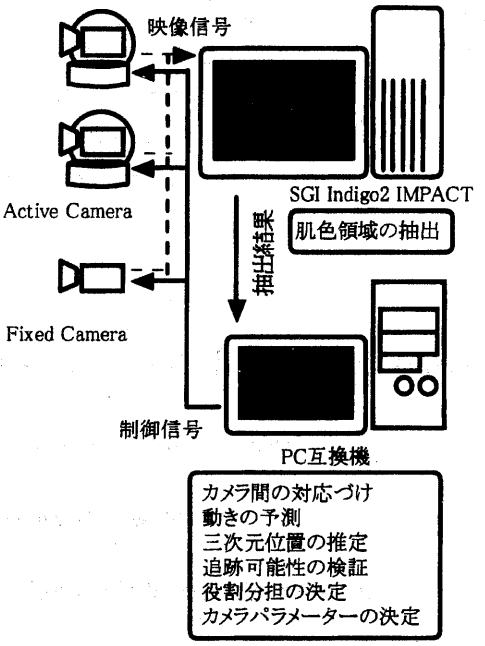

図 1:全体の構成

2.2. カメラモデル

図 2に本研究で使用するアクティブカメラを示す。このカメラはパン・チルト・ズームの制御が計算機から可能である。

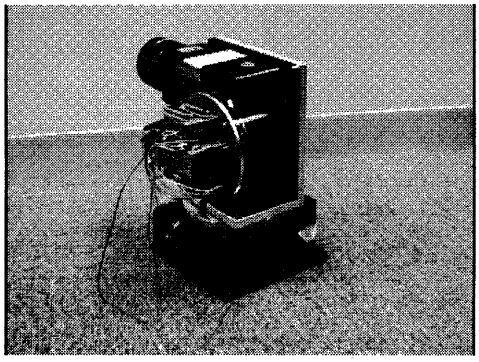

図 2:アクティブカメラの外観

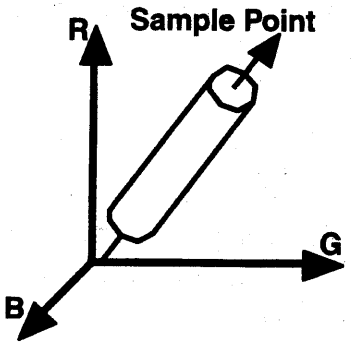

図 3:肌色の分布モデル

カメラは内部パラメーターとして焦点距離 f [mm]、f の制御範囲 f_{max}, f_{min}[mm]、単位時間当たりの速度 f_{speed}[mm/frame] を持ち、f が制御可能である。

外部パラメーターとして、初期位置姿勢をあらわす同次変換行列 T、パン角 θ [deg]、チルト角 ϕ [deg]、それぞれの制御範囲 $\theta_{max}, \theta_{min}, \phi_{max}, \phi_{min}$ [deg]、単位時間当たりの速度 θ_{speed}, ϕ_{speed}[deg/frame] を持つ。θ, ϕ が制御可能である。

ピンホールカメラモデルを用い、ワールド座標で $p_o=(x_o, y_o, z_o, 1)^T$ に存在する点は、カメラでは $p_c=(x_c, y_c, z_c, 1)^T = R_y(\phi) R_z(\theta) T p_o$ とすると、$x= fx_c/(f+ z_c)$, $y= fy_c/(f+ z_c)$ に射影される。ただし、R_y, R_z はそれぞれy軸, z軸についての回転行列である。

3. 人物の抽出

ジェスチャーを認識するためには顔、手の領域を撮影することが必要である。入力された画像から色情報を用いて顔と手の領域を抽出し追跡する。

肌色のサンプルを与えることで、現在の環境・光源での肌色モデルを作る。サンプルとして与えられた色と黒を結ぶ直線を長軸とする円柱内の色を肌色として抽出する。(図 3)

4. 人物の追跡

4.1. カメラ間の対応づけ

複数の対象を区別して追跡するためには、それぞれの対象についてカメラ間での対応関係を知ることが必要である。そこで、エピポーラ拘束と時系列拘束を用いて、肌色領域それぞれについて対応関係を求める。

- エピポーラ拘束による対応づけ

ある時刻 t にカメラ i で観測された肌色領域 $d_i(t)$と別のカメラ j で観測された領域 $d_j(t)$について、$d_i(t)$のエピポーラ線と $d_j(t)$のエピポーラ線の距離があるしきい値以下なら $d_i(t)$と $d_j(t)$は同一の対象とする。

- 時系列拘束による対応づけ

ある時刻にカメラ i 観測された領域 $d_i(t)$ と 1 フレーム前に同一のカメラ i で観測された $d_i(t-1)$ の見かけの角度があるしきい値以下なら $d_i(t)$ と $d_i(t-1)$ は同一の対象とする。

4.2. 動きの予測

対象が追跡可能かどうかを判定するために、t_f フレーム後の対象の位置を予測する。現在の時刻を t とする。

- 三次元動き予測

対象が時刻 t、t-1 でともに2台以上のカメラで観測されている場合、三次元動き予測が可能であり、予測位置 $p(t+t_f)$ は以下の式で与える。

$$p(t+ t_f)=p(t)+ t_f (p(t)-p(t-1)) \qquad (1)$$

- 三次元位置予測

対象が時刻 t で2台以上のカメラで観測され

ている場合、動き情報が利用できないので三次元位置を予測する。

$$p(t+t_f)=p(t) \qquad (2)$$

と予測する。

● 二次元位置予測

対象が時刻 t で 1 台のカメラでしか見えていない場合、動き情報も奥行情報も利用できないので、予測位置は現在の観測方向とする。

$$p(t+t_f)=\lambda p(t) \qquad (3)$$

ただし、λ は未知の値である。

4.3. 三次元位置の推定

各カメラがそれぞれの対象を視界内に収めることができるかどうか判定するためには、カメラから見た対象の位置を知ることが必要である。

対象が複数のカメラから観測されている場合は、ステレオ視により三次元位置がわかるので任意の視点から見た時の位置を求めることができる。しかし対象が1台のカメラからしか観測されていない場合、あるいは対象がまったく観測されていない場合は対象の位置を求めることができないため、カメラが対象を視界内に収めることができるかどうか判定できない。そこで、三次元位置の予測を行う必要がある。

対象がカメラの視界内にないことから、カメラがもっとも長時間観測しなかった位置にあると推定する。ボクセル $v_i(x,y,z)$ を用いて、

・カメラ i の視界内に点 (x, y, z) がある場合、$v_i(x,y,z)$ を 0 とする。

・カメラ i の視界内に点 (x, y, z) がない場合、$v_i(x,y,z)$ を +1 する。

として、カメラごとにボクセルの観測時刻を記録する。

対象の位置を推定する場合は、

・対象のエピポーラ線がわかっている場合、エピポーラ線が通るボクセルのうちで $v_i(x,y,z)$ が最大のボクセルの中心を推定位置とする。

・エピポーラ線がわからない場合、すべてのボクセルのうちで $v_i(x,y,z)$ が最大となるボクセルの中心を推定位置とする。

以下に未発見の対象の位置を推定する例をあげる。(図 4) 各ボクセルの値は、ボクセルがそのカメラによって何フレーム前に観測されたかを表す。現在カメラの視界内にあるボクセルは 0 である。値が大きいほど長時間観測されていないことを意味する。

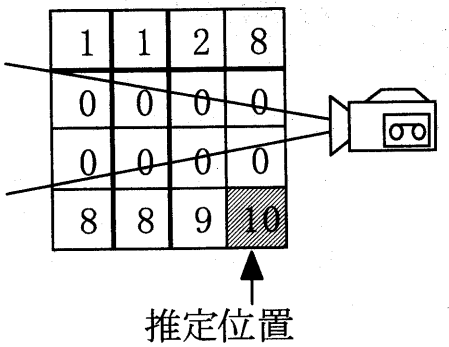

図 4:位置推測の例 1

最大の値を持つボクセル、この例では、右下のボクセルの中心にあると推定する。

次に、1 台のカメラから見たエピポーラ線がわかっている場合の例を示す。(図 5)

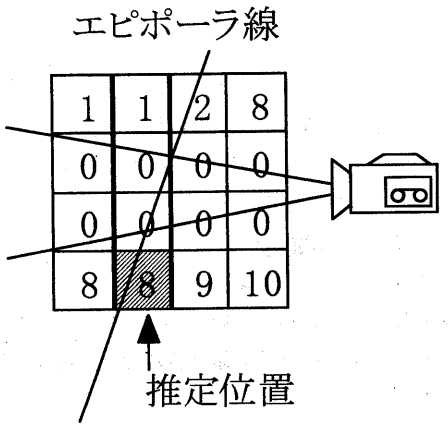

図 5:位置推測の例2

エピポーラ線がわかっている場合は、エピポーラ線上でもっとも長時間観測されなかったボクセルの中心に対象が存在すると推定する。

いずれの場合でも最大の値を持つボクセルが複数ある場合は、追跡のためのカメラの移動

量を減らすために、カメラの視界中央に最も近いボクセルの中心を推定位置とする。

4.4. 追跡可能性の検証

観測対象の予測された位置を利用して、各カメラ j は同時に追跡可能な対象の集合 T_j を生成する。

1. 追跡対象の集合 A の閉包 A^*を求める。
2. 閉包 A^* から要素 $a_i(\#(a_i)=\max(\# A^*)$ を取り出す。取り出す要素がなければ終了する。
3. a_i が追跡可能集合 T_j の閉包 T_j^*に含まれていないことを確かめる。
4. a_i に含まれる追跡対象が同時に追跡可能かどうか判定する。
5. 可能であれば a_i を追跡可能集合 T_j に加えて、その閉包 T_j^*を計算し直す。
6. 2 に戻る。

追跡可能性の判定

追跡可能性は、カメラが持っている最大性能で与えられた対象をすべて視界内に捉えることができれば追跡が可能であるとする。

1. 与えられたすべての対象の重心を、カメラの画像平面に射影する。このとき、3次元位置がわからない対象については4.3節の方法で3次元位置を推定する。
2. 投影された重心の外接長方形 R を求める。R の中心にカメラを向けたとすると、カメラの移動量 $\Delta\theta,\Delta\phi$ は、以下のように定義される。

$$\Delta\theta = \mathrm{Tan}^{-1}\frac{c_x}{f}$$
$$\Delta\phi = \mathrm{Tan}^{-1}\frac{c_y}{f} \qquad (4)$$

ただしカメラの速度と可動範囲の制限から、

$$\theta_{\min} \le \theta + \Delta\theta \le \theta_{\max}$$
$$|\Delta\theta| \le \theta_{speed}$$
$$\phi_{\min} \le \phi + \Delta\phi \le \phi_{\max} \qquad (5)$$
$$|\Delta\phi| \le \phi_{speed}$$

である。

より多くの対象を追跡するように、常にズームアウトする。ズームの変化量 Δf は以下のように定義される。

$$\Delta f = -f_{speed} \qquad (6)$$

ただし、

$$f_{\min} \le f + \Delta f \le f_{\max} \qquad (7)$$

3. カメラを回転させたと仮定して、外接長方形 R がカメラの視野に入るか計算する。視野内に入れば、追跡可能であるとする。

4.5. タスク割り当て

タスク割り当ては各カメラが目標として与えられた観測タスクの要素がなくなるように、各カメラの追跡可能集合の中から行動を選択することで実行される。

1. 各カメラは以前に選択したタスクを観測タスクへ戻す。
2. 追跡可能集合 T_j から追跡行動 a_i を一つ選ぶ。
3. a_i に含まれる追跡対象が、少なくとも一つでも観測タスクに入っているか確認する。もし、なければその追跡行動はタスクを満たすのに貢献しないので 2 へ戻る。
4. カメラの行動として a_i を選択し、観測タスクから a_i の要素(追跡対象)を引く。

4.6. カメラパラメーターの決定

カメラが追跡すべき対象が決まったので、カメラの制御パラメーターを決定し更新する。外接長方形 R の中心を画面の中心に捕らえるように式(4)(5)より回転角度 $\Delta\theta,\Delta\phi$ を決定する。

次にズームの制御量 Δf を決定する。追跡すべきすべての対象が、視界の中心部1／4にあるときは、より詳しい情報を得るためズームインする。対象が中心部1／4の外にあるときは、対象が視界から外れるのを防ぐためズームアウトする。

ただし、

$$f_{\min} \le f + \Delta f \le f_{\max} \qquad (8)$$

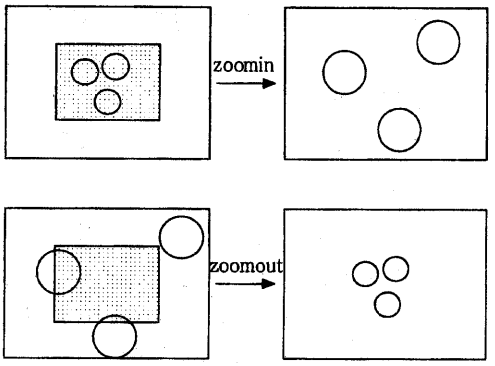

図 6:ズームインとズームアウト

以上に基づいてカメラパラメーターを更新する。

$$\theta_i(t+1)=\theta_i(t)+\Delta\theta_i$$
$$\phi_i(t+1)=\phi_i(t)+\Delta\phi_i \qquad (9)$$
$$f_i(t+1)=f_i(t)+\Delta f_i$$

5. 実験評価

5.1. シミュレーションによる実験

本アルゴリズムの有効性を検証するため、シミュレーションによる実験を行った。一辺2mの立方体の空間を設定し、3つの対象が立方体内を等速直線運動している。3台のアクティブカメラ、1台の固定カメラを配置し3つの対象をそれぞれ2台のカメラで追跡するタスクを与えた。(図 7)

アクティブカメラはパン・チルトともに±45度の範囲で5 *[deg/frame]* の速度で回転できる。全てのカメラは、$f=5.9\sim47.2\,[mm]$ の範囲でズームの制御が可能である。対象の移動速度は3 *[cm/frame]* 程度である。

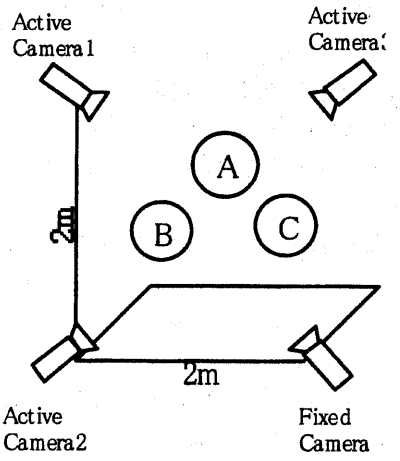

図 7:シミュレーションの初期配置

追跡した結果を図(図 8)に示す。横軸は時間、縦軸は各対象をそれぞれ何台のカメラで捉えることができたかを示す。

実験では、7フレーム目まではすべての対象を2台以上のカメラで捉えることができていた。8フレーム目で対象Cを 2 台のカメラで捉えることができなくなったために、対象 C の三次元位置を得ることができなくなった。

そこで、他のカメラが対象 C のエピポーラ線をなぞるように動くことで16フレーム目に対象 C を捉え、再び三次元位置を得ることができた。

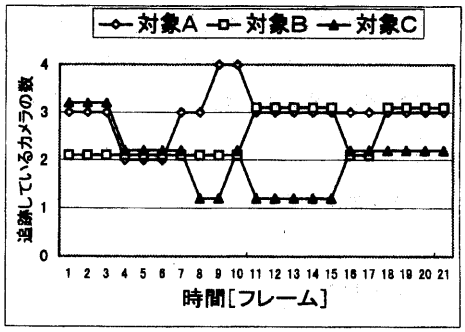

図 8:動的な役割分担による追跡結果

比較のため、静的な役割分担による追跡を行った。最初の1フレーム目で役割分担を決定し、以後その役割分担のまま追跡を行った。結果を図(図 9)に示す。

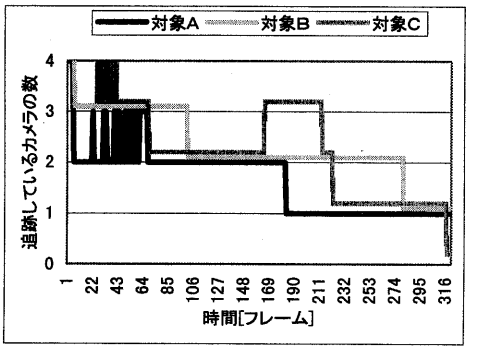

図 9:静的な役割分担による追跡結果

静的な役割分担では追跡対象がカメラの視界から外れそうになった場合、他のカメラで追跡することができない。時刻160[frame]で対象Cを追跡するカメラが2台から3台に増えているのは、対象Cが偶然カメラの視界に入ったからである。追跡するカメラが1台だけになり3次元位置がわからなくなった場合、再び2台以上のカメラで捉えることはできなかった。

5.2. 実環境での実験

2台のアクティブカメラを用い、実環境で追跡の実験を行った。アクティブカメラの特性は、前節と同じである。処理速度は1[*frame/sec*]程度である。

その結果を図(図 10)に示す。システムに対しては人物の右手を2台のカメラで追跡するようにタスクを与えた。

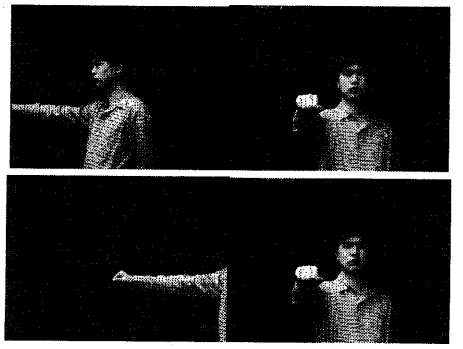

図 10:実環境での実験

最初、右手は1台のカメラでしか捉えられていなかった。(図 10上)しかし、エピポーラ線の情報を受け取ってカメラが回転することにより、2台のカメラで右手を観測し、システムに対する要求を満たすことができた。(図 10下)

6. まとめ

複数の能動的カメラを用いて複数の対象を追跡する方法について示した。対象の動きの予測を行い、三次元位置がわからない対象については位置の推定を行った。各カメラごとに追跡が可能な対象の集合をもとめ、要求されたタスクを満たす組み合わせを求めた。

複数のカメラの役割分担を動的に決定することで、要求されたタスクを満たす最適な組み合わせを求めることができた。静的な役割分担ではカメラの視界範囲から外れた対象を再び捉えることができなかったのに対し、対象が一台のカメラでしか見えていないときに三次元位置を推定することによって、他のカメラで対象を追跡することができた。

現在はソフトウェアによる画像処理で対象を発見しているので、処理速度は1[*frame/sec*]程度であり人物の動きに追従できないことがある。そこで今後の課題としては、画像処理を専用ハードウェアで行うことにより、処理速度を高速化することが挙げられる。

また、追跡中にカメラの数が動的に増減したとき、どのように処理するかも今後の課題である。

7. 参考文献

[1] Katsuhiro Matsuo, Chil-Woo Lee, Sabuto Tsuji, "Recognition of Facial Expression with Potential Net", Proc. of ACCV '93, pp. 504--507, 1993

[2] K. Sutherland, D. Renshaw, P. B. Denyer, "Probablistic Pattern Analysis for Facial Recognition", Proc. of ICARCV '92, Vol.1, CV-18.3. 1992

[3] クランポン-ユーニバン、木下 宏揚、酒井善則 "スティックモデルを用いた手振りの認識" 信学論(D-Ⅱ)、vol.177-D-Ⅱ,no.7, pp.51-60, Jan.1994

[4] 中嶋 正之、柴 広有 “仮想現実世界構築のための指の動きの検出法” 信学論(D-Ⅱ), vol.177-D-Ⅱ, no.8, pp.1562-1570, Aug.1994

[5] Yoshio Iwai, Yasushi Yagi, Masahiko Yachida, "Estimation of Hand Motion and Position from Monocular Image Sequence" Proc.of ACCV '95, pp. 79--84, 1995

[6] I. J. Kuch, T. S. Huang, "Virtual Gun-a vision based human computer interface using the human hand" Proc. of MVA '94, pp. 196--199, 1993

[7] Takashi Matsuyama, "Cooperative Distributed Vision", 1st International Workshop on Cooperative Distributed Vision, pp. 1-28, 1997

[8] 森 大樹、内海 章、大谷 淳、谷内田 正彦 “多数カメラによる人物位置・姿勢推定” 信学技報(MVE97-23), pp. 21-26, 1997