ウェアラブルイメージングシステムによる実時間イベント推定

澤畠 康仁† 堀 鉄郎† 相澤 清晴†
† 東京大学大学院新領域創成科学研究科

1 はじめに

本研究は、日常生活における体験を、映像として記録・蓄積することを目的とする。小型カメラの普及、ストレージ技術の進歩、およびウェアラブルコンピュータの台頭により、人が一生の間に目にするすべての映像を記録することは、もはや不可能ではない。この際、膨大な長さの映像から、必要なシーンを見つけ出すための技術が不可欠となる。映像の効率的な取得を行うためには、映像の内容に関するメタ情報を記述することが有効である [1, 2, 3]。本稿では、映像・音声に加え、GPS、加速度センサやジャイロなどのセンサ情報を同時に記録し、それらを統合的に処理を行うことによる、実時間でのメタ情報抽出手法について検討する. 具体的には、ユーザがどういう動作をしているのかというイベント検出、および映像中の顔の検出を行う。これらにより、「どこで」「何をしているとき」「誰と」を組み合わせたものをクエリとした、映像の検索が可能となる。

2 体験映像蓄積の特長と課題

ウェアラブルなカメラを身につけ、日常生活において目にする映像をすべて記録するということについて考える。人間の一生を 70 年とすると、常時記録による映像のデータ量はどれほどになるであろうか。いくつかの動画圧縮方式をもちいてその量を概算した結果を表 1 に示す。なお、 1 日は 16 時間として計算している。

テレビ電話品質での蓄積に注目してみれば、70 年間の体験映像の記録は、わずか 11[TBytes] で足りることになる。今日におけるストレージ技術は、驚くべき速さで向上を続けている。現在の HDD を用いても、200GB の HDD が 50 台あれば 70 年間の映像を余すところなく記録することが可能である。将来の HDD 技術の進展を考慮すれば、70 年分の映像を一つの HDD に蓄えることも遠い未来の話ではないといえよう。

表 1: 体験映像の量

quality	**rate**	data size for 70 years
TV Phone quality	64 kbps	11 TBytes
VCR quality	1Mbps	183 TBytes
Broadcasting quality	4Mbps	736 TBytes

センシングデバイスである CCD カメラや CMOS カメラもまた、高機能・小型化が進められている。カメラが搭載されている携帯電話も珍しいものではなくなっており、写真や動画をいつでもどこでも取得できるという環境がそろいつつある。

ウェアラブルコンピュータを謳った商品も市場に出始め、これまでデスクトップでのみと限定されていたコンピューティング環境は、その制限が解かれつつあり、人とコンピュータとの関係に大きな可能性を生み出すものとして大きな注目を集めている。

ハードウェア的観点から見た場合、これらの技術がうまく融合することにより、人間の一生分にも及ぶ長い映像を記録するということは、近い将来十分可能なことであるといえる。

体験映像の記録のおける利点と欠点を以下のようにまとめる。

利点
- 残したい瞬間を逃さず記録ができる
- 過去の体験をリアルに追体験・追想することができる
- すでに忘れてしまったことを思い出すことができる
- 見逃したシーンを見ることができる
- 自分がしたこと、しなかったことを証明できる

欠点
- 忘れたいことも残してしまう
- 他人のプライバシーを侵してしまう

3 各種センサを統合したウェアラブルイメージングシステムの構築

ウェアラブルイメージングシステムは、ウェアラブルカメラによって体験映像と各種センサ情報の同期記録、およびインデキシングを行うシステムである。センサとしては、ウェアラブルなカメラ、マイクに加え、位置情報を取る GPS、動き情報を取るジャイロ及び加速度センサを利用している。図 1 にウェアラブルイメージングシステムの概要を示す。カメラは帽子の先に取り付け、ユーザが見ているものに近い映像を取得する。ジャイロ、加速度センサは帽子の後方に取り付け、頭（カメラ）の動き、ユーザの前後方向、左右方向の動きを取得する。また、小型タイピン型マイクおよび GPS レシーバを肩に装着している。

現在入手可能なウェアラブル PC では、ストレージデバイスの容量と処理能力が未だ十分ではないため、本システムでは、ノート型 PC を用いて映像の記録と各種処理を行っている。カメラやその他のセンサは、USB や PC カードスロットなどを通し、PC に直接接続されており、すべての情報は PC の HDD に直接

Real-time Event Detection for Wearable Imaging System

† Yasuhito Sawahata (sawa@hal.t.u-tokyo.ac.jp)
† Tetsuro Hori (t_hori@hal.t.u-tokyo.ac.jp)
† Kiyoharu Aizawa (aizawa@hal.t.u-tokyo.ac.jp)
Department of Frontier Informatics, The University of Tokyo(†)

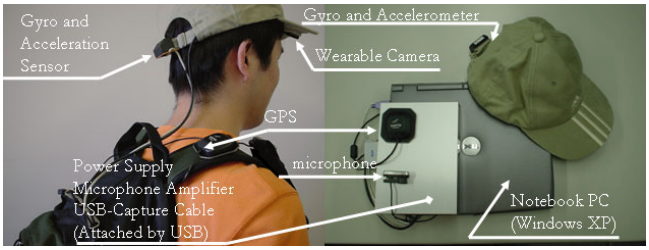


図 1: ウェアラブルイメージングシステム

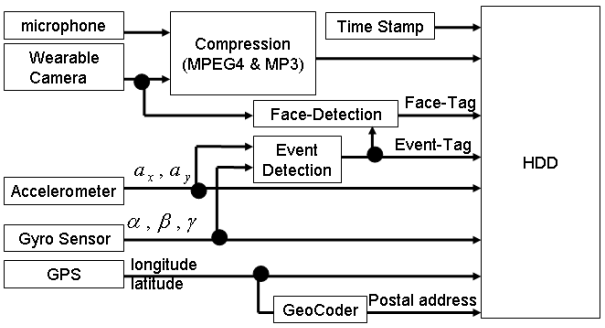


図 2: ソフトウェアのブロック図

記録される。各センサの電源は、すべて PC のバッテリーから供給するよう加工し、なるべく簡便に体験映像の記録が行えるように努めた。

ウェアラブルイメージングシステムを構成するソフトウェアは、Visual C++により記述し、Windows 上で動作する。ノート型 PC 上で走っているソフトウェアのブロック図を図 2 に示す。映像と音声に関しては、MPEG4 と MP3 にリアルタイムで圧縮しながらの記録を行っている。これにより、記録する映像の内容にもよるが、24 時間の映像を約 15[GB] から 20[GB] 程度の容量で記録することができる。すべてのセンサ情報が HDD に直接記録されるのと同時に、映像取得時のユーザの状態を把握するために、加速度センサとジャイロの情報は 4 節で述べる方法で処理が施され、その結果をインデックスとし、映像に対し付加する。同時に、ユーザが何を見ているのかを知るために、オブジェクト検出などの画像処理も行っている。また、GPS からの緯度経度情報は、データベースを参照することで番地情報に変換し、インデックスとして利用している。

4　イベントの推定

4.1　体験映像のインデキシング

人間は日常生活の中で、膨大な量の映像、音声、その他の情報を記憶している。それでも過去の出来事を、すばやく回想することができるのはなぜだろうか。このような疑問に対し、過去に行われた出来事の内容を思い出すためには、「いつ」、「誰と」、「どこで」、「その場所でなにが行われていたか」、「どのように感じていたのか」というような体験時の状態（コンテキスト）が、回想のためのキーとして重要な役割を果たしているという報告がある [4]。

このようなことから、本節ではカメラ、マイクを含めた各種センサ情報からコンテキストを推定し、それをインデックスとして予め映像に付加することにより、記憶の回想に近い自然なクエリを可能とする手法について検討する。特に「どこで」「ユーザの動作」「顔（誰か）」の検出に注目する。「どこで」は GPS にて取得した情報を参照する。「ユーザの動作」の検出は、センサの情報参照することで行う。これらは、ユーザの動作と高い相関を持つため、高精度かつ軽い処理での検出が見込める。ユーザが何を見ていたかは、ユーザのコンテキストを知る上で重要である。画像からのオブジェクト検出アルゴリズムは多数あるが、ウェアラブルでの限られた計算資源の中それらを同時に実行するのは不可能である。ここでは、日常生活の中での利用ということを考慮し、オブジェクトとして顔の検出を試みる。

4.2　動作イベント検出手法

ここでは、コンテキストとして、ユーザが「静止」「歩いている」「走っている」などの基本的な動作イベントの検出を目標とする。イベントの検出は、時系列に得られるセンサ情報を隠れマルコフモデル（HMM）によりモデル化することにより行う。

まず、加速度センサおよびジャイロセンサの出力から、特徴ベクトルを作成する。これらのセンサにより、x 方向（前後方向）の加速度、y 方向（左右方向）の加速度、x, y, z 軸回りの回転角が取得できる。これらの情報をもちいて、式 (1) のような特徴ベクトルを作成する。

$$FeatureVector = \left[\begin{array}{ccccc} a_x & a_y & \Delta\alpha & \beta & \gamma \end{array} \right]^t \quad (1)$$

本システムでは、この特徴ベクトルを毎秒 30 サンプル作成するように設定している。

HMM による学習の手順を以下に示す:

1. 取得した特徴ベクトル列を、K-Means 法を用い、C 種のシンボルからなるシンボル列に変換する。さらに各クラスタの重心ベクトル $\overline{FV}_j (0 \leq j < C)$ を保存しておく。
2. 特徴的なシンボル列を選択し、それぞれにイベント $E_i (0 \leq i < N)$ としてラベル付けを行う。N はイベントの総数を表す。
3. イベント E_i に対応する N 個の HMM λ_i を作成し、HMM パラメタの学習を行う。HMM の形状は left-right とし、状態数を S として与える。

HMM によるテストの手順を以下に示す:

1. 取得した特徴ベクトルと $\overline{FV}_j$ との距離を調べ、それを適切なシンボルに変換する。
2. シンボルをバッファに入れ、長さ L のシンボル列 $\mathbf{O}$ となるまで、ステップ (1) を繰り返す。

3. すべての HMM に対して、$P(\mathbf{O} \mid \lambda_i)$ を計算する。$P(\mathbf{O} \mid \lambda_M)$ が最大の尤度を示す場合、イベント E_M を検出する。

4.3 顔の検出

体験映像からの顔画像検出に求められる用件として、筆者らは、ウェアラブルコンピュータの限られた計算資源で実時間で処理が可能であることを設定する。すなわち、できるだけシンプルな方法での検出が望ましい。このような用件から、顔画像の検出アルゴリズムは、肌色検出をおこなうことで顔領域を抽出するという、川戸らの方法 [5] を参考にした。

その概要を以下に述べる。まず、画像をブロックに分割する。ブロックのサイズに対する肌色と判断された画素数の割合がしきい値を超えたとき、そのブロックが肌色領域であると判断する。ブロックのサイズおよびしきい値は可変であり、ブロックのサイズおよびしきい値を大きなものから小さなものへと徐々に変化させることで、そのブロックが肌色領域か否かを判断する。

肌色の検出は、RGB 値を式 (2) により ab 値に変換することで行う。

$$a = r + \frac{g}{2}, \quad b = \frac{r\sqrt{3}}{2} \tag{2}$$

但し、$r = R/(R+G+B), g = G/(R+G+B)$。

ab 値を用いることにより、RGB による輝度値の影響を抑えることができる。肌色を示す ab 値は、a_0, b_0 を中心とする正規分布と仮定している。a_0, b_0 に対する入力画素値の分散が、前もって定めたしきい値以下となるとき、その画素が肌色領域にあると判断する。フレームに顔が含まれているかどうかの判定は、さらにフレームを占める顔領域割合が、しきい値を超えたかどうかで判断する。

5 実験と検証

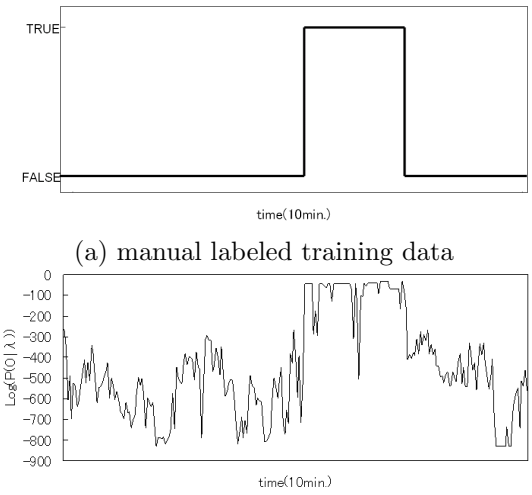
5.1 ユーザの動きの推定

HMM の学習のために、大学キャンパス内で、約 1 時間分のデータを取得した。データの前半部分を学習に、後半部分をテストとしてイベント検出を試みた。データ取得中ユーザは、静止、歩く、走るなどの動作を行っている。それぞれの動作が行われているシーケンスに対し、手作業にてラベル付けを行い、ラベルに一致する HMM の学習を行った。

図 3、図 4 は、映像をみて手作業でラベル付けしたもの及び、対応するモデルの $P(\mathbf{O} \mid \lambda_i)$ の値である。学習シーケンス長 L、状態数 S、クラスタ数 C はそれぞれ、60、10、30 に設定した。($L = 60$ ということはすなわち、シーケンス長が 2[s] ということを意味する。) なお、ここでは縦軸に対数を用いている。対数を使うことによって、尤度の計算時に観測確率および遷移確率の積算を L 回行うことで引き起こされ得るアンダーフローを回避している。したがって、図 3、図 4 における低い値のばらつきは、より強調されて見えていることになる。これらを考慮すると、手作業でラベル付けしたデータと実際の観測確率を比較すると、非常に高い相関があることが読み取れる。他のモデルとその観測確率を比較することで、イベントの種類を判断することが可能であることが分かる。

表 2: 推定したイベントの確かさ

Events	Accuracy [%]
Walking	97.1
Running	88.8
No Move	89.7



(a) manual labeled training data

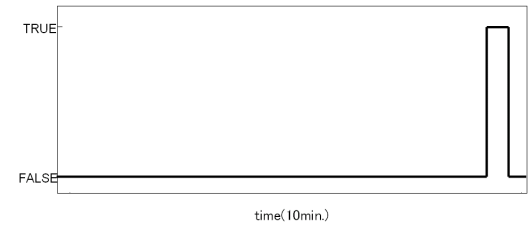
(b) likelihood

図 3: ラベルと P($\mathbf{O} \mid \lambda$="no move") の値の関係

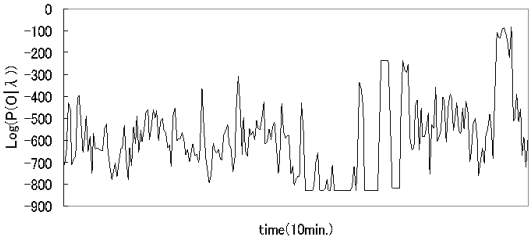
表 2 に、HMM により推定したイベントと実際のイベントとの関係を示す。この関係は、イベント E_X とラベル付けされたシーケンスの数に対する、HMM によりイベント E_X と判定されたシーケンスの数の割合を示す。歩いているシーン、走っているシーン、静止しているシーンなどは非常に高い精度で検出できていることがわかる。

5.2 位置情報を用いた映像へのアクセス

位置情報は、検索を行うためのクエリとして、非常に有用である。そのため、ウェアラブルイメージングシステムでは、GPS レシーバを用いて位置情報を取得している。図 5 に位置をベースとしたビューワを示す。GPS により取得したデータにより、ユーザが実際に移動した軌跡を地図上に表示している。マウスを用いて地図上の軌跡を範囲指定することで、その範囲の映像を取得することが可能である。また、GPS からの緯度・経度情報を、データベースを参照し番地情報に変換することにより、地図からだけではなく、「○○県△△市××町×丁目の映像」をクエリとして映像の検索を行うことも可能である。緯度・経度の情報を



(a) manual labeled training data



(b) likelihood

図 4: ラベルと P(**O** | λ=“running”) の値の関係

図 5: 位置に基づくビューワ

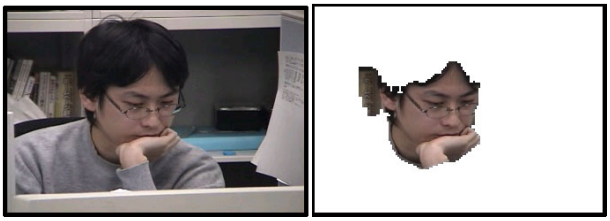
階層的な構造を持つ番地情報に変換することは、膨大な量のデータを扱う上で有効である。

5.3 顔の検出

図 6 に顔領域の抽出結果を示す。フレームを占める顔領域の割合が、しきい値を上回る時に顔を検出するが、このしきい値は可変として、ユーザが GUI を通して設定できるようにしている。しきい値を下げすぎると、顔領域が小さい場合も検出できるようになるが、誤検出の影響も顕著となり、顔画像を捉える精度が著しく低下する。

顔が検出されたフレームは、GUI 上に列挙され、ユーザはそれを選択することにより所望の映像を取得することができる。

顔検出の精度向上の手法として、ユーザが静止しているシーンにのみ顔検出を試みるという手法が挙げられる。これは、ユーザが何かをじっと見ているときは、ユーザは静止しているという仮定に基づく。これにより、誤検出を低減することができると同時に、顔検出処理を施すシーンをユーザが静止しているシーンに制限することができ、それにかかる計算量を大幅に減らすことができる。



(a) 元画像　　(b) 抽出結果

図 6: 顔の検出結果

6 おわりに

本稿では、体験映像の記録と同時に各種センサ出力を同期記録し、効率的な映像取得を行なうためのウェアラブルイメージングシステムの概要について述べた。インデキシングを行う際のメタデータとして、ユーザ状態を用いるため、人間の記憶の回想に近い形で映像取得を行なうことができる。GPS により取得した位置情報、HMM により検出した動作イベント、肌色検出による顔検出をそれぞれ組み合わせることにより、「どこで」「どんなとき」「誰か」をクエリとして映像に対してアクセスが可能である。

今後は、新たなセンサを導入し、特徴ベクトルを工夫することで、より複雑なイベントの検出を行うことを目標とする。さらに、映像の検索を行うための GUI の設計を進める予定である。

参考文献

[1] Kiyoharu Aizawa, Ken-Ichiro Ishijima, Makoto Shina,“Summarizing Wearable Video,” Proceedings of ICIP2001, pp 398-401, Oct。2001

[2] Haung Wei Ng, Yasuhito Sawahata and Kiyoharu Aizawa, “SUMMARIZATION OF WEARABLE VIDEOS USING SUPPORT VECTOR MACHINE,” Proceedings of ICME 2002, IEEE, Aug。2002

[3] 澤畠 康仁、相澤 清晴、“センサ情報からイベント検出を行うウェアラブルイメージングシステム,” 電子情報通信学会技術報告, MVE2002-81, pp.81-86, Nov。2002。

[4] Mik Lamming and Mike Flynn, “‘Forget-me-not’ Intimate Computing in Support of Human Memory,” Proceedings of FRIEND21, ’94 International Symposium on Next Generation Human Interface, Feb。1994

[5] 川戸 慎二郎, 鉄谷 信二, “顔領域抽出を目的とした肌色モデルと肌色領域抽出,” 電子情報通信学会技術報告, PRMU2002-59, pp143-148, 2001.