

個人行動記録システムにおける注目シーンの検出

注目シーン検出の高精度化と環境カメラの利用

久保田敏司[†] 中村 裕一[†] 大田 友一[†]

[†] 筑波大学 機能工学系 〒305-8573 茨城県つくば市天王台 1-1-1

E-mail: [†]{kubota,yuichi,ohta}@image.esys.tsukuba.ac.jp

あらまし 我々の提案する個人行動記録システムでは、頭部に装着したカメラから得られる個人視点映像を自動要約し、利用者が目的のシーンにアクセスしやすいようにインデキシングする。本稿では、そのための注目シーン検出に関して新しい手法を提案する。具体的には、カメラの動き推定のモデルを切り換えながら計算を行う手法と、変化領域の検出手法を提案する。また、室内に取りつけた視野の広い環境カメラを併用することにより、大局的な情報と同期した個人行動記録を得る方法を提案する。

キーワード 個人視点映像, 映像要約, 注目シーン, 動き推定, 環境カメラ

Detecting Scenes of Attention from Personal View Records

— Motion estimation improvements and cooperative use of a surveillance camera

Satoshi KUBOTA[†], Yuichi NAKAMURA[†], and Yuichi OHTA[†]

[†] IEMS, University of Tsukuba Tennoudai 1-1-1, Tsukuba, Ibaraki, 305-8573, Japan

E-mail: [†]{kubota,yuichi,ohta}@image.esys.tsukuba.ac.jp

Abstract This paper introduces a novel method for analyzing video records captured by a head-mounted camera. For this purpose, we previously reported that *scenes of attention* can be good indices for summarizing those videos. This paper introduces two new approaches for the performance improvement and for the extension of the potential applications. One is a new method of two-step motion estimation that adaptively uses either of the 2D affine model or the 3D rigid-body motion with central projection model. The other is the use of a wide-angled surveillance camera. The view from the camera is cooperatively used for delineating the location where the user acted, and for clearly presenting the situation.

Key words head-mount camera, video summarization, scenes of attention, motion estimation, surveillance camera

1. はじめに

個人の見聞きした情報を映像を用いて記録し、必要に応じてそれを再現することができれば、記憶の補助、記憶の共有等、多くの用途に利用できる [1] [2] [3]。我々はこのような研究の一つとして、利用者の頭部に小型カメラを装着し、得られた映像から、比較的粒度の細かい行動記録・要約を実現する手法を提案してきた [4] [5]。ここでは、“何をしたのか”だけではなく、“どのようにしたか”が簡単に分かるように映像を構造化することを目的としている。

そのための基本的なアイデアは、カメラ装着者の頭部の動きと画像中の物体の動きを検出し、その相互関係により、重要なシーンを検出することである。この重要なシーンとしては、

カメラ装着者が一カ所にとどまっていたり、何かに注目した場合を対象とし、それを注目シーンと呼ぶ。

注目シーン検出のために我々が提案してきた手法 [4] [5] は、概ね良い結果を出すことができるが、誤検出や計算量の多さなどの問題が残っていた。そのため、本研究では、注目シーン検出をより安定かつ一般的に使えるように、以下の2つの新しい手法を採り入れた。

- 画像間の動き推定を行う際に、2次元モデルと3次元モデルを切り換えながら計算することによって、精度の向上と計算時間の短縮を実現した。さらに、誤差の分布を考慮した2段階の変化領域検出を行うことにより、エッジや細かいテクスチャ部分が誤検出されるのを軽減した。

- 室内を大域的に観測する環境カメラからの映像を用いることによって、個人視点映像だけでは捉えにくい大域的な情報を利用可能にした。

以下、本稿では、まず我々の注目シーン検出の考え方について説明し、その改良方法、さらに、環境カメラの利用について述べる。

2. 注目シーンとその検出

2.1 注目シーン

個人視点映像中の注目シーンを、本研究では次のように定義している

積極的注目シーン： 人は何かに興味を持つと、その対象をじっと見続けようとする。見続けようとする典型的な動作としては、図 1(a) のように人が静止したまま静止している対象物をじっと見る場合や、図 1(b) のように動いている物体を首を振りながら、または体を動かしながら視界に入れ続けるよう追いかける、その他、図の (c) や (d) 等といった動作がある。本研究では、これらの動作が起った場面を積極的注目シーンと呼ぶ。

停留シーン： デスクワーク、他人との会話、講演の聴講のような場合には、同じ場所で比較的長時間、同じ出来事を見続ける。この場合は、図 1(a) に体の揺れが複合的に加わったものになる。この場合、シーンの長さが長時間におよぶ。本研究ではこれを停留シーンと呼ぶ。

各々を検出する手がかりは以下ようになる。

背景と異なる動きをすること： 積極的注目シーンを検出するためには、頭部装着カメラの移動による背景の動きと異なる動きをしている領域を検出する。例を図 2 に示すが、このような領域が画像上の一定の位置にあり続けるときに、これを注目対象の候補とし、そのシーンを注目シーンとする。ここで、物体がただ目前を横切った場合のように、実際には注意を向けられていない可能性が高い部分を除くために、この領域が画像中の一定の部分に存在し続けることを条件としている。

背景の動きが小さいこと： 停留シーンを検出するためには、背景の動きが小さいことを検出すればよい。これを本手法では 2 つの方法で検出する。まず、2 枚の画像の差分をとり、その値が小さい場合に、動きがないとする。差分値がある程度以上の場合には、以下に述べる動き推定を行うが、その値が小さくとどまっている場合には、動きがない場合と同様、停留しているとする。

2.2 動き推定処理を用いた注目箇所の抽出

本研究では、注目シーンの検出を以下のような手順で行う。

1. まず、頭の動きに起因する見かけの動きを推定するために、時間的に 1～数フレーム離れた 2 枚の画像間の対応関係を求める。
2. 求めた対応関係によって、視野の移動が小さいと判定される場合には、隣接するフレームをまとめて“停留シーン”

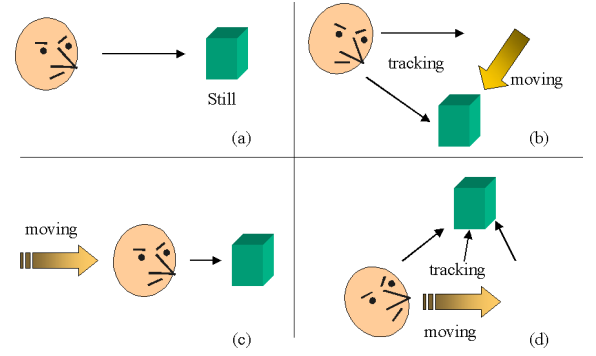


図 1 注目行動の例

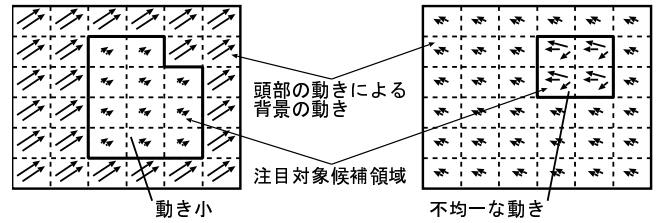


図 2 注目行動によって現れる画像特徴

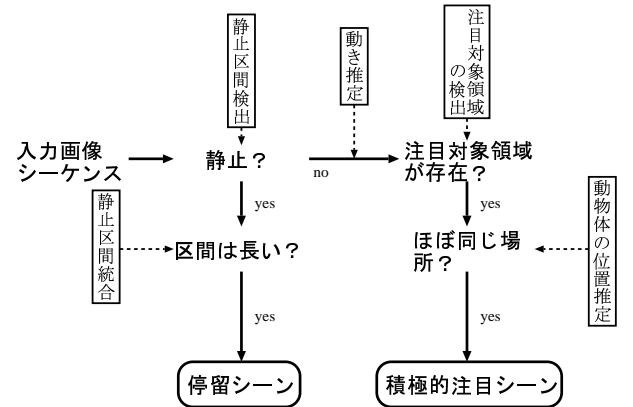


図 3 注目シーン検出の概略

とする。

3. 停留シーンと判定されなかったフレームについて、注目対象の検出を行う。注目対象が存在し、その見かけの位置が大きく変化しない部分を、“積極的注目シーン”とする。

概略を図 3 に示すが、詳細は [4] を参照して頂きたい。

上記の手法では、連続するフレーム間の動き推定に中心投影による 3 次元剛体運動モデルを使用し、シーンの奥行きを求めながらカメラの動きパラメータを推定する。

画像内の点 x の動き $u(x)$ は、カメラの並進運動 $t = (t_1, t_2, t_3)$ 、回転運動 $\omega = (\omega_1, \omega_2, \omega_3)$ を用いて、以下の式 (1) で表すことができる [6]。

$$u(x) = \frac{1}{Z(x)} A t + B \omega \quad (1)$$

$$A = \begin{bmatrix} -f & 0 & x \\ 0 & -f & y \end{bmatrix}$$

$$B = \begin{bmatrix} (xy)/f & -(f^2 + x^2)/f & y \\ (f^2 + y^2)/f & -(xy)/f & -x \end{bmatrix}$$

ただし、 f は焦点距離、 $(x, y, Z(x))$ は中心投影における点 x の 3 次元座標点である。

時刻 T における点 x における輝度値を $I(x, T)$ とすると、理想的には

$$I(x, T) = I(x - u, T - \delta t) \quad (2)$$

となることから、以下のエラー E が最小になるようなパラメータを決定する。

$$E = \sum_{x,y} (I(x, T) - I(x - u, T - \delta t))^2 \quad (3)$$

処理手順は次のようになる。

1. 平滑化とダウンサンプリングによる多重解像度画像の生成を行う。ここでは、入力画像に対して、その一辺の比が各々 $1/2$, $1/4$ の大きさの画像を生成する。
2. 以下を最も高い解像度まで順に行う。
 - カメラ運動パラメータの初期値を付与する。ここでは、一つ粗い解像度で推定された動きパラメータを初期値とする。ただし、最も粗い解像度の初期値は、1 つ前のフレームに対する動きパラメータとする。
 - 式 (3) のエラーが最小値になるパラメータを繰り返し計算により求める。本研究では、Levenberg-Marquardt 法を用いる。

種々の動き推定処理があるなかで、このような方法をとったのは以下のような理由からである。

- 一般的に、頭部にカメラを装着すると、動きが激しい画像となる。また、大型のカメラを装着できないため、質の悪い動画となる。そのため、局所的な特徴が複数のフレームに渡って安定してとらえられる保証がない。
- 目の前に様々な奥行きが物体が存在する。つまり、最も基本的な image mozaicing [7] で考えられているような条件が成り立たない。そのため、単純な 2 次元射影変換のようなモデルでは不十分である。

2.3 従来手法の問題点と解決策

上で述べた方法でも比較的良好な結果は得られているものの、次のような問題点があった。

- 画像間の視差がほとんどない場合でも、視差を計算するモデルを用いること。
- 動き推定が良好に計算され、その誤差が小さい場合でも、変化の激しいテクスチャやエッジ部分を動物体領域 (変化領域) として誤検出してしまうことがある。
- 個人視点のカメラでは、重要な情報を落している場合がある。例えば、どこで作業を行っているのかが判断しにくい場合や、背後で他人が重要な行動をしている場合もある。

このような問題に対し、本研究では次のような対策を行うことで、安定した動き推定とより広い場面での応用を可能にする。

- 2 次元アフィン変換モデル (以後、2 次元モデルと呼ぶ) を併用し、2 次元モデルと 3 次元モデルを切り換えながら動き推定を行う。
- 2 段階の変化領域検出を行う。1 段目で変化領域候補となった部分についても、もう一度マッチングを行い、誤検出されている部分を減らす。
- 環境カメラからの映像を同期して記録し、それを利用することにより、人物の位置や移動軌跡を映像のインデクスとして利用する。また、その映像自体を情報提示に利用する。

3. 新しい動き推定処理

3.1 動き推定モデルの切り換え

本研究では 2 次元アフィンモデルと中心投影による 3 次元剛体運動モデル (以下、3 次元モデルと略す) を切り換えることにより、精度と計算量の両面での性能を改善する。

まず、2 次元モデルにおける画像内の点 x の動き $u(x)$ は、次の式 (4) で表す。

$$u(x) = A_1 X(x) + A_2 \quad (4)$$

ただし、

$$A_1 = \begin{bmatrix} a_1 & a_2 \\ a_4 & a_5 \end{bmatrix} \quad X(x) = \begin{bmatrix} x \\ y \end{bmatrix} \quad A_2 = \begin{bmatrix} a_3 \\ a_6 \end{bmatrix} \quad (5)$$

2 次元モデルは計算が単純で、前節で述べた繰り返し最適化を用いて式 (3) を最小化しても、比較的少数回の繰り返しで収束する。しかし、我々の想定する環境では、アフィンモデルの前提条件が成立しない場合も多く、全ての画像対に対して単純に適用することはできない。そのため、まず 2 次元モデルによって計算を行い、残差又は推定された動きが大きい場合に 3 次元モデルに切り換える方法をとる。

処理の流れを図 4 に示すが、動き推定の切り換え、パラメータ変更は以下のように行う。

- 残差が大きい場合には、2 枚の画像間のフレーム間隔を小さくする。
- カメラの動きが大きく且つ、2 次元モデルを使っている場合には、3 次元モデルに切り換える。
- 動きが小さい場合には、フレーム間隔を大きくする。

この手法により、視差がない場合や動きが激しい場合等でも安定した動き推定を行うことができる。

3.2 2 段階の変化領域検出

従来の方法では、注目箇所を検出する際の誤検出が多いことが問題の一つであった。そこで、一度変化領域^(注1)となった部分についても、もう一度チェックを行うことにより、エッジや

(注1): 背景と異なる動きをした移動物体の領域を変化領域と呼ぶ。注目箇所の候補として検出される。

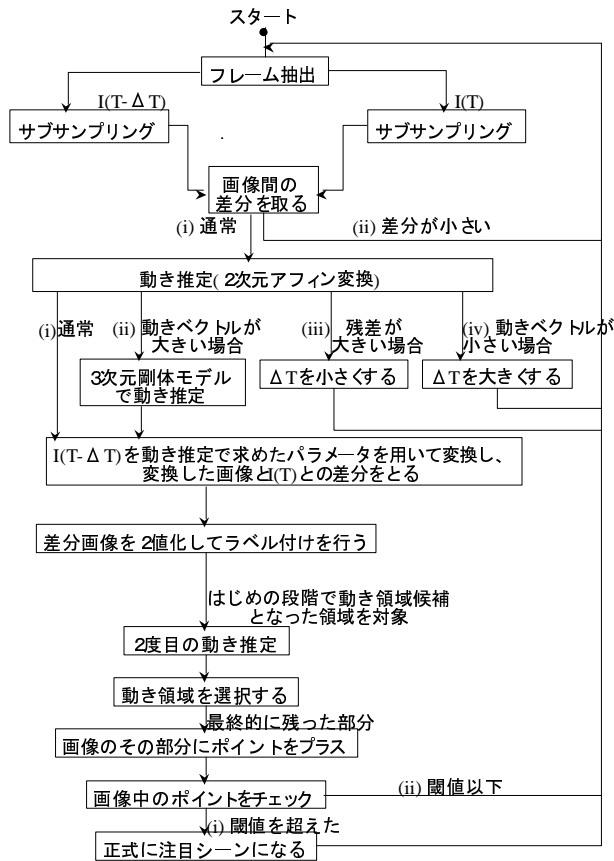


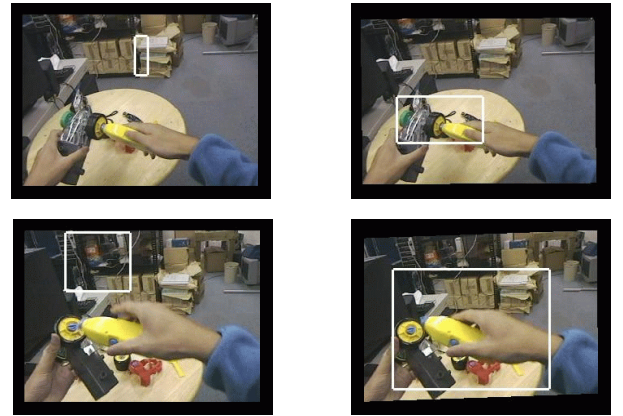
図4 動き推定処理の流れ

細かいテクスチャが抽出されてしまう問題を軽減する。

そのための前準備として、我々は、2次元モデルと3次元モデルの動き推定の精度に関して実験を行い[8]、それぞれのモデルを用いた場合の推定誤差を比較した。推定方法は以下のようになる。

1. 画像 I_i と奥行き画像 D_j の組を用意する。実験では、普段対象となることの多い室内シーンのテクスチャを I_i とし、室内、廊下、近接物体があるシーン等を簡略化した奥行きを D_j として選んだ。
2. 正解となる動きパラメータのセット (P_k) をあらかじめ多数用意する。そのため、実際に頭部の動きを磁気センサで計測し、歩行時、作業時の頭の動きを動きパラメータのセットとして用意した。
3. D_j を仮定し、 P_k によって I_i を変換することによって、新たな見え方画像 I'_{ijk} を生成する。また、この際に画像内の各点の移動ベクトル $v_a(x)$ を計算し正解データとする。
4. 既存の画像 I_i と新たな見え方画像 I'_{ijk} 間の動き推定を行ない、これによって推定された動きベクトル $v_b(x)$ と正解データ $v_a(x)$ を比較することによって、動き推定の評価とする。

紙面の都合上詳細は省くが、モデルの性質から容易に推測されるように、2次元モデルを用いた動き推定はカメラの動きが小さいときには安定しているが、カメラの動きが大きくなると



(a) 従来手法による結果

(b) 新しい手法による結果

図5 提案手法により結果が改善された例

精度が悪くなる。また、3次元モデルを用いた場合には、カメラの動きが大きいつきでも比較的结果が安定している。平均的な場面として、作業を行う程度の頭の動きを想定した結果、画像中心部で3画素～数画素、周辺部ではその倍程度の誤差となった。

本研究では、この動き推定誤差の範囲内で、もう一度変化領域と周辺領域との相関を確かめる。その結果、相関値の高い部分が周辺に見つかった場合には、その変化領域を候補から取り除く。これにより、厳密には小さく移動した物体の境界などが消えてしまう可能性があるが、本研究では、誤検出を減らせる利点の方が大きいと考える。

このような手法によって注目箇所の検出精度が改善された例を図5に示す。左上の四角で囲われた部分は、エッジや細かいテクスチャ部分であるため、従来手法では変化領域として検出されているが、改善手法では同じ映像区間に対して、正しく手や部品の移動部分が検出されている。また、計算時間が半分程度になり、大幅に計算時間が短縮された。

4. 環境カメラとの協調

個人視点のカメラでは、重要な情報を落としている場合がある。例えば、どこで作業を行っているのかが判断しにくい場合や、背後で他人が重要な行動をしている場合もある。このような情報を残すために、室内に取りつけた視野の広い環境カメラを用いて人物の行動を追跡し、個人視点カメラと同期をとって記録する。

得られる画像の例を図6に示す。このような画像に対して、以下のような処理を行う

- 人物(動物体)の検出と追跡
- 人物が停留していた部分の抽出
- 停留部分と、個人視点カメラから検出できる注目シーンの両方を併用した行動記録の作成

基本的には、差分を用いて変化領域(移動物体領域)を検出するが、背景差分のみでは、背景の変化に対応することができず、フレーム間差分のみでは、停止した人物を検出できないという問題がある。そのため、本研究では以下のように、適応的



図 6 環境カメラから撮影された画像

に背景を更新する．

$$B_t(x, y) = g(I_t, B_t)B_{t-1}(x, y) + (1 - g(I_t, B_t))I_t(x, y) \quad (6)$$

ここで、 $B_t(x, y)$ は時刻 t での背景画像、 $I_t(x, y)$ は時刻 t での観測画像、 $g(I_t, B_t)$ は背景画像と観測画像の差から更新の度合いを決める重み付け関数 (ゲイン) である．この g として、変化の小さい箇所に対しては、背景画像の値を随時更新し、変化の大きい部分についてはゆっくりと更新するように関数形を設定する．寝ている場合を除き、人間はある程度動いているので、変化が小さい画素だけを更新することにより、安全に背景を更新できるためである．また、人が動かした物体などは、しばらく時間が経つことによって背景となる

次に、背景差分により得られた移動物体領域を人物として、追跡を行う．ここであげる実験例 (図 6) では魚眼レンズを用いているため、中心から放射方向に延びた小さな移動物体領域を統合する処理等を行い、まとまった人物像が得られるようにする．

5. 実 験

本研究の有効性を評価するために、注目シーンの検出と環境カメラの利用について、簡単な実験例を紹介する．使用した映像は 14 分強 (25709 フレーム) からなるもので、記録された行動は「部屋 1 に入り、自動車のおもちゃを組み立て、部屋 1 を出て廊下をわたり部屋 2 へ入り、論文集を読んでインターネットで検索をし、部屋 2 を出て廊下をわたり部屋 1 へ入り、プロジェクターを組み立て、片付けをする」というものである．

これまで提案した手法で検出された注目シーンを並べた結果を図 8 に示す．この例では 47 個の注目シーンが得られているが、表示のデバイスの都合上、多くの積極的注目シーンが検出された場合には最後の数枚を重ねてある．比較のために、映像中から等間隔に 50 枚の画像を選択した場合を図 9 に示す．これらを比較すると、提案手法では、冗長な部分が省かれ、分かりやすい構造化が実現されていることがわかる．

次に、環境カメラを用いた実験例を図 7 に示す．この例では、人物の軌跡を求め、環境カメラからの画像上にプロットしたものである．この画像を見ることで、各々の人の移動軌跡と、停留していた場所が分かる．また、環境カメラから取得された画像内に置かれている灰色のマークは、その地点で個人視点映像

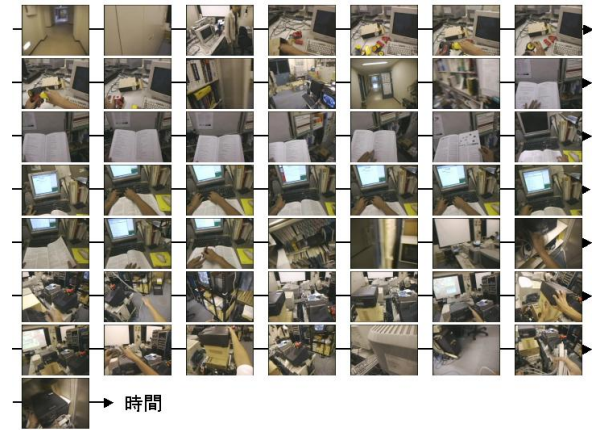


図 9 映像から等間隔に画像を抽出した例

から注目シーンが検出されたことを表している．右側のウィンドウは、各々の注目シーンを表し、これをクリックすれば、どのような作業が行われたかを確認することができる．以上のように、環境カメラからの映像を用いることで、ブラウジングの利便性が大幅に増強されることが分かる．

また、これらの実験結果をユーザに見せたところ、以下のような要求があった．

- 場所 (空間的) を移動するとき、例えば、部屋～廊下、廊下～部屋といった移動がわかるシーンがほしい
- 作業をする際の、その作業に使用する物をとる瞬間 (物のある位置) がわかるシーンがほしい
- 指先での作業の詳細のシーンがほしい
- 作業を行い、その後物が完成したその物が写っているシーンがほしい

これらの要求を満たす処理手法を開発することが今後の課題となっている．

6. ま と め

頭部に装着されたカメラから得られる個人視点映像を処理し、個人行動記録を生成する手法、また、外部の環境カメラと同期して記録することにより、より使いやすい記録を得る方法を提案した．本稿では手法の提案と簡単な実験について報告したが、今後様々な映像を使い、ユーザによる主観評価等の実験を行っていく必要がある．その際に、実験例の最後で述べたような種々の不足面を補うように、手法の改良を行っていく予定である．さらに、個人視点映像と環境カメラ映像を組み合わせた種々の処理について、実現可能な手法を探っていく予定である．

文 献

- [1] 石川陽一, 飯島俊匡, 川嶋稔夫, 青木由直: “エピソード映像の時空間的階層呈示による記憶想起”, 信学技法, PRMU98-186, 1998
- [2] Jebara, T., Schiele, B., Oliver, N., Pentland, A., “DyPERS: Dynamic Personal Enhanced Reality System”, MIT Media Laboratory, Perceptual Computing Technical Report #463
- [3] 石川陽一, 飯島俊匡, 川嶋稔夫, 青木由直, “日常生活空間における視点映像の階層的セグメンテーション”, 信学技報, PRMU****, 1999 .
- [4] 大出 純哉, 中村 裕一, 大田 友一, “ビデオ映像による個人行動

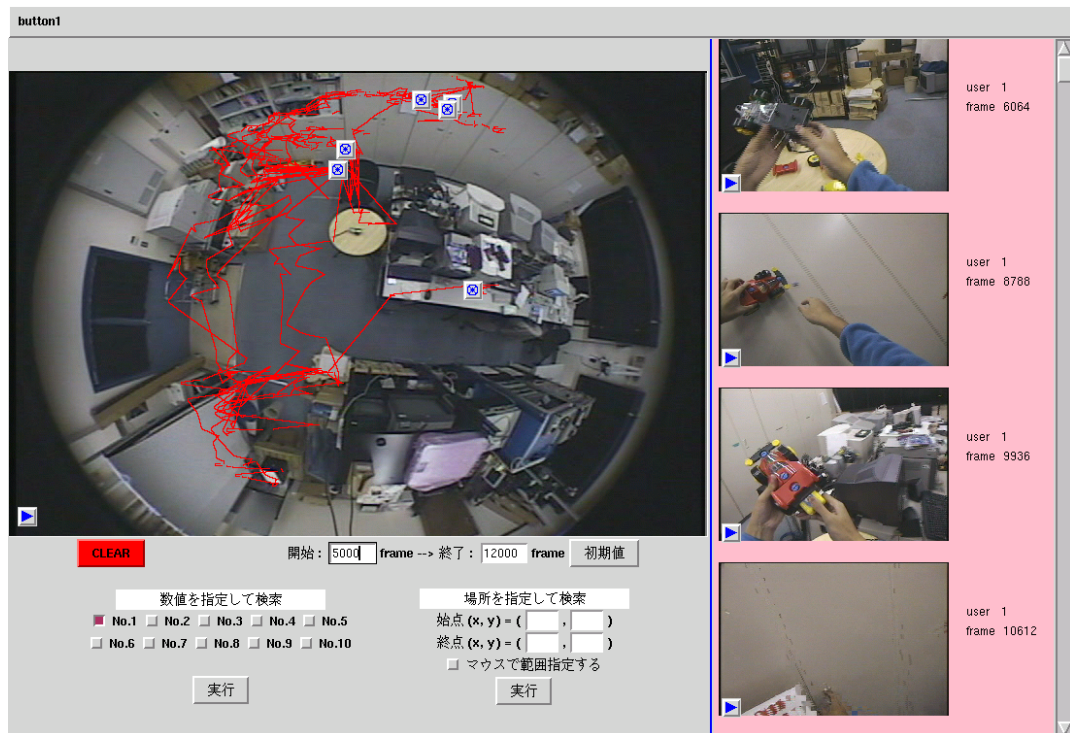


図 7 環境カメラ映像を利用した行動記録ブラウザ

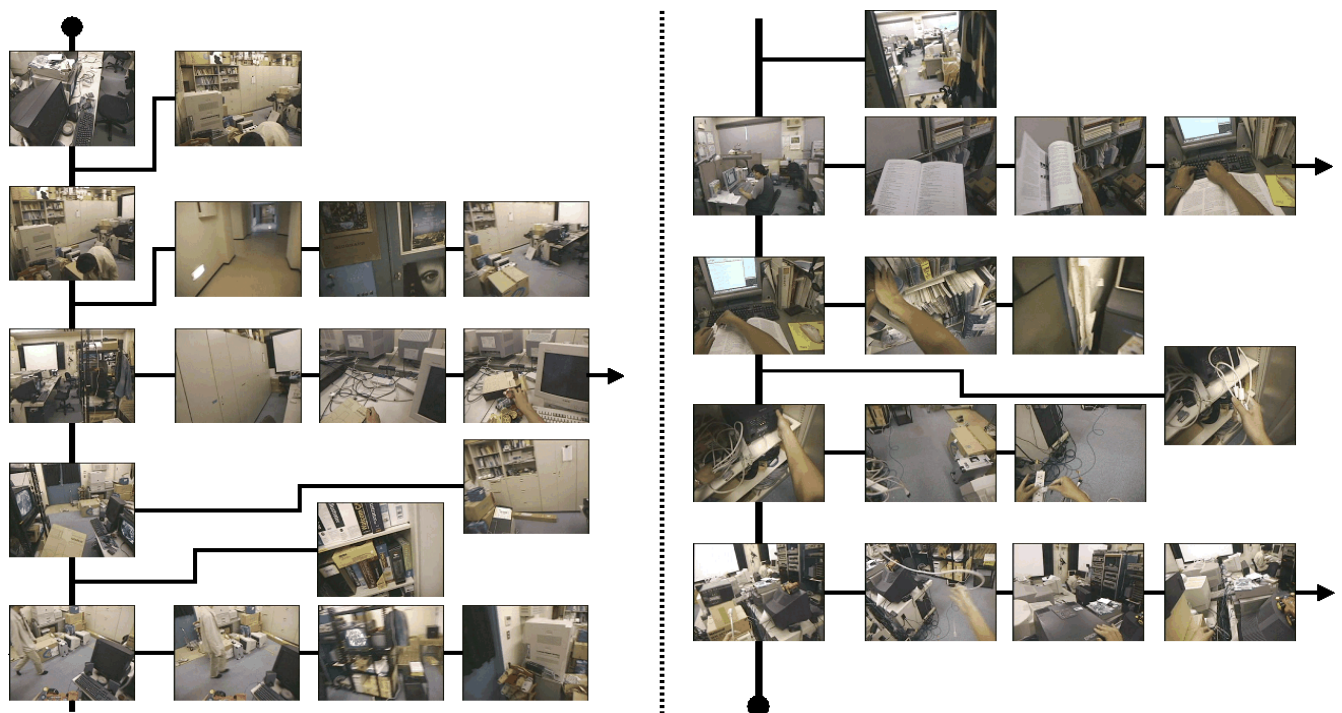


図 8 注目シーンによる映像の要約例: 縦軸に停留シーンを配置し、大まかなシーンの流れを表現する。横にその区間におきた積極的な注目シーンを配置し、その場で起きた出来事を説明する。

- 記録・要約システム -注目シーンとその安定化-”，第 5 回知能情報メディアシンポジウム，pp.83-90，1999．
- [5] 大出 純哉，中村 裕一，大田 友一，“映像による個人行動記録・要約システムとその評価 -注目シーン検出と要約の評価-”，MIRU2000，pp.499-504，2000．
- [6] Bergen,J.，Anandan,P.，and Hanna,K.，“Hierarchical Model-Based Motion Estimation”，Proc．ECCV，pp.237-252，

1992．

- [7] R.Szeliski,J.Shum,“Creating Full View Panoramic Image Mosaics and Environment Maps”，Proc.SIGGRAPH，pp.251-258，1997．
- [8] 久保田 敏司，中村 裕一，大田 友一，“個人行動記録システムにおける注目対象検出 -動き推定モデルの検討-”，電子情報通信学会ソサイエティ大会，2001-9．