オーディオビジュアル複合情報処理　38－28
(2002. 11. 15)

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

澤畠　康仁†　　相澤　清晴†

† 東京大学大学院新領域創成科学研究科　〒113-8656 東京都文京区本郷 7-3-1
E-mail: †{sawa,aizawa}@hal.t.u-tokyo.ac.jp

あらまし　本稿では、ウェアラブルコンピュータを用い、人間が日常生活において目にするすべての映像（体験映像）の記録を試みる。この際、膨大な映像データの中からいかにして自分の見たい映像を効率よく取得するのかという、インデキシングが非常に重要となる。本稿では、体験映像に付加するインデックスとして、映像取得時におけるユーザの状態（コンテキスト）を用いた方法を提案する。映像、音声に加え、コンテキストを抽出するための各センサ情報を記録することにより、効率的なインデキシングおよび映像取得・閲覧を行うウェアラブルイメージングシステムについて述べる。特にユーザのコンテキストとしては、ユーザの動きに注目し、その検出には隠れマルコフモデルを用いて行っている。

キーワード　ウェアラブル、イベント検出、センサ、インデキシング

Wearable Imaging System: Detecting Events from Sensor Information

Yasuhito SAWAHATA† and Kiyoharu AIZAWA†

† Department of Frontier Informatics, University of Tokyo　Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-8656 Japan
E-mail: †{sawa,aizawa}@hal.t.u-tokyo.ac.jp

Abstract　Digitization of lengthy personal experiences would be made possible by constant recording using wearable video cameras. It is conceivable that the resulting amount of video content would be extraordinarily large. In order to retrieve and browse the desired scenes, a vast amount of video would need to be organized with structural information. In this paper, we attempt to develop a "Wearable Imaging System" that is capable of constantly capturing data, not only from a wearable video camera, but also from various kinds of sensors. The data from these sensors are appropriately extracted and processed by Hidden Markov Model (HMM) to achieve efficient video retrieval and browsing.

Key words　Wearable, Event Detection, Sensor, Indexing

1. はじめに

近年、ウェアラブルコンピューティングに代表されるように、人間の環境に密着した新しいコンピューティング環境への期待が高まっている。これまで筆者らは、このようなウェアラブル環境でのセンシングに注目し、ウェアラブルカメラ・マイクを用いて日常体験を映像として常時記録するという試みを行ってきた[1], [5], [6]。体験映像の常時記録に際しては、膨大な映像データの中から、いかにして自分の見たい映像を効率よく取得するのかというインデキシングが重要な課題となる。

膨大な映像データを効率よくインデキシングするためには、カメラ、マイクのほかにユーザの状態（コンテキスト）を把握するためのさまざまなセンサを用いることが効果的である。映像に対する付加情報としてコンテキストを用いることで、記憶の回想を手本とした、人間にとって自然な形での検索が可能になる。

本稿では、映像、音声、各センサ情報を記録することにより、効率的なインデキシングおよび映像取得・閲覧を行うウェアラブルイメージングシステムについて述べる。特にユーザのコンテキストとしてユーザの動きに注目し、その検出は、センサからの情報を隠れマルコフモデルによって処理することによって行っている。

以下2節では、体験映像蓄積の利点と可能性、3節では体験映像を取得するためのウェアラブルイメージングシステムの概要を述べる。4節では体験映像のインデキシングを行うためのコンセプトを示し、5節でユーザ状態把握のための手法について述べる。6節で実験の結果について述べ、最後に7節でまとめる。

表 1　体験映像の量
Table 1　Amount of Wearable Video

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

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。100GB 前後の HDD は、15000 円〜20000 円程度で販売されており、現在の HDD を用いても、100GB の HDD が 100 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。

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、その範囲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 レシーバを肩の部位に装着している。

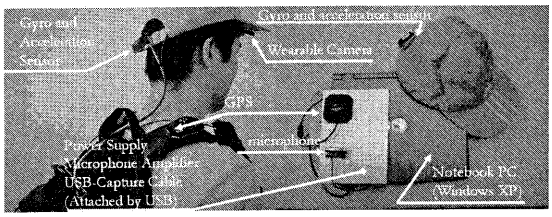


図 1　ウェアラブルイメージングシステム
Fig. 1　Wearable Imaging System.

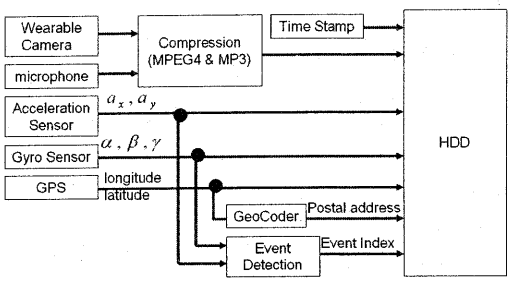


図 2　ソフトウェアのブロック図
Fig. 2　Block Diagram.

現在入手可能なウェアラブル PC では、ストレージ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 および 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 5. 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。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。

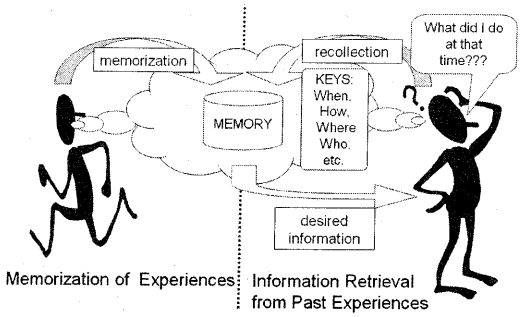


図 3　記憶の回想

Fig. 3　Human's memory recollection.

4. 体験映像のインデキシング

マルチメディアコンテンツのインデキシングの方法としては、MPEG-7 が有名である。MPEG-7 は、コンテンツの意味内容や、色情報や動き情報等の映像・音声のもつ低レベル特徴を記述するための規格である。一般的なマルチメディアコンテンツ向けに策定された技術標準であるが、対象を体験映像に特化して考えるとき、MPEG-7 で記述される特徴は、必ずしも適当であるとはいえない。体験映像として重要となる特徴は、何かについて考える必要がある。

ところで、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、効率よ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、過去に行われた会話の内容を思い出すためには、図 3 のように、「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 [2]。

そこでウェアラブルイメージングシステムは、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の利用を可能とする。センサ情報は、ユーザの状態と高い相関を持つため、映像・音声のみを用いるよりも高精度かつ軽い処理での特徴抽出が見込める。

例えば、会話のシーンの検出を考える。会話の検出には、音響情報を調べることが有効であるが、その処理を映像全体に施すことは効率的ではない。会話の最中はユーザの動きが少ないと仮定を設け、動きの少ないシーンを検出することで、処理を施すシーンを限定することができ、効率よく欲しいシーンの検出が可能となる。

以上のことから、体験映像のインデキシングを行う際、コンテキストの推定という技術が、非常に有効であると考える。

5. センサを用いたイベント検出

5.1 コンテキストの分類

コンテキスト (context) とは、ユーザの置かれている状況や状態を表す。ここでは、このコンテキストを、主観的なものと客観的なものの二つに分類して考える。

主観的なものとは、ユーザがどのように感じているのかというような、心理的な状態を表す。たとえば、「興味をもっている」「興奮している」「緊張している」などということに相当する。筆者らは、脳波の計測を行うことで、興味を持つか持たないかという、ユーザの主観的状態を直接的に抽出するという試みを行ってきた [1], [5], [6]。また、文献 [3] では、皮膚導電率を計測することでユーザの緊張状態を抽出し、映像のインデックスとして利用している。文献 [4] では、心拍変動量の計測により、ユーザの行動と心的変化の関連付けを行っている。このように、ユーザの心理的状態を抽出する手法がいくつか報告されている。

客観的なものとは、ユーザが何をしているのか、どこにいるのかというような、物理的に決定された状態を表す。これらは、動きを取得するセンサ、場所を取得するセンサなどを用いることで、高い精度で取得することができる。

ウェアラブル環境におけるセンシングは、環境の大きな変化に対してもロバストに行える必要がある。このようなことを鑑みると、生理的な信号の取得を要する心理状態の抽出は適切であるとはいえない。たとえば、脳波の計測では、ユーザはなるべく安静な状況にいることが望ましい。ユーザが激しく運動すると、それに起因して計測されるノイズにより、取得したデータの信頼性が大きく低下してしまう。一方で、客観的状態の抽出は、比較的環境の変化にロバストに行えるという特徴がある。

ウェアラブル環境での主観的状態を取得するのが不可能かというとそうではない。人間は、その仕草や様子などから、「楽しそう」「緊張してそう」などの心理状態を推定することができる。このような推量を行うための手がかりとなるのは、その人の動きや表情などの、客観的な情報である。すなわち、客観的な情報を多く取得することで、ユーザの主観的状態を抽出することが可能であることを意味する。

このようなことから、本稿では、主観の抽出は将来的な課題として据えておき、まずはユーザの客観的な状態を抽出に絞り、その手法について検討を行う。

5.2 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）[7] 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、図 4 のように、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。

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

ここで、a_x, a_y は x, y 方向における加速度、α, β, γ は、z, y, x 各軸回りの回転角を表す。また、$\Delta\alpha$ は、α の時間的な差分で

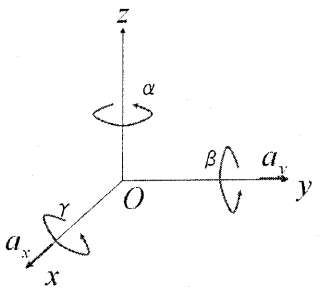


図 4　ジャイロ、加速度センサの出力
Fig. 4　Output of Gyro and Acceleration Sensor.

あり、$|\alpha_t - \alpha_{t-1}|$ とする。現在のシステムでは、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） センサから得られるサンプルを保存し、式 (1) で表される特徴ベクトルを作成する。

（2） K-Means 法を用いて、特徴ベクトル列をシンボル列に変換する。ここで、クラスタ数 C をパラメタとして与える。これによりすべての特徴ベクトルは C 種類のシンボルにクラスタリングされる。さらに各クラスタの重心となるベクトル $\overline{FV}_j(0 \leqq j < C)$ を保存しておく。

（3） 特徴的なシンボル列を選択し、それぞれにイベント $E_i(0 \leqq i < N)$ としてラベル付けを行う。N はイベントの総数を表す。すなわち HMM の総数に一致する。

（4） 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、すべての HMM パラメタは乱数により初期化しておく。ここでは HMM の状態数を、S として与える。

HMM によるテストの手順を以下に示す:

（1） センサからデータを取得し、式 (1) で表される特徴ベクトルを 1 つ作成する。

（2） 作成した特徴ベクトルをシンボル列に変換する。これは $\overline{FV}_j$ との距離を調べることによりおこなう。現在の特徴ベクトルが $\overline{FV}_K$ にもっとも近い場合、シンボル K に量子化される。

（3） 長さ L のシンボル列を作成する。シンボルはバッファに入れられ、バッファ内のシーケンスの長さが L に満たない場合は、ステップ (1)、(2) を繰り返す。L は、HMM に入力されるシーケンスの最大の長さであり、フレーム長に相当する。バッファ内のシーケンスの長さが L に達したとき、それらのシンボルを、観測シンボル列 $\mathbf{O}$ として定義する。

（4） 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。

文献 [8] では、HMM を用いたイベント検出を行っている。HMM を用いたイベント検出手法は、実際のイベントと推定されるイベントとの間に高い相関関係があり、有効な手法であることが報告されている。しかしながら、映像・音声から特徴ベクトルを作成しており、環境の大きな変化に対してロバストではないことが予想される。彼らの手法に対して、本システムでは環境の変化に強いセンサの出力を用いて特徴ベクトルを作成している。したがってそのイベントの検出精度も、環境の変化にロバストに対応可能であることが期待できる。

表 2　推定したイベントの確かさ
Table 2　Accuracy of Estimation

Events	Accuracy [%]
Walking	97.1
Running	88.8
No Move	89.7
Stairs	53.2

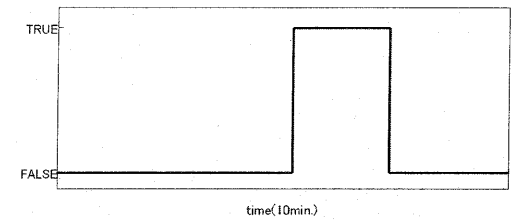
6. 実験と検証

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。

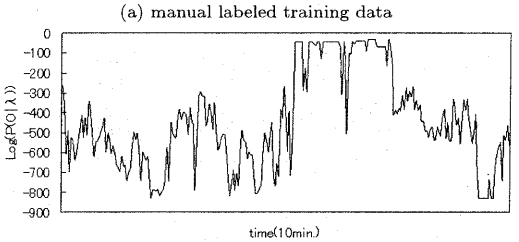
図 5、図 6 は、映像をみてマニュアルでラベル付けしたもの及び、対応するモデルの $P(\mathbf{O} \mid \lambda_i)$ の値である。学習シーケンス長 L、状態数 S、クラスタ数 C はそれぞれ、60、10、30 に設定した。($L = 60$ ということはすなわち、シーケンス長が 2[s] ということを意味する。) なお、ここでは縦軸に対数を用いている。対数を使うことによって、尤度の計算時に観測確率および遷移確率の積算を L 回行うことで引き起こされ得るアンダーフローを回避している。したがって、図 5、図 6 における低い値のばらつきは、より強調されて見えていることになる。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 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確率で検出できていることがわかる。一方で階段を上り下りしているシーンは、低い検出結果となった。階段のシーンは、同時に歩くという動作も行っており、それが検出結果の低下の原因である。現在の特徴量は、5 次元の単純なものであるため、いくつかの動作を含むような複雑なイベントの検出には十分ではないことが分かる。より複雑なイベントに対処するためには、さらにセンサを増やすなど、特徴量の再考が必要となろう。

また、体験映像の取得を動物園にてもおこなった。動物園では、動物を見るときは立ち止まり、他の動物を見るために移動したりと、人は移動と静止をしばしば繰り返す。ゆえに、ユーザが静止しているシーンを取得できれば、そのシーンに動物が



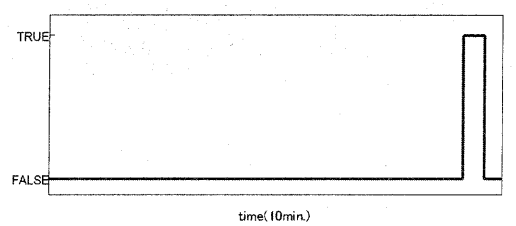
(a) manual labeled training data



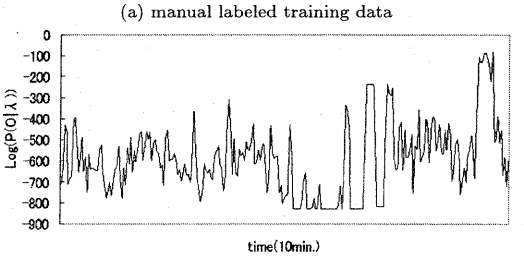
(b) likelihood

図 5　ラベルと P(**O** | λ="no move") の値の関係

Fig. 5　Manual labeled training data and transition of P(**O** | λ="no move").



(a) manual labeled training data



(b) likelihood

図 6　ラベルと P(**O** | λ="running") の値の関係

Fig. 6　Manual labeled training data and transition of P(**O** | λ="running").

移っている確率は高いことが予想できる。図 7 は、静止イベントに対応する HMM の尤度を表したものである。図 8 は、静止イベント以外のイベントから、静止イベントへ変化したシーンの最初のフレームを示したものである。ここで用いた HMM は、大学キャンパスで取得したデータにより学習されている。各キーフレームは、動物の檻の前、場内案内の看板、みやげ店などのシーンを捉えていることが分かる。

ウェアラブルイメージングシステムは、GPS をもちいて緯度・経度の情報を取得している。図 9 に位置をベースとしたビューワを示す。緯度・経度は、データベースを参照することで番地情報に変換される。緯度・経度の情報そのものでは、人

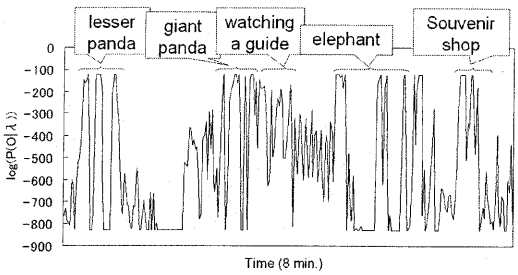


図 7　動物園のデータに対する静止イベントの出力結果

Fig. 7　A result of "no move" event detection performed for zoo data.

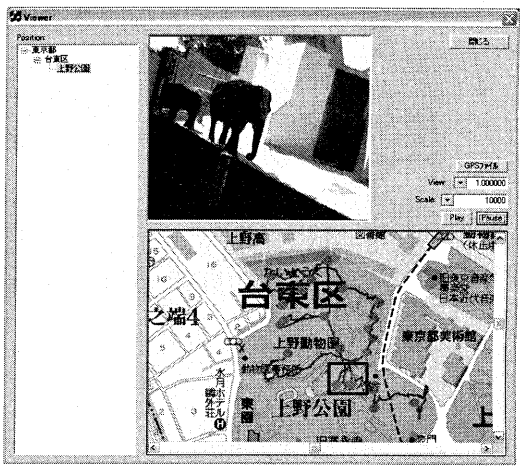


図 9　位置に基づくビューワ

Fig. 9　Location-based Viewer.

にとって可読性に欠く物であるうえに、構造的でなく管理が容易ではない。しかし番地情報は、階層的な構造をもっているので、膨大な量のデータを取り扱う上で効率が良い。番地情報は、図 9 内左側のツリービューに追加される。ツリービューの番地を選択することで、その付近の地図を表示するようになっている。緯度・経度の情報は、ユーザの移動した軌跡をプロットするためにも用いられる。地図上にプロットされた軌跡を選択することで、所望のシーンにアクセスすることができる。なお、軌跡上にある点は、イベントの変化を検出した点を表している。図 9 では、静止イベントを表している。

このように、ウェアラブルイメージングシステムは、検出したイベント、GPS による位置情報、時間情報を組み合わせることによって、「いつ」「どこで」「どんなときに」という情報をクエリとして、所望の映像にアクセスすることが可能である。

7. 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

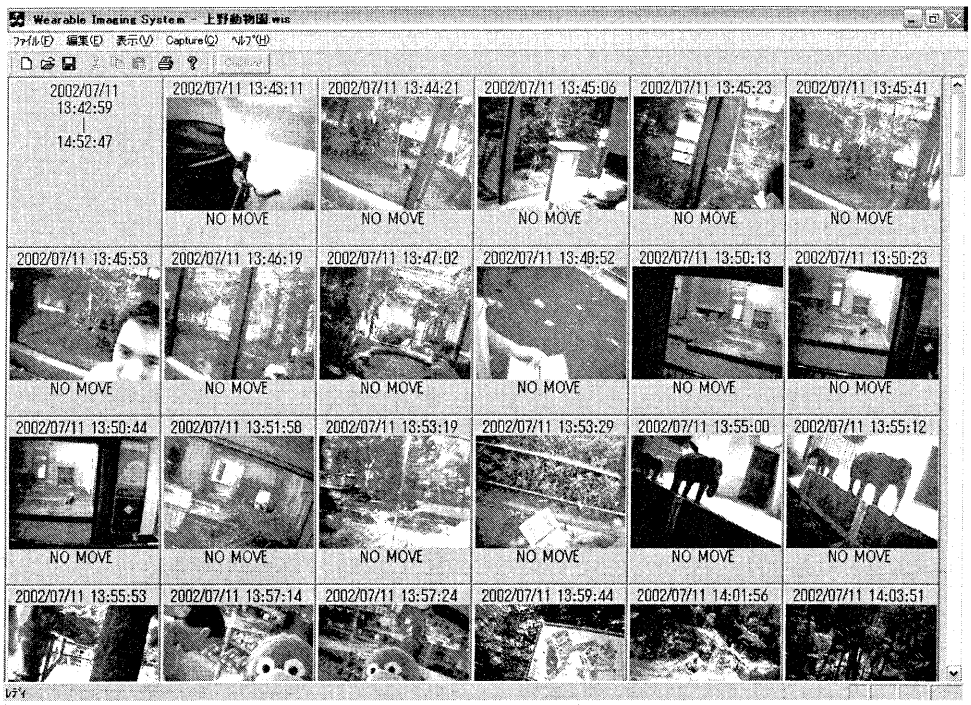


図 8　キーフレームビューワ
Fig. 8　Keyframe Viewer.

回想に近い形で映像取得を行なうことができる。センサ情報をHMMによりモデル化することによる、ユーザの動作イベント検出を試みた。

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。

文　献

[1] Kiyoharu Aizawa, Ken-Ichiro Ishijima, Makoto Shiina, Summarizing Wearable Video, Proceedings of ICIP2001, pp 398-401, Oct. 2001

[2] Mik Lamming and Mike Flynn, 'Forget-me-not' Intimate Computing in Support of Human Memory," Proceedings of FRIEND21, '94 Int. Symp. on Next Generation Human Interface, Feb. 1994

[3] Jennifer Healey and Rosalind W. Picard, A Cybernetic Wearable Camera, Proceedings of ISWC98, pp 42-49, Oct. 1998

[4] 上岡 玲子, 広田 光一, 廣瀬 通孝, "体験記録装置としてのウェアラブルコンピュータの研究," 日本バーチャルリアリティ学会第6回論文集, pp.149-152, September 2001.

[5] 澤畠康仁, ウン ハン ウェイ, 相澤清晴, 体験映像要約のための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, MVE2002-1, pp.1-6, June 2000.

[6] Haung Wei Ng, Yasuhito Sawahata and Kiyoharu Aizawa, SUMMARIZATION OF WEARABLE VIDEOS USING SUPPORT VECTOR MACHINE, Proceedings of ICME 2002, IEEE, to appear, Aug. 2002

[7] L. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE, 77(2):257-286, Feb. 1989

[8] Clarkson, B. and A. Pentland, Unsupervised Clustering of Ambulatory Audio and Video, Proceedings of ICASSP'99, 1999, http://www.media.mit.edu/ clarkson/icassp99/ icassp99.html