ウェアラブル・コンピュータ向け リアルタイム Personal Positioning System

青 木 恒†

ユーザの状況認識はウェアラブル・コンピュータにおいては重要な機能である．とくにユーザが環境の中でどこにいるのかという場所認識は必要不可欠である．本論文では，カメラを備えたウェアラブル・コンピュータをユーザが身につけている状態を想定し，そのカメラが撮影した画像を解析することにより，リアルタイムで場所を認識させる試みを行った．ある場所を一度訪れた際に画像の特徴量列から辞書データを作成し，再度その場所を訪れた際には，同様に作成された特徴量列との比較を行うことで辞書内の場所を通過しているかどうかを判別する．この判別には音声認識と同様の dynamic programming アルゴリズムを用いた．実験装置は Libretto のような携帯できるパソコンと帽子に装着するカメラという構成で動作できる設計になっており，場所認識率が最高で 85.7%の性能値を得，ウェアラブル・コンピュータ上での本システムの実現可能性を示した．

Real-time Personal Positioning System for Wearable Computers

Hisashi Aoki†

Context awareness is an important functionality for wearable computers. In particular, the computer should know where the person is in the environment. This paper proposes an image sequence matching technique for the recognition of locations and previously visited places. As in single word recognition in speech recognition, a dynamic programming algorithm is proposed for the calculation of the similarity of different locations. The system runs on a stand alone wearable computer such as Libretto 1100*ff*. Using a training sequence, a dictionary of location is created automatically. These locations are then be recognized by the system in realtime using hat-mounted camera.

1. はじめに

ユーザ位置の認識はウェアラブル・コンピュータにとって重要な機能の 1 つであり，2 つのアプリケーション例で有効である．1 つは自動日記システム，もう 1 つはユーザの状況に適応した情報入出力である．自動日記システムとは，ユーザがウェアラブル・コンピュータと小型カメラを身につけており，その小型カメラがとらえた一日の出来事を記録していくものである．ウェアラブル・コンピュータはカメラからの映像の要約を自動作成し「日記」を適宜作成していく[1),2)]．たとえば「午前 8 時 15 分：自宅を出る．午前 8 時 17 分：路上で鈴木氏と会う．午前 9 時 02 分：職場に到着…」といったものである．このような機能の実現のためには，コンピュータはユーザの現在地を知る必要がある．ユーザの現在地が検知できれば，カー・ナビゲーション同様に未知の場所に向かうユーザに進路を提供することもできる（パーソナル・ナビゲーション）．さらに，ユーザ位置に応じて場所に関連した情報を提示することができる．たとえば，あらかじめ音声メモで「牛乳がきれている」という情報をウェアラブル・コンピュータに記録しておけば，ユーザがスーパーマーケットにさしかかったとき，牛乳を購入するように促すことができる[3)~6)]．

ユーザの状況に適応した情報入出力については Starner らが 1998 年に以下のような例をあげている．

> **「たとえば，もしユーザが上司の部屋にいるのならば，それは要談であると推定できる．そんなときには緊急でない限り電話や電子メールの着信をユーザに知らせるべきではないだろう」**[7)]

このような機能を実現するためには，コンピュータはユーザの位置情報取得ができなくてはならない．屋外であれば GPS（Global Positioning System）を利用するのも一案である．屋内においては，無線タグ技

† 株式会社東芝研究開発センター
Corporate Research and Development Center, Toshiba Corporation

術がこれを実現する最も有力な手法である[8)~10)]．しかしながら，ユーザの行動範囲すべてに無線タグを配置するのは現実的には困難であり，一方で GPS は屋内ではほとんど機能しないという事実がある．

この問題を解決する 1 つのアプローチは，ユーザがカメラを身につけ，そのカメラからの画像を手がかりに場所を推定するというものである．無線タグの代わりにバーコードを取りつけておき，カメラがそれを認識するという方法も可能だ[10)]．仮にバーコードを用いなかったとしても，オブジェクト（物体）認識，あるいは看板などの文字認識を利用して場所推定の手がかりにするという手法もある[11)]．しかし，すでに提案されている多くの画像処理的（CV: Computer Vision）アプローチは，1 枚の静止画像である「フレーム」の認識に立脚するものである．そのため，認識処理を行うべきフレームの選択方法や，1 枚のフレームの画質などが大きな問題になっていた．既出の Starner らは，画像を HMM（Hidden Markov Model 隠れマルコフモデル）に基づいた推移確率を利用して処理することにより部屋の移動を検出する方法を提案している[7)]．Clarkson らは，音声と映像の組合せによってユーザの行動を場所ごとに分割し，すでに訪れたことのある場所に再来した場合にはその一致を見出すことができる方法を提案している[12)]．

本論文では，リアルタイムでユーザ位置を検出する Personal Positioning System（PPS）システムの提案とその手法の検証実験の結果を報告する．このシステムでは図 1 に示したような身につけたカメラからの映像を入力し，映像も身につけられたスタンドアロンの（ネットワークから隔絶された）PC 上で処理することを想定している．従来の方法のうち，無線タグを用いた方法では室内環境側にとりつけたタグのバッテリ交換や情報更新といった定期・不定期のメンテナンス作業が生じ，バーコードを用いた方法では人間の視界にバーコードが入るために，美観を損ねる場合があるという問題があった．また，双方ともきめ細かいサービスを実現するためには室内のあちこちにタグやバーコードを貼りつけなければいけないという点が不可避である．現実の生活では建物の中でも一個人が管理できる空間は限られており，他人の管理領域にタグやバーコードを貼りつけるためには折衝が必要となる場合が多い．本論文で取り組んだ方法においては，人間が立ち入ることができる場所であれば，環境側に何 1 つ取りつけなくても認識を実現することができる．ネットワークを利用しないため，無線通信の混信や情報の漏洩に関する問題も生じない．

同様のアプローチでは，音声認識で用いられる dynamic programming 手法[14)]を用いた場所認識手法が筆者らによってすでに提案されている[13)]．この提案手法では，ユーザの位置が認識されるだけではなく，その場所に接近している軌跡を弁別することができる．しかし文献 13) での実験では，100 秒程度の短い映像で処理しており，処理も SGI ワークステーションを利用してオフラインに行ったものだった．

本論文では，単体のウェアラブル・コンピュータ上で実現するよう文献 13) の手法のアルゴリズムを大幅に改善した．本システムではスタンドアロン PC でリアルタイムに毎秒約 7 フレームを処理することができる．その PC も，たとえば東芝 Libretto 1100*ff* といった実際に身につけて行動することが可能なプラットフォームで実現することができる（本論文での実験は MMX テクノロジ Pentium 166 MHz プロセッサを搭載したノート PC で実現しており，同 266 MHz を搭載した Libretto 1100*ff* は処理速度の面でそれを上回る性能を持つ）．また提案手法では，処理を行う PC は外部のデータベースとネットワーク経由でデータの照合を行う必要がないような設計になっているために，無線通信のバンド幅の小ささに起因する処理性能の低下や，通信環境の不安定さに起因する動作不安定を避けることができる．加えて，提案手法ではユーザは認識すべきタイミングをコンピュータに指示したり，ある部屋への入退場を手動で入力したりする必要がない．さらに本論文では，認識可能な場所のリストである「場所辞書」を自動生成する方法を提案している．最後に，本論文では場所辞書作成のためのトレーニング・データとして 16 分 30 秒，実証実験のためのテスト・データとして 7 分 30 秒に及ぶ長時間のデータを用い，手法の有効性を検証した．手法の有効性を示すために，既出の論文 7) で提案された精度指標を修正した新しい精度指標も導入している．

以下では，2 章でシステムの概要を説明した後に，3 章では映像の特徴量となる色相ヒストグラムの計算方法，および特徴量どうしを比較して類似であるかどうかを検証する手法について説明する．4 章では，場所辞書を自動生成する方法について説明する．最後に 5 章において，実証実験の結果と考察を行う．

2. システムの概要

システムは，ノート PC，PCMCIA ビデオキャプチャーカード，および（たとえば**図 1** のような）装着可能型カメラから構成される．これらはスタンドアロン環境で完全にリアルタイムで動作し，遠隔におかれ

図 1 装着型カメラ
Fig. 1 A small wearable camera.

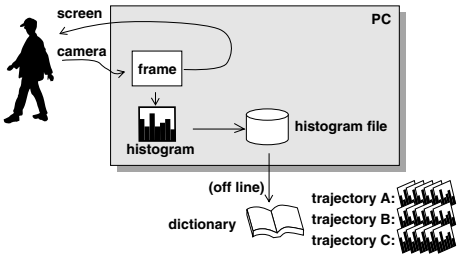


図 2 学習フェーズでのデータの流れ
Fig. 2 Data flow at recognition mode.

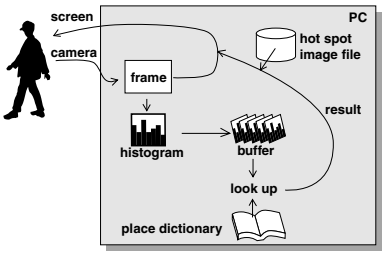


図 3 認識フェーズでのデータの流れ
Fig. 3 Data flow at analysis mode.

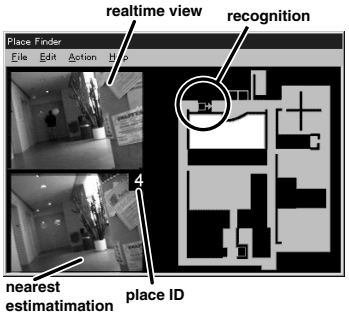


図 4 本システムの画面イメージ
Fig. 4 Screen shot of the proposed personal positioning system.

たワークステーションなど別のコンピュータからの支援をいっさい必要としない．ユーザは「学習フェーズ」「認識フェーズ」の 2 フェーズに分けてシステムを操作する.「学習」と「認識」の間に，システムが自動的に場所辞書をオフラインで作成するフェーズがある．まず学習フェーズではシステムを身につけたユーザは自分が実際に行動する，あるいは「しそうな」範囲を無作為に歩く．この間，PC 上では受信フレームの色相ヒストグラムを計算し，ハードディスク上に記録していく（**図 2**）．学習フェーズが終わると，システムは場所辞書に適した色相ヒストグラムの時間区間を探索する．適している度合いが上位であるものから規定の個数が選ばれ，場所辞書に掲載される「場所単語」となる．なお，この詳細は 4 章で説明する．

引き続き，ユーザは認識フェーズに切り替えて自由に歩く．システム内には過去数〜数十フレーム分の色相ヒストグラムが一時蓄積（buffering）されており，それらと場所辞書内の場所単語の類似度が総当たりで計算される．類似度が規定値よりも高い場所単語が見つかった場合，システムはユーザに対して特定の場所を通過しつつあることを認識結果として提示する（**図 3**）．

実験はノート PC，東芝 Tecra 740CT を用いて行い，画像データの収集はソニー製のハンディカムを用いた．実際に装着型の PC やカメラを利用しなかった理由は，アルゴリズム確立のために同じ画像を繰り返し実験する意図からであり，システムの構成上はノート PC が Libretto 1100*ff* に置き換わり，カメラが図 1 のようなカメラに置き換わっても機能するようになっている．**図 4** に，本システムの画面例を示す．

3. ヒストグラム類似度の計算

CV（Computer Vision）システムをウェアラブル・コンピュータに搭載する場合，次の 3 点が要件となる．

- カメラからの入力画像の明るさの変化やノイズ，傾きなどにロバストでなければならない．
- 十分なフレーム・レートを保証するために，各フレームに係る計算は高速でなければならない．
- 各フレームから抽出される特徴量は，できるだけ少ない情報量でなければならない．

第 1 点目に関しては，もしシステムが明るさの変化に影響を受けてしまうと，同じ場所に朝に訪れた際に場所辞書が作成され，認識が夜に行われた場合などに正しい認識ができなくなる恐れがある．また，当然ながらカメラ画像に割り込んでくるノイズは学習時と認識時で異なるものであるし，学習時と認識時でカメラの傾きが同じである保証もない．第 2 点目に関しては，計算が遅いためにフレームレートが低くなると深刻な性能低下が発生する恐れがある．たとえば毎秒 5

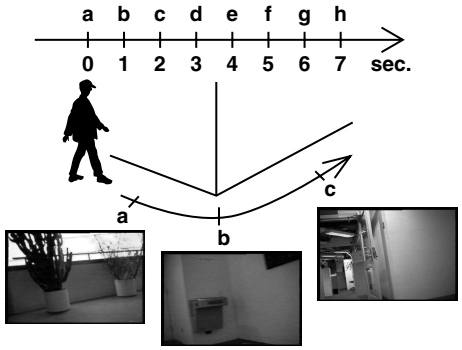


図 5 サンプリングレートが低いために生じる問題
Fig. 5 Sparse sampling causes bad performances.

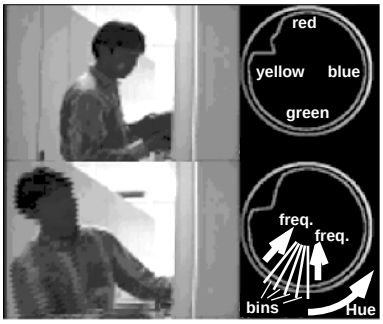


図 6 色相ヒストグラム
Fig. 6 Chromatic (hue) histogram.

フレーム程度でシステムが処理できる場合，同じ場所を学習時と認識時で通過したとしても，最大のタイミングのずれは 100 ミリ秒であり，100 ミリ秒の間にカメラがとらえる画像の変化は大きくない．しかし，もし 1 フレームに係る計算が膨大であるために 1 フレームあたり 2 秒の時間を要した場合，学習時と認識時のタイミングのずれは最大 1 秒に及ぶ．**図 5** に示したように，学習時には (a)(c)(e)…が記録され，認識時にシステムに入力される画像が (b)(d)(f)…となった場合，(a) と (b) の画像が違いすぎてもはや同じ場所として認識することが困難になる．

第 3 点目に関しては，フレームごとの特徴量が莫大になれば，それだけ計算量が増してしまうことは明らかであるが，それ以外に格納場所の問題が生じる．640 × 480 ピクセルのフルカラー画像は 900 キロバイトの大きさを持っているが，学習フェーズでこれをすべてハードディスクに記録していくと 114 フレームで 100 メガバイトに達する．これが本論文で後述するように各フレーム 36 バイトのサイズにおさめられれば，290 万を超えるフレームを同じ領域に保存することができる．

3.1 色相ヒストグラム

本システムでは，フレームの特徴量として色相ヒストグラムを用いた．画素は通常，赤，緑，青からなる RGB 値で表されるが，Hue（H），Saturation（S），Brightness（B）の 3 値でも表すことができる．フレーム全体の画素について，H 値から 36 の分解能を持つヒストグラムを計算し，36 次元の特徴ベクトル $\boldsymbol{f}$ を作成する．

図 6 には色相ヒストグラムの例を示してある．図ではヒストグラムが環状に示されているが，これは，H 値の両端は同じ色を表しているからである．すなわち，もし H 値を 0 から 1 の数値で表すとすると，0 から数字が増えるにつれ，示している色は青，赤，黄色，緑と移り変わり，$H = 1$ においては再び青に戻る．図中では，内側の多角形が中心に近い頂点を持つ成分ほど，ヒストグラム上では多くの画素をカウントしたことを意味している．図 6 に示しているように，色相ヒストグラムは画像中の物体の動きにロバストであり，さらに画面全体の明るさの変化に対してもロバストである[15]．

フレーム s からフレーム e までに関する色相ヒストグラム $\boldsymbol{f}$ のグループを，下記のように $\boldsymbol{F}$ で表すことにする．

$$\boldsymbol{F}_{se} = (\boldsymbol{f}_s, \boldsymbol{f}_{s+1}, \cdots, \boldsymbol{f}_e)$$

本システムでは各ヒストグラムの（あるいは各ベクトルの）成分は 0 から 255 の数値を持つように変換される．したがって，1 フレームに係る特徴量は 36 バイトで記述されている．

3.2 複数のフレームの類似度測定

上述の $\boldsymbol{F}$ を用いて，2 つのヒストグラム・グループの類似度を比較する方法を以下で示す．2 つのグループとは，場所辞書内に登録されている場所単語の数〜数十フレーム分と，認識フェーズでカメラからリアルタイムに入力されて buffering されている数〜数十フレーム分のことである．まず，それぞれのグループのヒストグラムの距離（非類似度）を総当たりで計算した距離行列 $\boldsymbol{D}$ の成分を以下のように計算する．

$$(\boldsymbol{D}_{ij})_{kl} = d(\boldsymbol{f}_k, \boldsymbol{f}_l) = |\boldsymbol{f}_k - \boldsymbol{f}_l|^2$$
$$k = s_i, s_i + 1, \ldots e_i$$
$$l = s_j, s_j + 1, \ldots e_j$$

これはグループ i とグループ j の距離行列を示している．グループ i はフレーム s_i から e_i までのヒストグラム列であり，グループ j はフレーム s_j から e_j までのヒストグラム列である．i と j が同じグループでなくても，2 グループが同じ場所を同じ速度で通過したフレーム群を示している場合には，やはり対角成

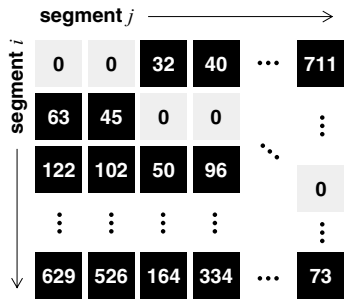


図 7 2つのグループの速度が異なるときの距離行列の例
Fig. 7 Difference in speed.

分は 0 か，あるいは低い数値になることが期待される．

2 グループが同じ場所を通過したフレーム群でありながら，グループ j はグループ i よりも遅い速度であった場合には，**図 7** に示したように上側に谷が寄った形の行列になると想定できる．

したがって，グループ i と j の各フレームの対応関係は，$(1,1)$ 成分から $(e_i - s_i + 1, e_j - s_j + 1)$ 成分に向かう「谷」（周囲に比べて数値の低い成分）を探索することで求められる．これは上で示したように必ずしも対角成分ではない．谷の探索には下記の制限を定めた．

- (m,n) 成分からは $(m+1,n)$，$(m,n+1)$，$(m+1,n+1)$ 成分のいずれかにしか進めない．すなわち，グループ i と j の間に速度の差は認めても，逆行は認めない．
- 最終的に求められた「谷」経路に沿っての成分の和が最小になる．

このようにして求められた対応関係を「最小経路」と呼ぶことにし，最小経路に沿っての行列の成分和を T で表す．たとえば，$\boldsymbol{D}$ が以下のような行列であった場合，

$$\boldsymbol{D}_{ij} = \begin{pmatrix} 0 & 3 & 2 & 9 \\ 3 & 4 & 5 & 2 \\ 7 & 9 & 2 & 0 \end{pmatrix}$$

最小経路は $(1,1)$–$(2,2)$–$(3,3)$–$(3,4)$ となり，T は $0+4+2+0=6$ となる．$(1,1)$ の次の成分として $(2,2)$ が選ばれていることに注目して欲しい．$(2,2)$ 成分は $(1,2)$ 成分や $(2,1)$ 成分より大きいが，$(1,2)$，$(2,1)$ 成分のいずれを経由した経路も最終的には $(2,2)$ 成分を経由した場合よりも T が大きくなる．このような dynamic programming 手法は音声認識で多く用いられる[16]．すなわち，話者や状況によって発話される速度が異なるのを補正する効果を持っており，本論文の場合では同一の場所を通過する速度の差を吸収することができる．

本システムでは，場所辞書内の場所単語として 40 フレームぶんのヒストグラムを保持し，認識フェーズでの buffering も同様に 40 フレームを用いて行っている．したがって，場所辞書が 16 の場所単語を保持するとき，システムは $40 \times 16 = 640$ のヒストグラム（各 36 バイト）を保持していることになる．40 フレームという数字は，人間の室内歩行の典型的な変化ポイントであるドアの通過というイベントを基準に設定した．実験プラットフォーム上での処理フレームレートにおいてドアに達する数歩前から，ドアを離れて数歩後までの画像をフォローできる経験的時間として設定したものである．場所辞書内の場所単語 P（場所 P に関するヒストグラム記述）を $\boldsymbol{F}_{dict(P)}$ で表し，認識フェーズの時刻 t の時点で buffering されていたヒストグラム・グループを $\boldsymbol{F}_{test(t)}$ と表すと，時刻 t の $\boldsymbol{D}$ と時刻 $t-1$ の $\boldsymbol{D}$ との間には以下の関係がある．

$$(\boldsymbol{D}_{dict(P)test(t)})_{kl} = (\boldsymbol{D}_{dict(P)test(t-1)})_{k,l+1} \quad (l = 1, \ldots, e_{test} - 1)$$

すなわち，時刻 t の $\boldsymbol{D}$ の成分のうち，第 1 列から第 $e_{test(t)} - 1$ 列までは，時刻 $t-1$ の $\boldsymbol{D}$ における第 2 列から第 $e_{test(t-1)}$ 列と同じである．したがって，新しいフレームが 1 つバッファに入るたびに計算しなければいけないのは時刻 t の $\boldsymbol{D}$ における第 $e_{test(t)}$ 列だけである．P のような場所単語がたとえば 16 カ所ぶん場所辞書に登録されている場合には，1 フレーム入力のたびに計算しなければいけないのは $40 \times 16 = 640$ の d の計算と，16 単語に対する T の計算だけで済む．

4. 場所辞書の作成

学習フェーズで記録されたヒストグラムのすべてが場所辞書の場所単語としてふさわしいわけではない．たとえば，単調な長い廊下の西側を 50 メートル歩いた部分のヒストグラム・グループは，同じ廊下の東側を 50 メートル歩いた部分のヒストグラム・グループと酷似していることが予想される．したがって，前者を場所辞書に登録しても，その廊下のなかのどこを歩いているかを弁別する性能は持たない．したがって，学習フェーズに記録されたヒストグラムから生成されるヒストグラム・グループのうち，ほかのグループとの違いが明確であるヒストグラム・グループを場所単語として採用すべきであるといえる．

このため筆者は，弁別性能を評価するために以下のような指標 M を導入した．下記でグループ i および j は双方ともに同じ学習フェーズで記録されたヒスト

グラムから生成されるヒストグラム列である．

$$M_i = \sum_{all\ j} T_{ij}/(number\ of\ segments)$$

number of segments は，学習フェーズで記録されたヒストグラムから生成されるヒストグラム列の数であり，Σ による加算の回数と同じである．たとえば，学習フェーズで 1,000 のヒストグラムが記録され，それを連続する 40 のヒストグラムごとにグループにしたとすると，i および j は 1 から 25 の値をとりえ，このとき *number of segments* は 25 となる．辞書作成に際しては，学習フェーズでのユーザの行動に制限を加えていないことから，ユーザは 1 回の学習フェーズ内で同じ場所を 2 回以上通過している可能性はあるが，先験的な知識なしにシステムが同じ場所への再来を認識することはできない（つまり，仮に学習フェーズのヒストグラムのなかに類似の部分が 2 つあったとしても，それは本章の先頭で述べたように「違う場所が同様のヒストグラム列をなす」場合であるのか，それとも「実際に同じ場所を 2 回通過している」場合であるのかを知る知識がシステムにはない）．厳密には上記の M は「同じ場所の通過」は除外して計算すべきであるが,「同じ場所の通過」を示すグループ数は全体のグループ数に比べて十分小さいので，この際無視することができるものと考える．

M が計算されると，学習フェーズの中で他に対して際だっているヒストグラム・グループを M の順で列挙することが可能になる．しかし，この段階ではたとえばフレーム 301 から 340 までのグループが第 1 位に選ばれたときに，フレーム 302 から 341 が第 2 位に選ばれるなど，オーバラップしているグループがまとまって列挙されることが考えられる．したがって，このようにオーバラップしているグループの中からは 1 つだけを選び出すようにする．

場所辞書として選び出された場所単語に関しては，その場所を通過中の画像フレーム 1 枚がヒストグラム列とともに辞書に記録される．この画像フレームは図 4 にあるように，システムが最も有力と判断している場所候補として画面左下に表示される．

一方，あらかじめ画像として入手したフロアプラン上での場所単語の座標と方向を手作業で登録する．これは図 4 のウインドウの右にある地図上の矢印として，場所認識が行われたときに表示される．

5. 実験と考察

学習フェーズおよび認識フェーズの映像は MIT 内の研究所建物の 1 フロアで撮影された．学習フェーズの映像は 16 分 30 秒，認識フェーズの映像は 7 分 30 秒である．いずれも研究所建物の 3 階を中心として無作為に歩いたものであるが，いずれのケースも 3 階から出て 1 階ホールや地階ホールを歩いた区間を含んでいる．学習フェーズと認識フェーズは別の日に撮影され，学習フェーズの撮影は夕方に，認識フェーズの撮影は日中に行われた．場所辞書に載せる場所単語は 3 階部分だけから選んだ．**図 8** には，研究所建物 3 階の見取り図と，システムが選択した 16 の場所単語が示してある．4 章で述べた弁別性能指標の順は，13，9，11，4，14，5，8，6，10，16，2，1，7，15，12，3 の順である．単語 8 と 16 は同じ場所の同じ方向の通過を示しているが，学習フェーズにおいてこの場所は 2 回通過しており，そのいずれもが場所辞書に選ばれたことを示している．実験は，場所辞書に登録する場所単語の数を変化させて行った．

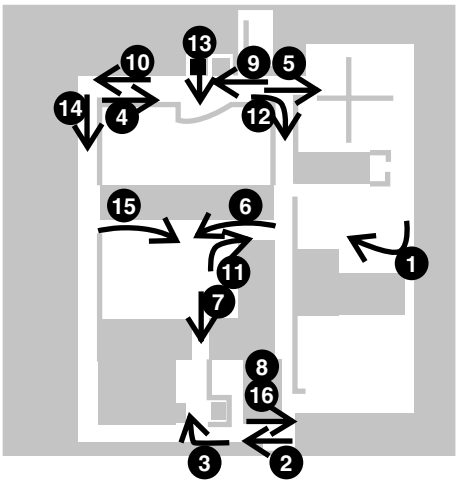


図 8 研究所建物 3 階の見取図と,「場所辞書」に登録された「場所単語」
Fig. 8 The floor plan and the chosen trajectories.

図 9 は認識フェーズでユーザが歩いた経路を示している．実線は認識フェーズの最初の 3 分間の挙動であり，ユーザはこのあと階段で下り，1 階および地階のホールを 2 分間歩いた．次にエレベータを使って 3 階に戻ってきたあとの挙動が破線で描かれている．破線区間は 2 分 30 秒である．**図 10** には認識結果を示した．普通に書かれている数字は認識されるべくして認識された「正認識」である．黒地に白で書かれている数字は，誤った場所と認識された「誤認識（substitution）」である．四角で囲われた数字は，認識すべきでない部分で認識された「過認識（insertion）」である．黒四角は場所辞書内に認識すべき場所単語があるにもかかわらず，何も認識されなかった「未認識（deletion）」

である．まったくの空欄になっている部分は，認識すべきでない部分が，実際に何も認識されなかった区間「正沈黙」であり，これも正解のうちである．したがって，正解とすべきものは「正認識」と「正沈黙」であり,「誤認識」「過認識」「未認識」はエラーといえる．たとえば，場所 6 はどの場合でも正認識されていることが分かる．一方，2 つ目の場所として訪れている場所 7 は，場所辞書内の場所単語数が 16 の場合のみ正認識され，単語数が 15 から 13 の場合には未認識のエラーとなっている．この場所 7 は弁別性指標の順で第 13 位にあるために，辞書内の単語数が 12 以下の場合には辞書内に存在しない．したがって，単語数が 12 以下の場合には場所 7 が検出されないのはエラーではない．図 10 では「//」によって省略されているが，ユーザが 3 階から離れている区間に関しては，どの単語も過認識されることはなかった．

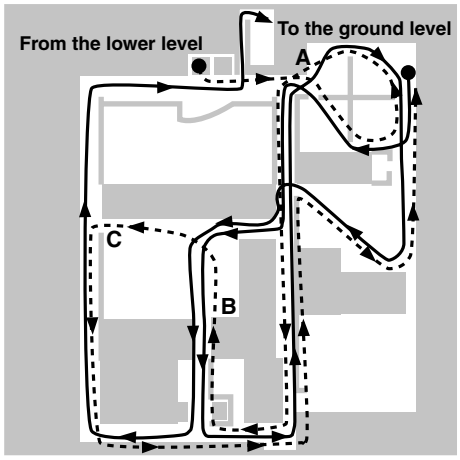


図 9 認識フェーズでのユーザの挙動
Fig. 9 The test walk.

表 1，**表 2** には数値による実験結果を示した．ここで用いられている指標は文献 7) で Starner らが用いたものを修正したものである．T および G は，それぞれ認識すべき場所，および沈黙すべき区間の数である．G にはユーザが 3 階を離れた区間（場所 4 と 13 の間）は 1 回として加えられており，区間が連続している場所 2 と 3 の間は加えられていない．D は未認識の数，S は誤認識の数，I は過認識の数である．文献 7) の実験では移動のどの瞬間においても，いずれかの部屋にいるというラベル付けがされているため,「どの場所にも属さない」という中間的状態がない．これに対し，本論文のケースでは移動行程のすべてに場所の正解があるわけではなく,「どの場所にも属さない」区間ではシステムはどの場所も検出しないこと，すなわち沈黙を保つことが要求される．したがって，文献 7) の評価基準からはこの点に変更を加え,「場所を検出すべき」認識率から「沈黙を保つべき」認識率を分離した．これらから計算される場所認識に関する精度 Acc_T および沈黙に関する精度 Acc_G は以下のように定義される．

$$Acc_T = \frac{T - D - S}{T}$$

$$Acc_G = \frac{G - I}{G}$$

文献 7) で述べられているように，過検出が多い場合には Acc_G は負の数もとりうる．また D と I は相関関係にあり，I を減らそうとすれば D が増える．

表 1，2 によると，辞書内に場所単語が 9 あるいは 10 登録されている場合に本システムの性能が最高になっている．なお，評価尺度に変更を加えているため

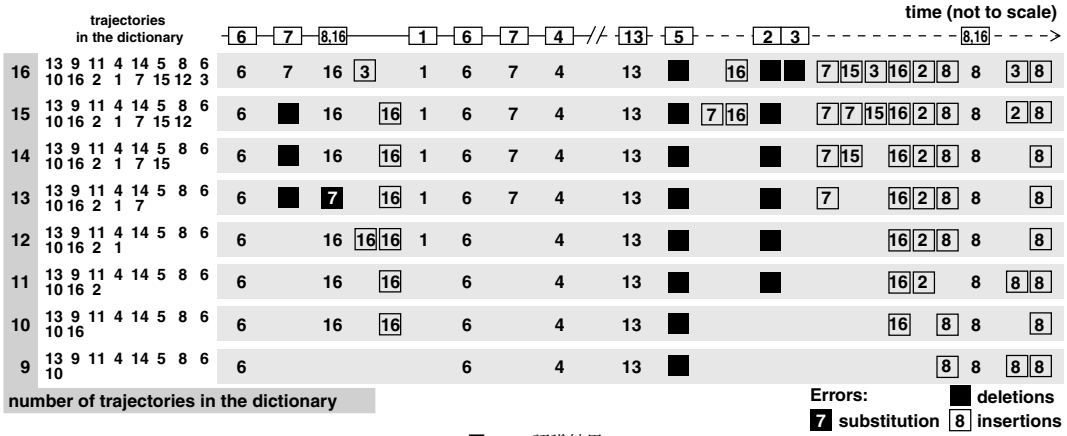


図 10 認識結果
Fig. 10 The results.

表 1 システムの性能（1）（#は場所辞書中の場所単語の数）
Table 1 System performance (1). (#: number of trajectories in the dictionary)

#	T	$T-D-S$	D	S	Acc_T%
16	12	9	3	0	75.0
15	11	8	3	0	72.7
14	11	8	3	0	72.7
13	11	7	3	1	63.6
12	9	7	2	0	77.8
11	8	6	2	0	75.0
10	7	6	1	0	85.7
9	6	5	1	0	83.3

表 2 システムの性能（2）（#は場所辞書中の場所単語の数）
Table 2 System performance (2). (#: number of trajectories in the dictionary)

#	G	$G-I$	I	Acc_G%
16	12	2	10	16.7
15	12	1	11	8.3
14	12	5	7	41.7
13	12	6	6	50.0
12	10	4	6	40.0
11	9	4	5	44.4
10	8	4	4	50.0
9	7	4	3	57.1

に単純比較することはできないが，参考として示すなら本論文の Acc_T 最高性能は 85.7%であり，文献 7) においては最高時の認識率 Acc は 81.82% である．

図 10 に示したように，場所 5 は一度も正しく検出されなかった．これは，学習フェーズで場所 5 を通過した際にはこの入口ドア A は開放で固定されていたが，認識フェーズでは閉まっており，ユーザが自ら手でドアを開けるという動作が含まれてしまったからと考えられる．また，ユーザが認識フェーズで図 9 の B 地点から C 地点に向かう間に大量の過検出が生じているが，これはこの区間を通過するのに 4〜5 秒を要し，場所辞書中の場所 6，7 および 15 と類似であるからである．場所 2，3，8，16 に関する過検出も同様の理由と考えられる．これらが場所辞書として選ばれてしまった理由は，これらどうしは類似であるにもかかわらず，これらと，それ以外との差が大きかったために M が大きな数字をとったことが原因と考えられる．

一方，このように自動生成された場所辞書は必ずしもユーザが認識させたいと望む場所ではないという可能性はある．これに関してはたとえば加速度センサを用いて歩数を計数し，現在地を補間推定することなどの方法で解決が期待できる[17]．

今回の実験では，辞書の作成と認識のための歩行が同一人物によって行われたが，3 章に述べたように，場所辞書のデータ量は小さいものであるため，それを複数のユーザで共用することができる．共用が実現すれば，一度も訪れたことのない場所についても初回から認識が可能になるなどのメリットが生じる．このような場所辞書をデータベース化しておけば，電車や自動車を利用した大きな移動に関しては GPS や PHS といった比較的解像度の粗い場所認識手段を用い，それらが使えなくなった状況やより高い解像度が求められる状況において本手法を用いるといった方法によって，きめが細かく，正確さも高いサービスが提供できる．本手法の特徴の 1 つであるプライバシー情報（自分の詳細な位置）がネットワークに流れないという面を活かすためには，建物の入り口に設置された情報キオスク端末で一括してその建物に関する場所辞書をダウンロードさせたり，FM 波などで場所辞書データを「館内放送」してもよい．このようにすれば，ユーザの PC からは現在地に関する詳細な情報をネットワーク上にアップロードすることなく，ユーザ自身だけが自分の位置を知ることができる．

6. おわりに

本論文では，ウェアラブル・コンピュータ向けに画像処理をもちいた場所認識手法を提案した．提案システムではスタンドアロンの PC を用い，外部のコンピュータと交信することなしにリアルタイムに認識を行うことができた．また，システムの性能を「場所を正しく認識する率」と「非検出となるべき区間で検出しない率」という 2 つの尺度を導入することで評価し，最高時には前者が 85.7%，後者が 50.0%という性能が得られた．今後の精度向上の可能性としては，あらかじめシステムに場所辞書の位置関係を持たせ「ある場所からある場所には（何秒以内には）行けない」などの制約を加えることが考えられる．また，場所辞書へ登録する場所単語の選択方法についても他のバリエーションを試す余地がある．さらに，今回 36 次元の色相ヒストグラムを用いた部分に関し，それ以外の画像特徴量に関しても実験を行う必要があると考えている．

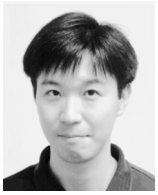
ウェアラブル・コンピュータがユーザの室内の現在地を知ることができれば，ユーザの行動に即した「気の利いた」情報提示が実現しよう．たとえば，会議室に入室すれば会議に必要な資料がパソコンに表示されたり，インターネット電話の着信音を鳴らさないように設定したりできるし，自動販売機の前に立つ回数が多ければ，健康のためにほどほどにするよう促したりできる．従来のコンピュータでは，サービスを受ける

ためにはユーザ自身が自分の要求を入力しなければいけないという問題が存在したが，場所認識により「ご主人様が何を欲しがっているのか」という気を回してくれる存在になる第一歩を踏み出すことができるだろう．

参 考 文 献

1) Lamming, M. and Flynn, M.: Forget-me-not: A human memory prosthesis–ubiquitous technology in support of everyday memory problems, *Proc. FRIEND21 International Symposium on Next Generation Human Interface* (1994).
2) Rhodes, B. and Starner, T.: Remembrance agent: a continuously running automated information retrieval system, *Proc. 1st International Conference on the Practical Application of Intelligent Agents and Multi Agent Technology* (*PAAM '96*) (1996).
3) Weiser, M.: The computer of the 21st century, *Scientific American*, Vol.265, No.3, pp.66–75 (1991).
4) Kakez, S., Conan, V. and Bisson, P.: Virtually documented environments, *Proc. 1st International Symposium on Wearable Computers* (*ISWC '97*) (1997).
5) Rekimoto, J. and Nagao, K.: The world through the computer: computer augmented interaction with real world environments, *ACM UIST'95* (1995).
6) Nagao, K. and Rekimoto, J.: Agent augmented reality: A software agent meets the real world, *Proc. 2nd International Conference on Multiagent Systems* (*ICMAS-96*) (1996).
7) Starner, T., Schiele, B. and Pentland, A.: Visual contextual awareness in wearable computing, *Proc. 2nd International Symposium on Wearable Computers* (*ISWC '98*) (1998).
8) Want, R. and Hopper, A.: Active badges and personal interactive computing objects, *IEEE Trans. Comsumer Electronics*, Vol.38, No.1, pp.10–20 (1992).
9) Orwant, J.: For want of a bit the user was lost: Cheap user modeling, *IBM Systems Journal*, Vol.35, No.3, pp.398–416 (1996).
10) Starner, T., Mann, S., Rhodes, B., Healey, J., Kirsh, D., Picard, R.W. and Pentland, A.: Augmented reality through wearable computing, *Presence*, Vol.6, No.4, pp.386–398 (1997).
11) Schiele, B. and Crowley, J.: Probabilistic object recognition using mutltidimensional receptive field histograms, *International Conference on Pattern Recognition* (*ICPR '96*), vol.B, pp.50–54 (1996).
12) Clarkson, B. and Pentland, A.: Unsupervised clustering of ambulatory audio and video, *Proc. International Conference on Acoustics, Speech and Signal Processing* (*ICASSP '99*) (1999).
13) Aoki, H., Schiele, B. and Pentland, A.: Recognizing personal location from video, *Proc. Perceptual User Interfaces Workshop* (*PUI '98*) (1998).
14) Ney, H.: The use of a one-stage dynamic programming algorithm for connected word recognition, *Readings in Speech Recognition*, pp.188–196 (1990).
15) Aoki, H., Shimotsuji, S. and Hori, O.: A shot classification method of selecting effective keyframes for video browsing, *Proc. ACM Multimedia 96*, pp.1–10 (1996).
16) Sakoe, H. and Chiba, S.: Dynamic programming algorithm for spoken word recognition, *Readings in Speech Recognition*, pp.159–165 (1990).
17) 杉山博史ほか：歩行状況センサを利用した歩行者用音声対話道案内システム，第 60 回情報処理学会全国大会論文集，3-227 (2000).

(平成 11 年 12 月 24 日受付)
(平成 12 年 7 月 5 日採録)



青木 恒（正会員）

1970 年生．1993 年東京大学工学部応用物理学科卒業．同年（株）東芝入社．同社にて画像認識，映像理解，ビデオ構造化，検索システムおよびヒューマンインタフェースの研究に従事．1998 年より 1999 年まで米国 MIT Media Laboratory 客員研究員として，画像認識技術のウェアラブル・コンピュータ応用を研究．現在，（株）東芝研究開発センター ヒューマンインターフェース ラボラトリー勤務．