研究テーマ　人物検出に関する研究

研究者　松村　博　メディアドライブ株式会社　共同研究員
冨永将史　財団法人ソフトピアジャパン　雇用研究員

フェーズI

1 研究の概要

人間の生活を支援する空間の構築を目指し、位置に依存しない人間の動作（ジェスチャ）を認識する手法を開発するための基盤となるデータベースを構築した。更に、構築したデータベースを基に、対象人物の要望・意図から適したサービスを提供する空間の実現を目指し、パーセプトルームを構築し、壁面に設置した複数のカメラから視体積交差法を用いた人物検出、位置推定および手サイン検知手法を確立した。

2 研究の目標

近年、情報化社会の進歩が目覚しく、一般の家庭環境にも自然にコンピュータが受け入れられる環境が整ってきた。一般家庭のリビングにおいて家電製品を操作する際通常はリモコンが用いられるが、多機能化に伴いリモコンの操作手順は複雑になり、手間も増えており、高齢者社会が進行する現在、人にやさしい機器制御を望む声もあがりはじめた。そのような中、人間の身振り手振りや表情に対応して、機器の操作を可能とする技術が確立できれば、快適な生活環境を提供することができるのではないかと考えられる。そこで、対象人物の要望を察知し、意図に適したサービスを提供する空間（パーセプトルーム）の実現を目的としている。

本研究では、室内空間の位置に依存しない人間の動作（ジェスチャ）を認識する手法を開発するため、壁面に複数のカメラを設置し、同期の取れた複数の映像を一度に収集することが可能なパーセプトルームを構築した。また、研究基盤としてパーセプトルーム内で撮影・構築した複数人ジェスチャデータベースを紹介する。更に、このデータベースを基に、パーセプトルーム内で、視体積交差法を利用した対象人物の位置を推定する手法を提案する。複数カメラ統合画像から背景差分により人物領域を推定し、更にフレーム間差分を行い動作領域の推定を行う。これらの結果から視体積交差法により人物領域および動作領域の同定を行い、情報を統合することで、手サインの提示タイミングの検知と、その際の手領域の抽出を行う。

3 実施内容

3.1 パーセプトルームの構築

部屋自体が人間の行動を見守り、生活を支援するシステムを一般的にインテリジェントルームと言う。我々はパーセプトルームと名づけ、複数カメラからの映像により人間の行動を把握し、支援を行う空間を構築した。

構築したパーセプトルームにおけるカメラ配置を図 1 に示す。同期の取れたカメラ 16 台を、1 辺 460 cm の空間内に内向きに設置した。各カメラは 45 度間隔で配置し、8 台を高さ 90cm の位置に水平に、残る 8 台を高さ 220cm の位置に下方 22 度を向くように設置した。このように設置することで、人物の位置に関わらず、最低 2 台のカメラから撮影することができる。人物位置の検出のみを目的とした場合、天井カメラが有効である。リビングなどの生活空間を想定した環境では、天井の高さを考慮すると、複数台もしくは魚眼カメラを必要とする。更に、本研究では取得した映像の手サイン検出や個人識別などといった処理への利用を考慮し、真上からのアングルでは頭部しか撮影できないことから、天井カメラを採用しないこととした。

図 2 にパーセプトルームにおけるシステム構成を示す。16 台のカメラは、それぞれ 1 台のカメラ PC に接続されており、ビデオレートのカラー画像（640 × 480）を取得できる。さらに、各カメラ画像は画面合成スイッチャを通すことで、同期のとれた状態でアナログ合成されたカラー画像

（640 × 480）がメイン PC によりリアルタイムで取得できる。人物検出、位置推定、イベント検出、認識部位の領域検出などは、メイン PC にてこの統合画像を用いて行う。図 3 に統合画像の例を示す。各画像は左上から右にカメラ番号順に並ぶ。

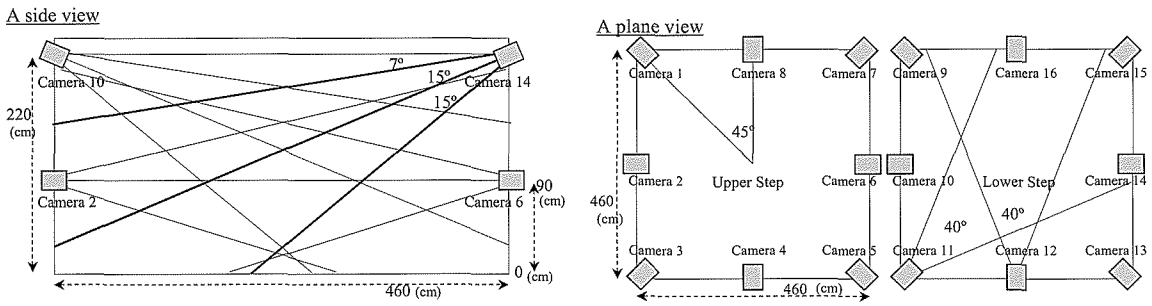


図 1. カメラ配置

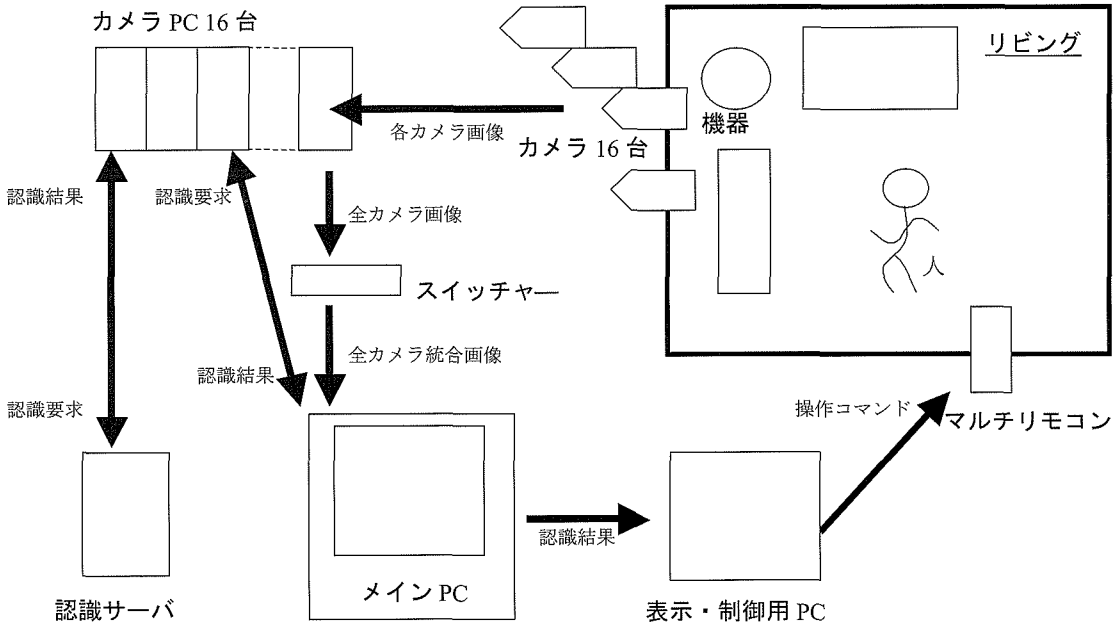


図 2. システム構成

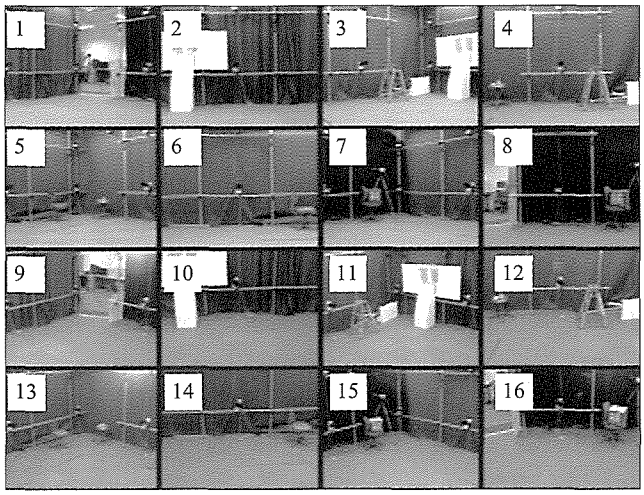


図 3. 16 台のカメラ画像からの統合画像

3.2 ジェスチャデータベースの構築

パーセプトルームにおける情報家電機器の操作を想定して、1名の人物が手サインを出すシーンおよび、複数の人物が動作するシーンを撮影し、研究基盤となるデータベースを構築した。

3.2.1 手サインデータベース

構築したパーセプトルーム内において、手を上げる前の状態から、手を上げサインを出し、手を下ろすまでの一連の動作を撮像した。表1に手サインデータベースの仕様を示す。

また、手サインの出し方に指示を出さないものとして、10回のテストデータを撮影し、その後、顔と手が重ならないようになどの指示を行い、正規データの撮影を行った。

表1. 手サインデータベース仕様

時間	4～5秒（1動作当り）
照明	明るさが均等になるよう設定
背景	ブルーバック
人数	60名（10代～60代までの6段階、男女各5名）
手サインの種類	グー、チョキ、パー、指さし（上下左右）、ポインティング、 親指を立てる、L字（親指と人差し指を立てる）の10種
提示方向	基準カメラに対して0度および22.5度の2方向
姿勢	部屋中心にて椅子の上に着席の状態
回数	2回（1動作当り）
服装	肌色、青色は避ける、袖の長さは自由、 眼鏡や腕時計などは可、手袋や帽子は不可
データフォーマット	フレームレート30F/S、フルカラー、640×480、 Motion-Jpegで圧縮（圧縮率1/3程度）
総データ数	3000（10種×2方向×1姿勢×2回＋10テスト）

3.2.2 複数人ジェスチャデータベース

構築したパーセプトルーム内において、決められたシナリオに従い複数の人物が動作するシーンを撮像した。表2に複数人ジェスチャデータベースの仕様を示す。また、表3に規定および5つの行動のシナリオ一覧を、表4および図4にシナリオ毎の行動指定を、図5にタイムラインを示す。なお、提示する手サイン（α、β、γ、δ、ε）は撮影によってシフトする（1回目のA：$\alpha \rightarrow \beta \rightarrow \gamma \rightarrow \delta \rightarrow \varepsilon$、2回目のA：$\varepsilon \rightarrow \alpha \rightarrow \beta \rightarrow \gamma \rightarrow \delta$、3回目のA：$\delta \rightarrow \varepsilon \rightarrow \alpha \rightarrow \beta \rightarrow \gamma$のように順に変更）。

加えてシナリオ6として自由動作（3分間 約690M）を撮影した。自由動作シナリオでは以下条件内で自由に手サインを出すよう指示した。なお、紛らわしい動作の排除、対象と手サインの明確化のため、記録の際のイベント情報は映像を見せながらの聞き取りを行った。

自由行動の条件：

① 手サインは表2で定めた5種のみ
② 手サイン提示時は対象を注目すること
③ 手サイン提示時は身体を停止すること（歩きながら等の手サインは認めない）
④ 手サインは一定時間（1秒以上）出すこと（瞬間的な手サインは認めない）
⑤ 手サインは必ず利き手で（両手で同時に出したりしない）
⑥ 寝ない
⑦ 他の人物と接触しない
⑧ 室内への出入りは自由（入室タイミングはある程度指示する）

以上、構築に際して、歩行、停止、座る、手サイン、対象物、人物をシーン毎に明記し、人物特徴として身長を記載した。また、リビングを想定するため靴は脱ぐこととした。

表 2. 複数人ジェスチャデータベース仕様

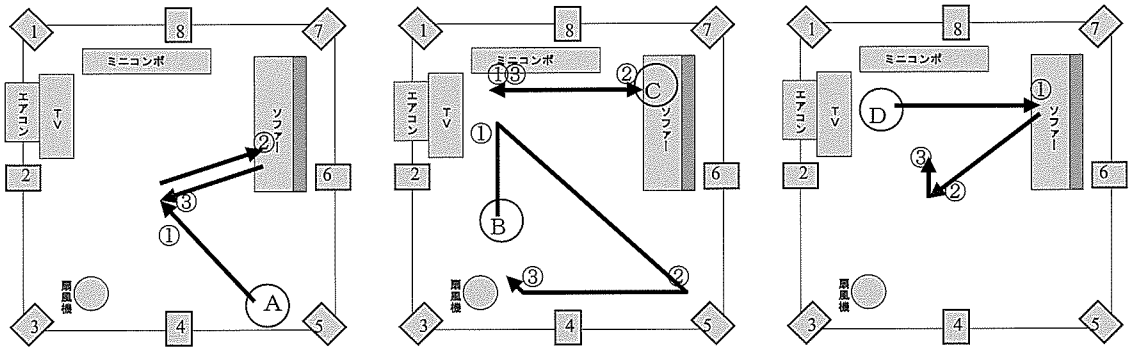
時間	1分（1パターン当り）
照明	蛍光灯のみ
背景	ブルーバック
人数	10名
手サインの種類	指0本（グー）、指1本（人差し指を立てる）、指2本（チョキ）、指5本（パー）、指さし（正面へ）、の5種
配置家電・家具	ソファー、TV、ミニコンポ＋テーブル、扇風機、エアコン（TVの上の空間に設置）、なおこれら家電は稼動しない
データフォーマット	フレームレート 30F/S、フルカラー、640×480、 動画 AVI Motion-Jpeg 圧縮（圧縮率 1/3 程度）、 プログレッシブスキャン、シャッタースピード 1/60
総データ量	35×データ量（1 AVI 約 230M）＋背景のみ 10 秒（38M） ＝ マシン1台につき約8G：　×16台 ＝ 合計約 128G ：＋自由動作シナリオ ＝ 総合計約 129G

表 3. 複数人ジェスチャデータベース シナリオ

シナリオ	人数	サイン提示者	静止	動作	備考	計
規定	1	固定 orA、D			×10人 ×2パターン	20
1	3	A	B C		×3パターン	3
2	3	A	B	C’	×3パターン	3
3	3	A		B’ C’	×3パターン	3
4	3	A D同時	B		×3パターン	3
5	3	D☆ A☆交互	B		×3パターン	3
6	10	全員	全員	全員	自由動作	1
（注☆シナリオ5： ☆の後に続く番号の動作を順に行う）					合計	36

表 4. 複数人ジェスチャデータベース シナリオ別行動指定

人物	行動	移動先	手サイン方向	手サイン	注
A		①部屋中心	ミニコンポ	α：指0本グー	☆2
		同位置	ミニコンポ	β：指2本チョキ	
		②ソファー座	TV	γ：指5本パー	☆4
		同位置	エアコン	δ：指1本	
		③部屋中心	扇風機	ε：指差し	
B		静止（部屋左下立）	無し	無し	
B’	三角形に移動	①～③	無し	無し	
C	ソファー（座る）	静止	無し	無し	
C’	ソファー（座る）とTV前の移動	①～③	無し	無し	
D		①ソファー座	TV	β：指2本チョキ	☆1
		同位置	エアコン	γ：指5本パー	
		②部屋中心	扇風機	δ：指1本	☆3
		同位置	扇風機	ε：指差し	
		③部屋中心上	ミニコンポ	α：指0本グー	☆5



(a) 人物 A の行動　(b) 人物 B、C の行動　(c) 人物 D の行動

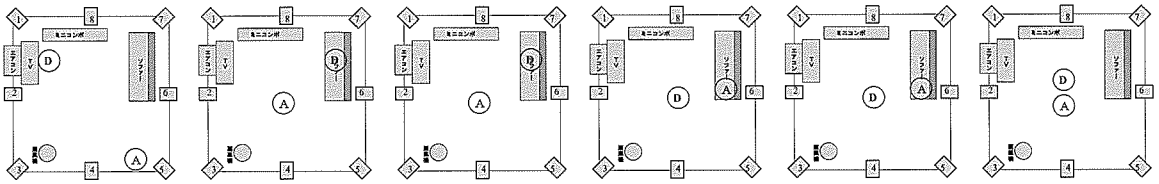
図 4. 複数人ジェスチャデータベース シナリオ別行動位置

シナリオ 1～3

A	①へ 移動	ミニコンポ グー		ミニコンポ チョキ	②へ 移動	TV パー		エアコン 指１本	③へ 移動	扇風機 指差し
B’	①へ 移動	②へ 移動	③へ 移動	①へ 移動	②へ 移動	③へ 移動	①へ 移動	②へ 移動	③へ 移動	①へ 移動
C’	①へ 移動	②へ 移動	③へ 移動	②へ 移動	③へ 移動	②へ 移動	③へ 移動	②へ 移動	③へ 移動	②へ 移動

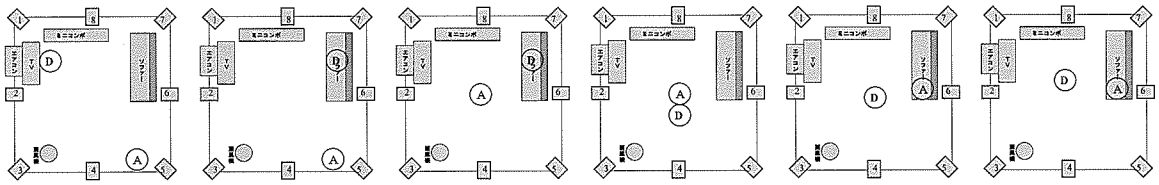
シナリオ 4

A	①へ 移動	ミニコンポ グー		ミニコンポ チョキ	②へ 移動	TV パー		エアコン 指１本	③へ 移動	扇風機 指差し
D	①へ 移動	TV チョキ		エアコン パー	②へ 移動	扇風機 指１本		扇風機 指差し	③へ 移動	ミニコンポ グー



シナリオ 5

A			①へ 移動	ミニコンポ グー	前へ 移動		②へ 移動	TV チョキ		
D	①へ 移動	TV パー			②へ 移動	扇風機 指１本			③へ 移動	ミニコンポ 指差し



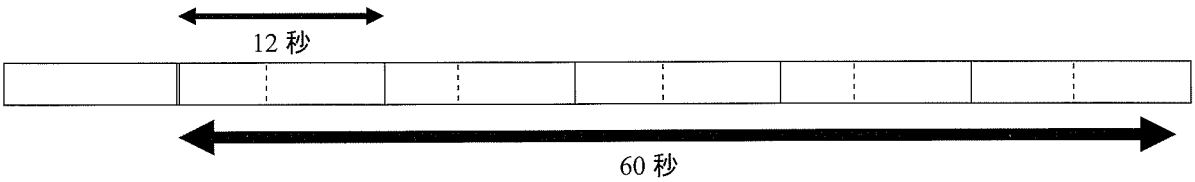


図 5. 複数人ジェスチャデータベース シナリオ別タイムライン

3.3 人物検出手法および位置推定手法の開発

室内空間の位置に依存しない人間の動作（ジェスチャ）を認識するためには、3 次元的な人物位置を特定する必要がある。そのため、パーセプトルームにおける人物検出および位置推定手法の開発を行った。本手法の開発では、パーセプトルームに設置した 16 台のカメラからのリアルタイム映像および、複数人ジェスチャデータベースの規定シナリオ（単数人の手サイン提示映像）を用いた。

16 カメラ統合画像に対して、事前に取得しておいた背景画像との差分、2 値化を行い、各カメラ画像の領域毎に x 軸方向、y 軸方向への射影処理を行う。射影ヒストグラムを走査し、x 軸方向、y 軸方向共に画素数が閾値を越えた矩形領域を人物候補領域とする。x 軸方向の候補領域において視体積交差法を適用して人物領域の平面位置を推定する。視体積交差法やシルエット法は、主に 3 次元形状復元に用いられる手法であるが、本研究の目的では、人物の位置推定と動作推定に用いるため、詳細な形状を得る必要はない。そのため分解能を荒く（総ボクセル数を少なく）設定でき高速化が期待できる。なお、視体積交差法を用いる際、全てのカメラパラメータ（カメラ位置、パン・チルト角、焦点距離等）は固定で既知とする。

視体積交差法ではカメラからの 3 次元的なボクセル投影（投票）を行うが、ここではまず x 軸方向の射影ヒストグラムから得られた人物候補領域において平面的な投影を行い、投票を行う範囲を限定する。

投影平面 F（平面サイズ $N\times M$）に対して、図 6 に下段 8 カメラから投影可能な領域 C_l(l=1,2,…,8 : カメラ番号)を示す。投影は各カメラ位置から、カメラに相対する壁側へ錐状の領域に行われる。各カメラからの投影可能領域 C_l は

$$C_l \subseteq F \tag{1}$$

であり、投影平面 F は 8 つのカメラのうち少なくとも 1 つにより投影、

$$C_1 \cup C_2 \cup \cdots \cup C_8 = F \tag{2}$$

が成り立つ。図 7 に投影した様子を示す。ここで濃淡値は、カメラからの投影領域の重なりを示す。

通常視体積交差法による形状復元では、精度を要求するため、投影可能領域の積

$$C_1 \cap C_2 \cap \cdots \cap C_8 \tag{3}$$

即ち全カメラから見える範囲である中央のみを有効とし、実際に投影される領域 T_l(l=1,2,…,8 : カメラ番号)の積

$$T_1 \cap T_2 \cap \cdots \cap T_8 \tag{4}$$

により物体形状を求める。これに対し本手法では、室内空間（投影平面）における位置の推定を目的とすることから、平面全体を網羅する式(2)の範囲を全て有効として投票を行う。

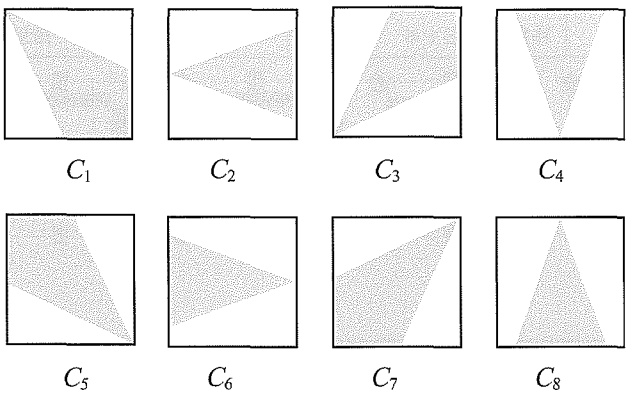


図 6. 下段 8 カメラの投影可能領域

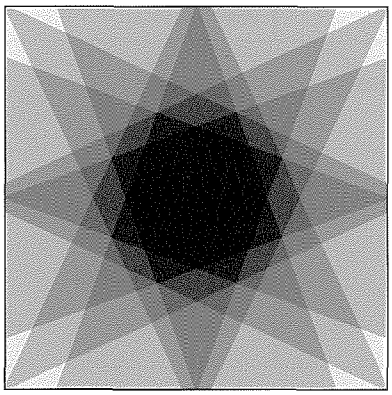


図 7. 8 カメラからの投影可能領域

さらに投影結果から領域を絞り込むため、射影領域範囲外への投影部分になる領域を削除する。各カメラからの投影可能領域 C_l のうち、投影領域 T_l を除いた非投影領域（例：図 8）

$$D_l = C_l - T_l \quad (l=1,2,\cdots,8) \tag{5}$$

はすなわち、抽出目的である人物が存在しない背景領域の範囲であるため、削除が可能である。

このようにして、限定した領域の空間にのみ 3 次元的な高さを考慮した視体積交差法による投票を行い、人物位置を求める。図 9(a)に 8 カメラからの投影領域 T_l の例を、図 9(b)に非投影領域 D_l を削除した投票限定領域 G を示す。

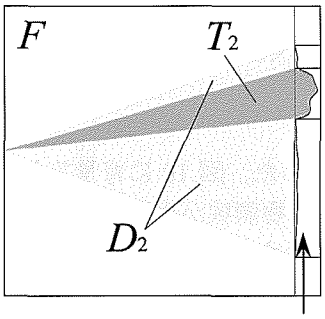
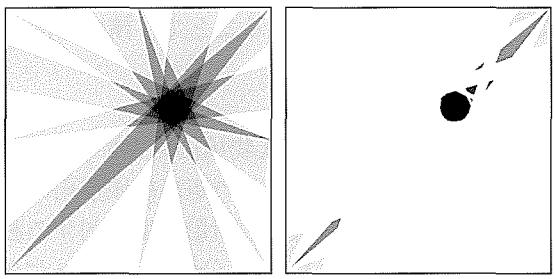


図 8. C_2 における削除可能領域 D_2 の例



(a) 投影領域 T_l　　(b) 投票限定領域 G

図 9.　8 カメラからの投影

投票によって得られた複数のボクセルに対し、ボクセル数 1 の孤立ボクセルを除去し、ラベリングを行うことで、固まり（ブロック）毎に分ける、更に、体積比較を行い、ブロックの体積が最大であり、前フレームでの人物重心位置に最も近い重心位置のブロックを抽出目的の人物とし、このブロックの重心位置を現在の人物位置とみなす。

16 カメラ統合画像において、射影により抽出された複数の矩形領域の同一人物対応を、視体積交差法による投票結果を基に推定する。各領域の中心(x_o,y_o)（もしくは領域内の数点）に対し、カメラ位置(x_k, y_k) $(k=1,2\cdots8$：カメラ番号)から伸ばした直線

$$y = \left(\frac{y_k - y_o}{x_k - x_o} \right) \times (x - x_o) + y_o \tag{6}$$

が最初に接触したブロックの座標(x_a, y_a)のラベルを領域に対応付ける。

$$region = Label\left[\min\left((x_k - x_a)^2 + (y_k - y_a)^2 \right) \right] \tag{7}$$

全ての領域に対して処理を行い、最終的に人物と見なされているブロックと同一のラベルに対応付けられた全ての領域が、各カメラ画像での同一人物領域となる。

3.4 手サイン検知手法の開発

人物位置に続き、更に手サインを認識するためには、3 次元的な手位置および、映像中の手領域を特定する必要がある。そのため、パーセプトルームにおける手サイン提示タイミングの検知手法および手領域抽出手法の開発を行った。

背景差分では検出できない動き領域の位置推定を行うために、フレーム間差分を用いて動き検出を行い、同一動領域の推定のため視体積交差法を使用する。このように動領域の室内空間位置を推定することで、背景差分のみでは得られない手領域の動きのような身体の部分的な動作が推定できると考えられる。

まず、各カメラ画像において、同一人物として対応付けられた矩形領域の面積比により移動状態か停止状態かを判別する。なお、時間推移における人物対応は、ブロック重心座標の変化が最も少ないブロックを移動後の同一人物として採用することとした。次に、矩形領域の重心位置を比較し、人物が停止状態の時に、高い座標位置に動領域の重心が現れた場合、上半身のみを動かしていると

みなして手サイン提示の前段階とし、この先頭フレームを手サイン提示開始タイミングとする。この前段階中に得られた動領域の位置から若干広い範囲を、手サイン提示中の手領域とみなす。

手サインの認識要求を出す際、正面に近いカメラ画像を取得するカメラ PC を選択し、このカメラ PC にのみ認識要求を出す。

視体積によって得た人物ブロックの、現フレームと 5 フレーム前での重心位置をもとに、移動方向を推定する。移動方向を人物正面とみなすことで、背後からのカメラ PC を選択対象から外す。また、動領域が検出されないカメラ画像領域に対応するカメラ PC も、手が映っていないとみなし、選択対象から外す。更に、人物重心位置と動作部分の重心位置の相対位置関係から手の提示方向を推察することも可能であると考えられる。

複数のカメラ PC に対して認識要求を出すことになるが、各カメラ PC の認識結果から多数決や重要度にもとづき情報を統合することで、精度の高い最終結果を導く。

4 結果

パーセプトルームにおいて人物検出および位置推定の実験を行った。図 10 に背景差分による候補領域抽出結果を示す。射影からの矩形領域抽出のみでは、同一人物領域が複数に分割される場合がある。図 11(a)にカメラ位置から射影領域に対する投影領域 T_1 を示す。このときボクセル空間サイズは $(X, Y, Z) = (46, 46, 22)$ とし、分解能は 10cm に設定した。なお図中には、理解し易いよう 2, 4, 6 カメラに正対する壁の位置に、カメラから得られた画像を表示した。図 11(b)に非投影領域 D_1 を削除した投票限定領域 G を示す。このときの有効空間への投票結果を図 12 に示す。空間中に縦長のブロックが確認できる。図 13 にブロックに対して領域を対応付けた結果を示す。カメラ 8 およびカメラ 14 で 2 つに分かれていた人物候補領域が同じ領域として対応付けられ、結合されているのが確認できる。

図 14 に図 10 と同時刻におけるフレーム間差分による候補領域抽出結果を示す。また、図 15(a), (b)に投影領域 T_1 と投票限定領域 G を示す。更に図 16 に有効空間に投票した結果を示す。このとき手の部分のみを動かしているため、ブロックは上方に浮いた形で抽出できた。図 17 にブロックに対して人物候補領域を対応付けた結果を示す。人物候補領域がうまく対応付けられていることが確認できた。

図 18(a)に手サイン提示開始タイミングとして抽出されたフレーム画像を、図 18(b)に手サイン提示中として抽出されたフレームを示す。図中白線で囲われた領域が最終的に手領域として抽出された領域である。太い白線で囲まれたカメラ画像領域は最終的に正面に近いカメラとして選ばれたものであり、これに対応するカメラ PC にのみ、抽出した手領域の座標を送り、認識要求することになる。本実験では、立った姿勢でのサイン提示は良好に抽出できたが、座った姿勢でのサイン提示では動領域が小さくなることから、若干不安定であった。

なお、今回の実験ではメイン PC として Pentium4 1700MHz 256MB Windows2000 SP2 を使用した。1 フレーム当りの処理時間は平均 150msec であった。

今後、パーセプトルームを現実のものとする上で、複数人に対応していく必要がある。

図 10. 背景差分による候補領域抽出結果

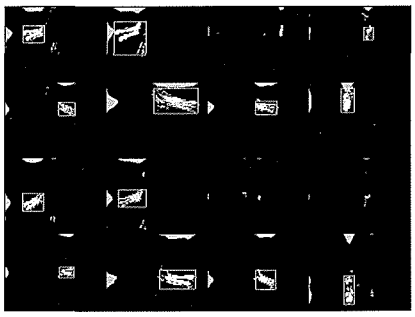
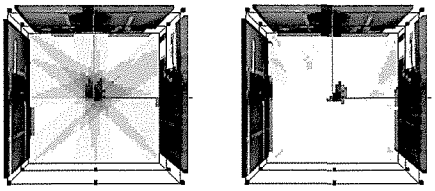
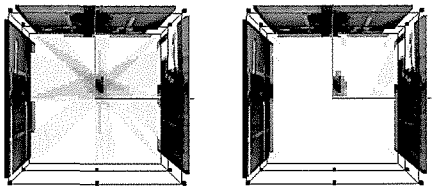


図 14. フレーム間差分による候補領域抽出結果



(a) 投影領域 T_I　　(b) 投票限定領域 G

図 11. 背景差分からの投影



(a) 投影領域 T_I　　(b) 投票限定領域 G

図 15. フレーム間差分からの投影

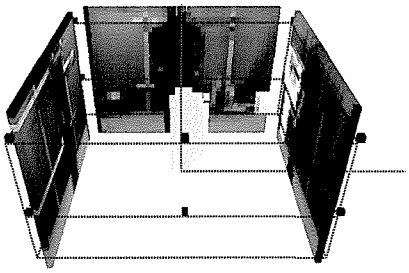


図 12. 背景差分からの投票結果

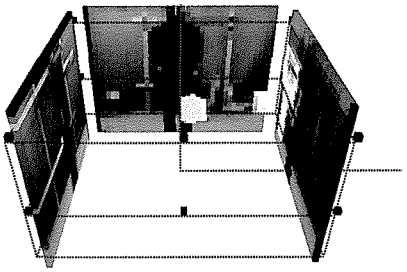


図 16. フレーム間差分からの投票結果

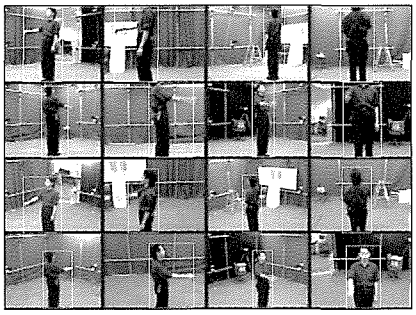


図 13. 背景差分による人物領域抽出結果

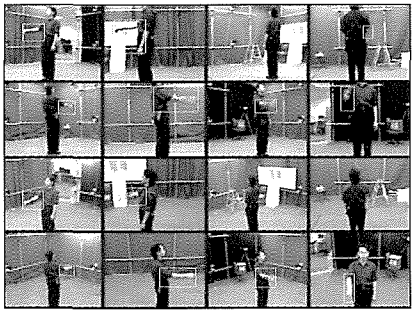
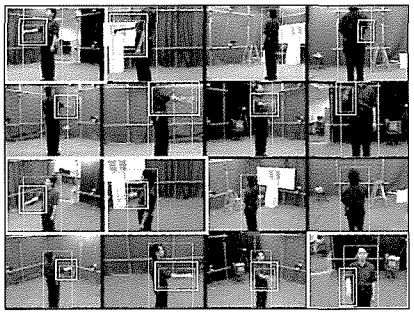
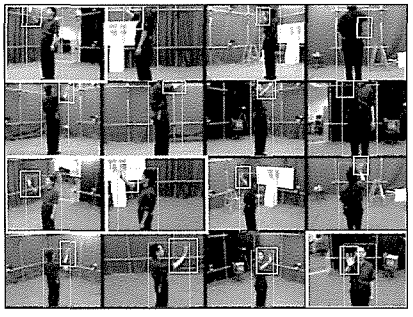


図 17.フレーム間差分による動域抽出結果



(a) 手サインの提示開始フレーム



(b) 手サイン提示中の手領域

図 18. 手サイン提示タイミング検知結果

フェーズII

1 研究の概要

フェーズIでは、パーセプトルームおよびジェスチャデータベースを構築し、視体積交差法を利用した対象人物の位置推定手法および、手サインの提示タイミングの検知と、手サイン領域抽出手法の技術開発を行った。次いでフェーズIIでは、パーセプトルームの実現へ向けて人物検出・位置推定手法を複数人に対応した。視体積交差法に存在確立マップを導入することで、ロバストな人物検出手法を確立した。同時に室内に存在する他の人物が動作している際の手サイン提示タイミングの検知および手領域検出精度を向上した。

2 研究の目標

リビングなど複数人が共存する空間において、人物のジェスチャ認識を行うためには、対象となる人物に注目し、その人物の空間的な位置と動作部位を特定・検出する必要がある。また、複数カメラを用いる場合、カメラ配置によっては身体全体が必ずしも一つのカメラ画像内に収まるとは限らないことや、認識に最適な画像が常に得られるとは限らないため、各人物領域と動作領域のカメラ間の同定を行うことが重要である。本研究では、室内空間の位置に依存しないジェスチャ認識を行うための領域検出手法として、視体積交差法に存在確率なる概念を取り入れることで、空間的な人物位置と動作部位を特定し、ロバストな複数人物の検出と、複数カメラ画像間の人物同定を実現する。

3 実施内容

3.1 存在確立による複数人物検出手法の開発

フェーズIで確立した視体積交差では雑音ボクセルの検出が問題となり、室内に複数の人物が居る場合、人物同士の分離や精度の高い検出が困難になる。そこで、視体積交差法により検出した各人物の人物ボクセルに対し、存在確率マップを適用することで、雑音ボクセルを削除し、人物毎のボクセルの固まり（人物ブロック）の安定した検出を可能とする。さらに、カメラ位置から人物領域に対して、最短距離の人物ブロックを探索することで、カメラに映る人物領域の同一人物の同定ができる。人物の位置推定を空間的に行うため、カメラ配置と人物位置によっては身体の一部しか撮影されないようなカメラ画像においても、カメラ間の同一人物の同定を行うことができる。また、この視体積交差を同定した人物領域毎に、フレーム間差分による動作領域に対して行い、各人物の動作ボクセルを同時に検出することで、複数の人物を同時に、動作部位の検出と、動作領域の特定を行うことができる。

事前に取得した背景画像との差分、2値化、ラベリングを行い、カメラ画像毎に人物候補領域を検出する。検出された人物候補領域において、ボクセル空間（空間サイズ $K \times L \times M$）に、視体積交差法による複数カメラからの3次元的なボクセル投票を行う。通常、視体積交差法による形状復元では精度を高めるため、全カメラから見える積の範囲である中央のみを有効とするが、本研究では、室内空間全体を網羅する必要があるため、和の全範囲を有効とする。そのため、人物位置によっては検出精度が下がり、人物以外の雑音の検出が増加する。複数人物の存在する空間において視体積交差を行うと、オクルージョンにより人物以外のボクセル（雑音ボクセル）を検出しやすくなる。さらに、人物同士の位置が近い場合や、ボクセル空間の分解能が粗い場合、人物ボクセルとオクルージョンによる非人物ボクセルが連結する問題が生じる。これらの問題に対し、ボクセル空間内での平面位置における人物の存在位置確率と、ボクセルの高さの累積を掛け合わせることによる、空間的な存在確率マップを提案する。この存在確率マップを用いることで、動作に伴うボクセルは損失せずに雑音ボクセルのみを除去し、安定した人物ボクセルの検出を行うことができる。

各ボクセルにおける平面位置の存在位置確率は、サンプリング間隔が人物の移動に対して十分短ければ、前フレームの位置に近いほど高く、離れるに従い低くなると仮定できる。さらに、人物同士が近接する場合、他の人物の存在位置確率が高い位置では、自分の存在できる確率はその分抑制

されると考えられる。そのため、相手の存在によって各自の存在位置確率は減少すると仮定する。このような仮定から、ある人物 i が位置(x, y)で存在する可能性を $P_i(x, y)$ (i=1,2…N：人数)としたとき、各ボクセル位置(x, y)における人物の存在位置確率 $PM(x, y)$を、

$$PM(x,y) = \max_{1\le k\le N} P_k(x,y) \times 2 - \sum_{i=1}^{N} P_i(x,y) \qquad (8)$$

のように定義する。分布のモデル化は、分解能の粗さによる効果の期待値と処理速度を考慮し、前フレームの人物位置を基準に単調減少に設計した。図 19 (a)に 2 名を例に模式的に表した確率分布を示す。さらに、確立分布に対し移動方向を考慮した変形を行う。人物の推定位置を基準にして、確立分布の適用範囲をスライドするように設計した場合、人物の急激な移動方向の変化に対応できなくなる。また、推定位置や推定方向に存在する雑音や他の人物に強く影響をうける可能性がある。そのため存在確立分布の適用範囲を、現在の位置から 1 フレーム以内に人物が移動可能な距離を十分に含むように設定し、存在確立のみを推定位置を基準に変形する。すなわち、図 19 (b)に示すように、現在の位置から移動可能な範囲に相当する円錐の底辺は変形させず、存在確立の最高位置である頂点を次フレームにおける推定位置になるような変形を行う。以上のような存在確立分布により、低解像度のボクセル空間においても、人物同士の分離が良好に働き、複数人物の検出を精度よく行うことができる。

次フレームにおける人物位置の推定では、ボクセル空間の分解能を粗くした場合、瞬間速度と加速度を用いると 1 フレームでの移動ボクセル数が 1 ボクセルに満たない事があり、安定した推定を行うため、過去 3 フレームの平均速度(v_x, v_y)と平均加速度(a_x, a_y)を算出する。存在位置確率の分布は、前フレームでの人物位置(fx, fy)より、t フレーム後の予測人物位置

$$(dx, dy) = \left(fx + v_x t + \frac{1}{2} a_x t^2, fy + v_y t + \frac{1}{2} a_y t^2 \right) \qquad (9)$$

から、この座標点を存在位置確率の頂点とした単調減少の形をとる。存在位置確率の適用範囲は、移動速度に関わらず一定とし、存在する確率のみを変形させる。以降このように生成した存在位置確率の分布を存在位置確率マップと呼ぶ。図 20 に存在位置確率マップの例を示す。

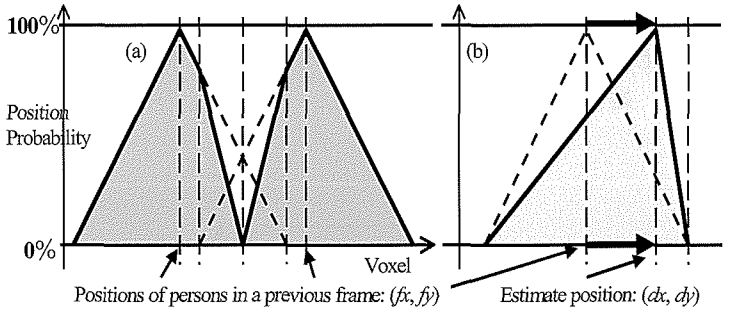


図 19. 複数人物による存在位置確率

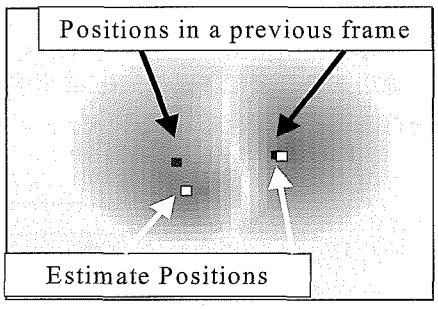


図 20. 存在位置確率マップ

高さの累積（図 21）を、現フレームのボクセル投票結果から求める。平面的な各座標位置(x, y)において、ブロックの存在するボクセルの高さ $V(x, y, z)|_z$（ボクセル空間サイズ $K\times L\times M$）の累積

$$HM(x,y) = \sum_{j=1}^{M} V(x,y,j)\Big|_j \quad s.t. \quad V(x,y,j) \neq 0 \qquad (10)$$

を算出する（以降、高さ累積マップと呼ぶ）。図 22 に高さ累積マップの例を示す。

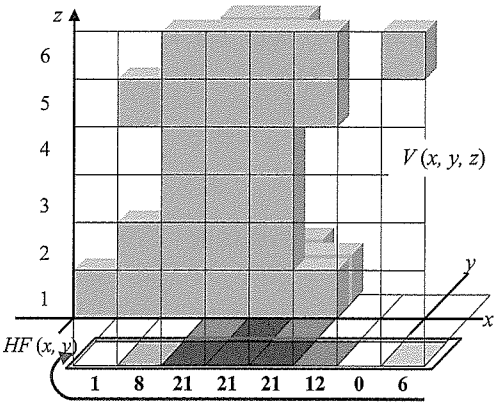


図 21. 高さの累積

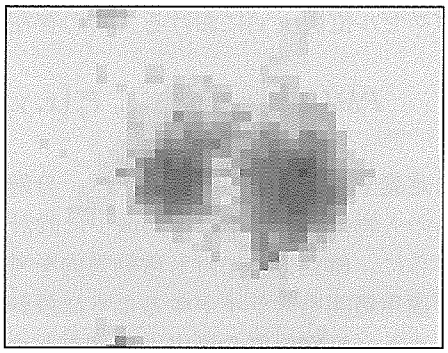


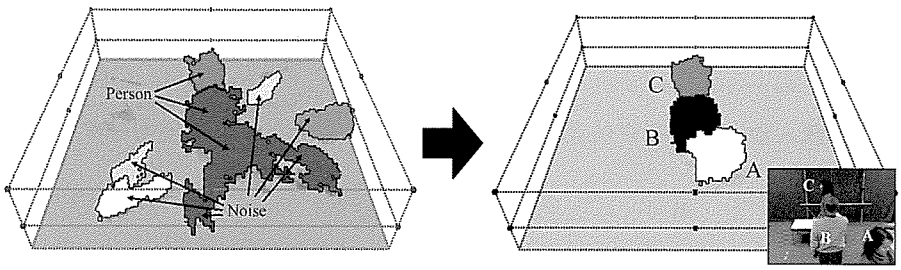
図 22. 高さ累積マップ

これら存在確率マップ *PM* と高さ累積マップ *HM* を掛け合わせ、存在確率マップを構築し、閾値 *th* 以下

$$EM(x,y) = PM(x,y) \times FM(x,y) \leq th \tag{11}$$

となる座標位置(*x*, *y*)に存在するボクセルを除去することで、雑音の除去を行う。高さ累積マップの影響により、空中に浮かんで現れる身体の一部（腕など）は除去されず、床近くに存在する物や人物の影などによる雑音が除去できる。なお、カメラ配置により、位置によって撮像可能なカメラ数が異なり、投票度数に格差が生じる。そのため、撮像可能カメラ数でボクセル空間の正規化を行った後、閾値処理を行う。閾値 *th* はボクセル空間の分解能に対して、検出対象となる人物のサイズと設置カメラ台数から適切に設定する必要がある。閾値 *th* を低く設定すると検出可能な物体のサイズが下がるため、雑音ボクセルを含めた誤検出が発生しやすくなる。以上の処理を施したボクセルに対し、ラベリングを行うことで固まり（ブロック）毎に分ける。このブロックのうち、前フレームでの人物重心位置に最も近い重心位置のブロックを、検出目的の人物のブロック（人物ブロック）とし、このブロックの重心位置を現在の人物位置とみなす。

視体積交差で得られた結果に対して、存在確率マップを適用しない場合と適用した場合の結果をそれぞれ図 23 (a)、(b)に示す。図 23 (a)に示すように、存在確率マップを適用しない場合、人物同士の近接時に雑音ボクセルが発生し、人物同士のボクセル分離ができない。対して、存在確率マップを適用することで、図 23 (b)に示すように、雑音ボクセルの除去と共に、人物同士のボクセル分離を行うことができる。



(a) Without Existence probability map　　(b) With Existence probability map

図 23. オクルージョンによる雑音と分離結果

複数の人物が存在する場合、カメラアングルによってオクルージョンが発生じ、人物の映り方に影響する。対象人物の動作を認識するためには、その人物が撮像されている領域を特定することが必要である。

本手法では各カメラ画像において、背景差分により得られた人物領域の同一人物同定を、視体積交差法による投票結果を基に推定する。図 24 に示すように、カメラから投影される画像の各人物候

補領域の中心$(x_\alpha, y_\alpha, z_\alpha)$ 方向に、カメラ位置$(x_\beta, y_\beta, z_\beta)$ (β=1,2…16：カメラ番号)から伸ばした直線

$$\left(\frac{x-x_\alpha}{x_\beta-x_\alpha}\right)=\left(\frac{y-y_\alpha}{y_\beta-y_\alpha}\right)=\left(\frac{z-z_\alpha}{z_\beta-z_\alpha}\right) \tag{12}$$

が最初に接触した（カメラ位置に最近の）ブロックの座標$(x_\gamma, y_\gamma, z_\gamma)$ (γ=A,B…：ブロックラベル)のラベルγを領域に対応付ける。

$$region_{\_rabel}=\gamma \quad s.t.\left[\min\left(\left(x_\beta-x_\gamma\right)^2+\left(y_\beta-y_\gamma\right)^2+\left(z_\beta-z_\gamma\right)^2\right)\right] \tag{13}$$

以上の処理を全カメラ画像の人物候補領域に対して行うことで、各カメラ画像間での同一人物の対応付けを行うことができる。しかし、複数人物が一つの人物候補領域に結合している場合、対応付けが正しく行われない。そこで、領域内の画素全て（もしくは数画素間隔）に対して求めることで、各画素における人物同定を行う。

3.2 動作イベント検出手法の開発

本研究では、室内に存在する複数人物の動作の推定を行うために、動作の区切と考えられる「動作停止」、「動作開始」、「挙手」を動作イベントとして検出する。

背景差分による人物ブロックの検出と同様、フレーム間差分により、視体積交差によるボクセル投票を行うことで、人物毎の動作によるボクセルの固まり（動作ブロック）を検出できる。各カメラ画像において人物検出手法の適用によって同一人物と対応付けられた領域毎にフレーム間差分を行い、各人物の動作候補領域を検出する。検出された動作候補領域において、それぞれ人物ブロックの検出と同様に視体積交差を行い、動作ブロックを検出する。動作ブロックのラベリングは投票時の人物と同じラベルを付ける。このように動作ブロックを人物毎に検出することで、対象人物の動作部位を特定し、この動作を分析することで、動作認識が可能となる。

図 25 に人物ブロックおよび動作ブロックの検出例を示す。図中、床面に映る濃淡は存在確率マップである。検出した人物ブロックと動作ブロックを比較することで、対象人物の状態や行動を推定することができる。人物ブロック *HB* が検出され、同じ位置に同程度の体積の動作ブロック *MB* を検出した場合、移動（歩行）状態と判定でき、未検出の場合、停止状態と判定できる。また、図中黒枠で示したブロックのように、空中に浮かぶ動作ブロックを検出することで、上半身の一部の動作状態を検出することも可能である。ジェスチャに対し、実験データを基にした人物ブロックと動作ブロックの判定パラメータを適切に定めることで、ジェスチャの推定が可能となる。さらに、検出した動作ブロックに対して、各カメラ画像における領域を特定することができる。したがって、同一人物の動作領域をカメラ毎に検出することで、例えば手の部分のみが映るカメラ画像も、誰がどのような動作を行っているかの判断に有効に利用できる。

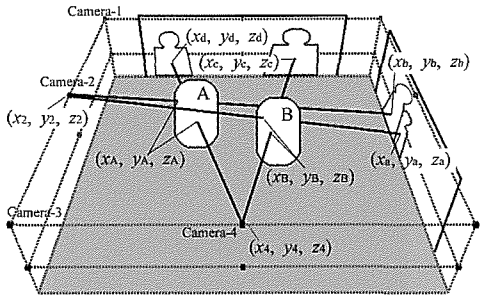


図 24. 人物領域と人物ブロックの対応づけ

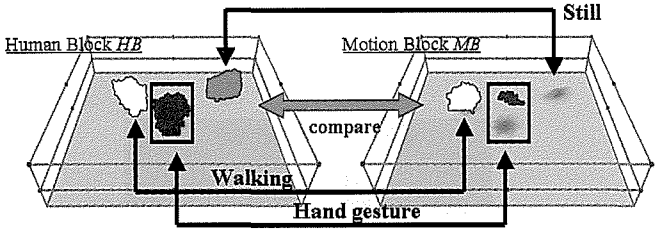


図 25. 人物領域と人物ブロックの対応づけ

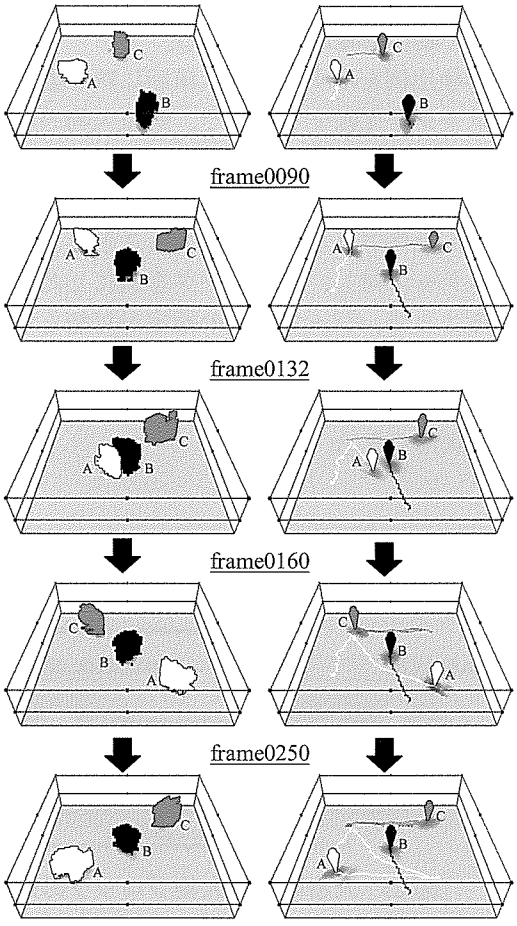
4 結果

複数人物の検出、画像間の人物同定および、動作イベント検出の実験を行った。

ボクセル空間サイズは $(X, Y, Z) = (92, 92, 22)$ とし、水平方向の分解能を 5 cm に設定し、垂直方向の分解能は、上下に異なる人物として分ける可能性が少ないことから、幾分粗く 10 cm に設定した。また、存在確率マップにおける人物位置からの適用範囲は 12 ボクセル（60 cm 相当）に設定した。これは人間の歩く速度を 1 m / sec、人間の身体の幅を約 40cm としたとき、100msec（1.5 フレーム）要する距離である。

被験者 3 名が室内で移動し、内 1 名が途中立ち止まり手サイン提示の動作を行った。図 26 に一連の検出結果（250 フレーム : 16.65 秒間）を示す。図 26 (a)列は人物ブロックの推移を示し、図 26 (b)列はブロックの重心位置から求めた人物検出結果（円錐は人物位置）を示す。図中において同一ラベル及び同色のブロックが同一人物である。人物 A は三角形を描くように移動し、人物 B は室内中央で停止、人物 C は左右に往復移動を行う、移動軌跡が確認できた。また、図 27 (a)〜(c)に人物同定を行った結果を示す。図中、白線矩形領域が各人物に同定された領域を示す。2 番カメラにおいて、人物 A の後ろに見える人物 B の腕のみの領域も、人物 B として正しく同定された。

本実験において初期状態（開始後数フレーム）は、存在確率マップを適用せず、雑音ボクセルの含まれる結果から位置を推定し、以降この推定位置を基に存在確率マップを展開した。そのため、初期時の人物配置によっては雑音ボクセルが多く発生し、人物検出および位置推定に失敗する場合もある。失敗時における再検出、再位置推定には、人物の動きに着目し、検出された人物ブロックと共に、フレーム間差分による動作ブロックなどを利用することが考えられる。これら初期状態の適切な設定法と失敗時の回復方法は今後の課題である。

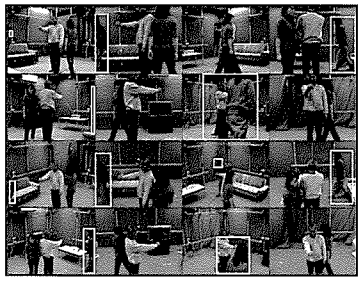


(a) Blocks of person (b) Results of the tracking

図 26. 複数人物検出結果

(a) Regions of the person A

(b) Regions of the person B



(c) Regions of the person C

図 27. 複数人物の検出結果

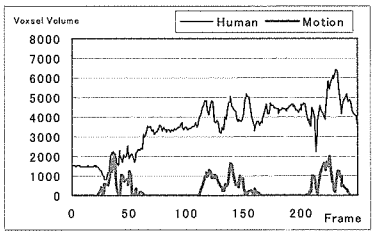
図 28 に各人物の人物ブロックと動作ブロックの体積の推移を示す。人物の動作に伴い動作ブロックの体積の増減が確認できた。位置の変動に伴う、動作ブロックの発生フレームを動作開始イベント、消失フレームを動作終了イベントとすると、動作の推移は図 29 のようになる。ブロックの体積の変動は、位置と人数と動作に大きく影響を受けるため、適切な閾値の設定が必要となる。なお実験では、閾値を 0 とし、 連続 3 フレーム以上の検出を動作イベントの変わるタイミングとした。

人物ブロックの最上値 HB_h は身長と定義可能であることから、HB_h を基準に人物ブロックと動作ブロックを比較することで、行動を推定することができる。人物 C において検出ブロックの最上値 HB_h（図 30）を参照すると、フレーム 80～125 間は座っていると推定できる。また、人物 B において、動作ブロックの重心の高さ（図 31）を参照すると、フレーム 109～116 間に段階的な上昇が、フレーム 160～169 間に下降が確認できる。例えば人間のサイズを八頭分すると、腕の存在する範囲は身体上半分の頭部を除いた三頭分の位置（腰以上肩以下）となる。この範囲内のみで動作ブロックが検出された場合（胸部のみが自律的に動くことは無いと考えると）、腕のみが動作していると考えられる。従って、人物ブロックの最上値 HB_h はそのままその人物の身長であるから、動作ブロックの最上値 MB_h と最下値 MB_l が HB_h から導かれる三頭分の範囲

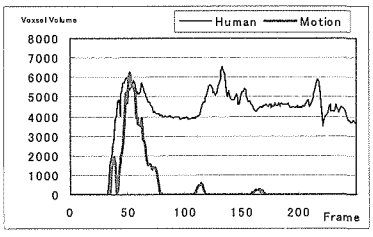
$$HB_h/2 \le MB_h \le HB_h \times 7/8 \tag{14}$$

$$HB_h/2 \le MB_l \le HB_h \times 7/8 \tag{15}$$

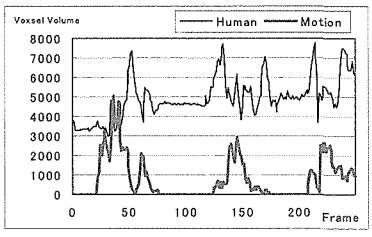
に現れた時、腕の部分が動いたと考えることができる。その際、瞬間的に動作ブロックが現れた場合は、移動量が少ない時に現れるノイズとし、連続して現れた場合のみ腕の動作とすることで、手サイン提示イベントが検出できる。なお、手サインの提示開始および終了は、先のように動作ブロックの重心の高さ $MB_g(x, y, z)|_z$ が順に上がっていく（もしくは下がっていく）ことで判断する。図 32(a)に手サイン提示イベント時の動作ブロックを、図 32(b)に対応する 2 番カメラの動作領域を示す。空中に浮かぶ腕の部分の動作ブロックが検出できた。更に、図 33 に frame115 における全カメラ画像での動作領域の同定結果を白線矩形で示す。このように、システムの対応するイベントに対し、実験データを基にした人物ブロックと動作ブロックの判定パラメータの決定が、イベント検出に重要となる。これら動作停止、開始、手サイン提示の各動作イベント検出時に、個人識別、顔向き推定、手サイン認識の各認識処理を行う。個人識別や顔向き推定には、顔領域が必要なため、人物ブロックにおける頭部に相当するブロック（八頭分の上から二頭分程度）の同定領域を、手サイン認識には、動作ブロックにおける同定領域を、認識対象領域とする。その際、実験で用いた統合画像では画像解像度が低いため、詳細な認識には不十分であると考えられ、解像度の高い画像を取得しているカメラ PC に領域座標を伝えることで、より適切な認識処理を行うことができる。



(a) Person – A



(b) Person – B



(c) Person - C

図 28. 検出ブロックの体積

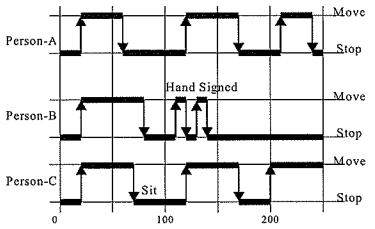


図 29. 動作イベント状態推移

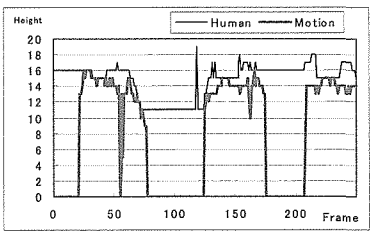


図 30.最上値（人物 C）

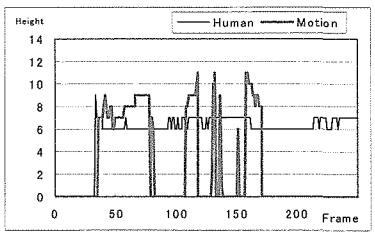
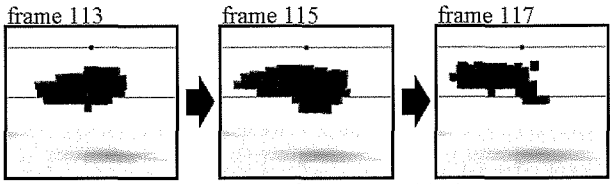
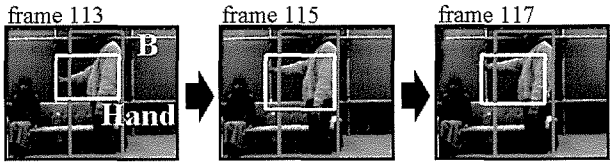


図 31.重心の高さ（人物 B）



(a) Movement of the motion block



(b) Motion region (Camera-2)

図 32. 動作ブロックの推移（人物 B）

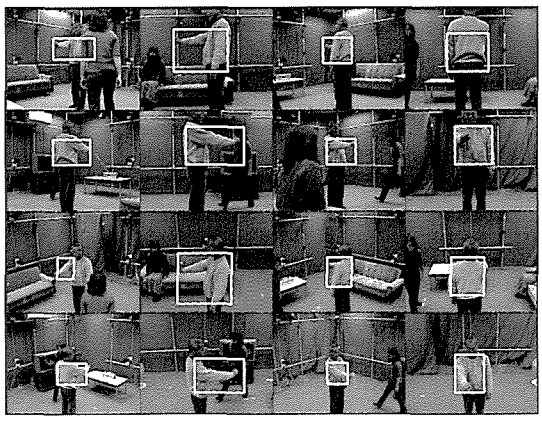
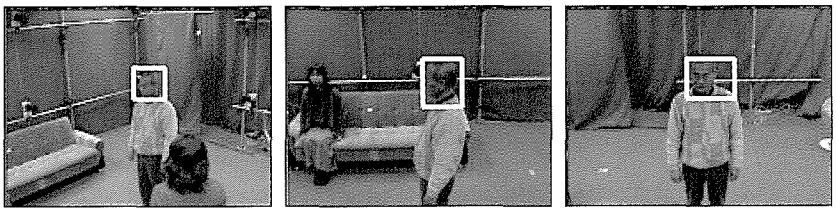


図 33. 動作領域の同定結果（人物 B）

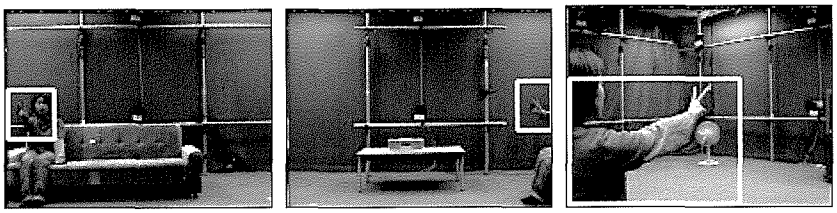
図 34 (a)～(c)に停止時の 9 番、10 番、16 番カメラの顔領域を白線矩形で示す。同一画面に複数人が映る場合でも、複数のカメラ画像で同一人物の顔認識を行うことができ、各認識結果から最終的な推定精度を向上することができる。

図 35 (a)～(c)にソファーに座り「チョキ」の手サインを提示したイベント時の 2 番、4 番、7 番カメラの手領域を白線矩形で示す。正面に当たる 2 番カメラでは、カメラから距離があるため、カメラ PC の画像でも認識は若干困難となる可能性がある。しかし、各カメラの手領域は同一人物と同定されているため、顔の映らない、4 番カメラによる横方向からの画像や、7 番カメラによる後方からの画像の認識結果を用いて検証することで、推定精度を向上することが可能である。



(a) Camera-9　(b) Camera-10　(c) Camera-16

図 34. 同一人物の顔領域



(a) Camera-2　(b) Camera-4　(c) Camera-7

図 35. 同一人物の手領域

本実験では、Pentium4, 1.7GHz の PC 1 台を使用した。 1 フレーム当り平均約 1.5 秒の処理時間を要した。これは一人あたり約 0.5 秒を要したためである。

プラズマディスプレイ、扇風機、エアコンなどの家電を配置したパーセプトルームにおいて、検出した手サイン提示タイミングの手領域を、ソケット通信によりカメラ PC に送信し、手サイン識別を行うことで、家電製品を制御した。図 36 に手サインにより TV を起動した様子を示す。複数人になると時間的な問題で手サインの検知が困難になるが、一人の場合は良好な家電制御が確認できた。今後、複数人における動作イベントのタイミング検知を実時間で処理し、各家電に対して適切な処理を行うことが課題である。

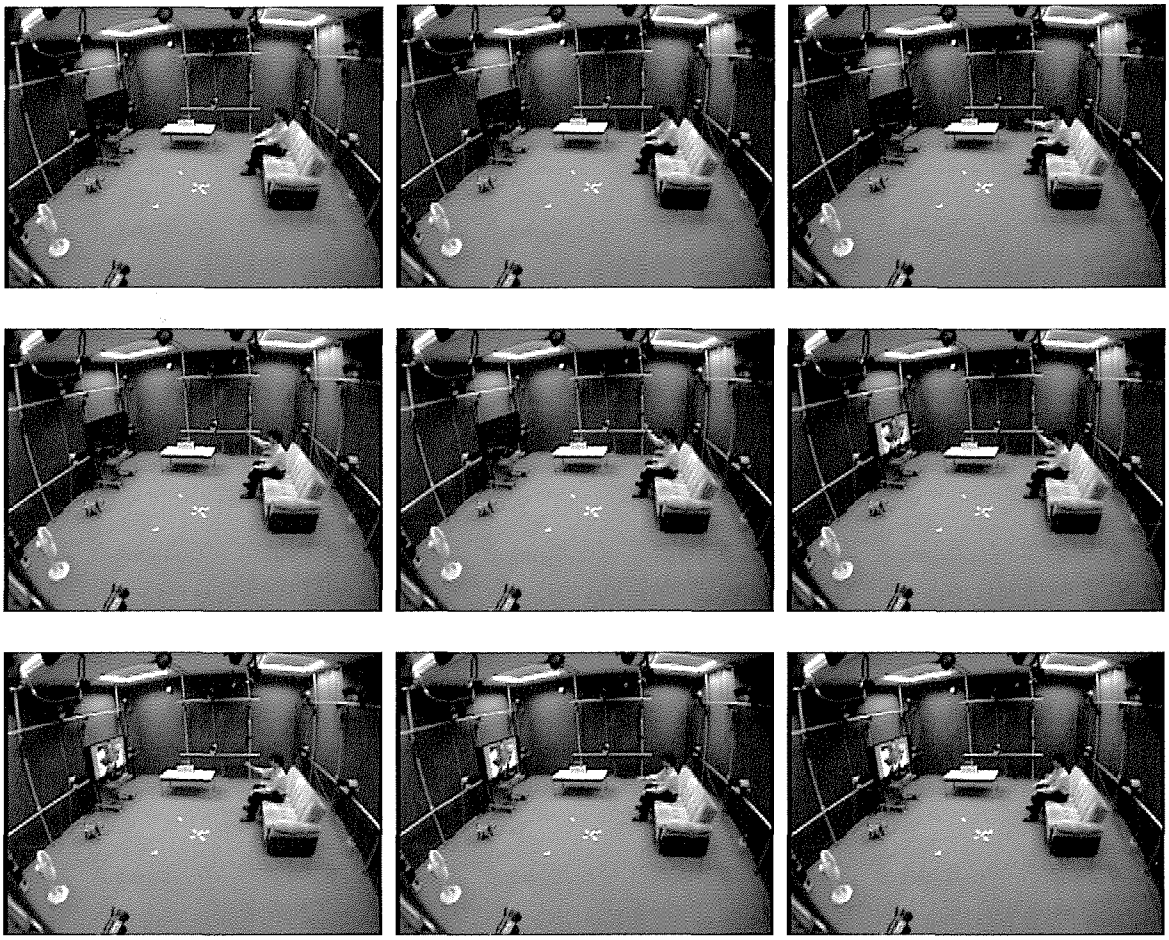


図 36. パーセプトルームにおける家電制御（TV の起動）

フェーズ III

今後の取り組み

これまでの研究において、人物検出技術および人物検出技術とパーセプトルームの有効性が示された。今後の取り組みとして、アルゴリズムの効率化や、適切なボクセル空間の分解能の設定、サンプリング間隔による存在確率マップの適用範囲の適切な設定を行うことで、高速化が期待でき、実利用は可能であると考えられる。また、本提案手法で常に安定した結果を得るためには、背景差分の雑音を軽減することが重要である。実験的な空間と異なり、一般的な環境を想定した場合、窓の外の情景変化や、家具の位置の変化などにより雑音ボクセルが増加し、人物の正確な位置推定が困難になる。そのため、実用化にあたっては、適切な背景更新法や背景差分法を導入する必要がある。更なる実現に向けて、照明や背景の変動に対する頑健性の向上や、検出精度の向上といった技術面ばかりでなく、家電制御に適した自然なジェスチャとは何か？といった生活姿勢面での検討も重要であると考えられる。また、カメラの台数や設置箇所にも考慮し、例えば壁面から一方向へのカメラのみを用いた場合などの考察も必要と考えられる。