

映像インデキシングのための手と把持物体のロバストな認識と追跡

Simple and Robust Tracking of Hands and Objects for Video Indexing

伊藤 雅嗣 尾関 基行 中村 裕一 大田 友一

Masatsugu Itoh, Motoyuki Ozeki, Yuichi Nakamura, Yuichi Ohta

筑波大学 機能工学系

IEMS, University of Tsukuba

E-mail: ozeki@image.esys.tsukuba.ac.jp

Abstract

本稿では、机上作業プレゼンテーションにおける話者の手と把持物体を検出・追跡する手法を提案し、それを基にした映像インデキシングの枠組みについて述べる。本手法では、可視光カメラ・赤外線カメラ・ステレオカメラの3つの画像センサを用いることにより、登場する物体に関する事前知識がなく、話者以外の人物が背後に出入りするという条件下においてもロバストな物体追跡を実現する。さらに物体追跡の結果に人物の動作認識結果を組み合わせることによって作業内容を識別し、映像インデキシングに応用する例を紹介する。

1 はじめに

映像に様々な情報を付加することにより、視聴者の目的に応じた検索や要約、提示などが行える柔軟なマルチメディアコンテンツを実現することができる[1]。マルチメディア技術の発達によって映像が簡単に扱えるようになったことにより、一般企業や教育機関でもこのような映像コンテンツの制作に関心が集まっている。しかし映像コンテンツの制作には、カメラワークや編集についての専門技術が必要とされる上、映像にインデックスを付加するには多大な時間と労力がかかる。このような背景から、撮影の時点からインデックスとして有用な情報を自動的に取得し、様々な目的で利用できる映像コンテンツを簡単に制作できる枠組みが必要とされている。

この問題に対し、我々は、科学実験や機材組み立てなどの机上作業プレゼンテーションを対象とした映像コンテンツ取得システムを構築している。システムの概要を図1に示す。このシステムでは、複数のカメラでプレゼンテーションを自動撮影・編集すると共に、物

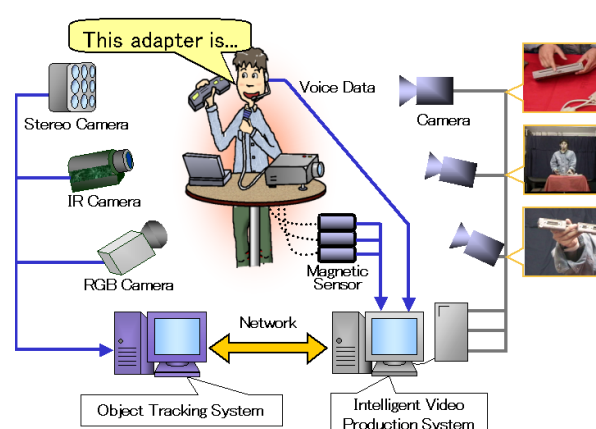


図1 映像コンテンツ取得システム

体の位置やテクスチャ、話者の動作などを検出し、撮影された映像にインデックスとして付与する。映像教材や作業マニュアルにおいて、部品・道具などの物体についての説明やそれらの使い方は重要な情報であるため、物体に関する情報を基に映像を構造化することで、欲しい情報に柔軟にアクセスできる映像閲覧や映像検索が可能となる。

そこで本稿では、複数の画像センサを用いることで、個々の物体についての事前知識を用いることなくロバストに物体を追跡する手法を提案する。また、物体追跡の結果に人物の動作認識結果を組み合わせることによって作業内容を識別し、映像インデキシングに利用する例を紹介する。

2 映像インデキシングへの利用と物体追跡の条件

2.1 映像インデキシングへの利用

例えば、作業などの映像マニュアルを考えた場合、部品や操作、またそれに対する説明が最も重要な情報となる。このような映像マニュアルを利用する際に、重

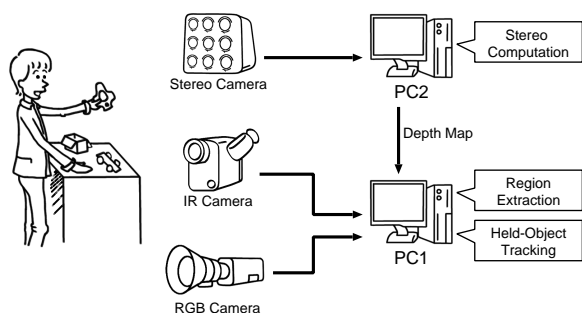


図2 把持物体追跡システムの概要

要な部品や操作が説明されたり、使われたりしている部分に簡単にアクセスすることができれば便利だろう。

このような映像コンテンツを実現するためには、作業に登場する物体について、以下のような情報を自動的に取得し、インデックスとして映像に付加しておく必要がある。

- 物体の位置とテクスチャ
- 物体を使った作業の内容と区間
- 物体への注目度

物体を検索キーとして使用するためには、個々の物体の各時点での位置、それが使って行われた作業の区間、物体が注目されていた区間などの情報が映像をインデックスとして記録しておく必要がある。ここで物体のテクスチャが得られていれば、例えば、映像閲覧の際に物体像のアイコンをクリックすることで、物体が説明されている部分や操作に使われている部分にアクセスすることができる。また作業の内容がインデックスとして記録されていれば、作業内容をキーとして映像を検索することができる。

我々の映像コンテンツ取得システムは、撮影時にこのようなインデックスの取得を行うことを目的としているが、その構成は把持物体追跡システムと知的撮影システムからなる。把持物体追跡システムでは、話者が物体を持っている状態を検出し、その位置とテクスチャを獲得する。知的撮影システムでは、複数のカメラによりプレゼンテーションを撮影し、得られた映像を話者の動作認識を利用して編集する[2][3]。物体追跡の結果と話者の動作認識の結果を合わせることで、その物体を使った作業の内容と区間を知ることができる。

このように全体的なシステムとしては、自動撮影・編集と人物動作の認識も行っているが、本稿では図2に示す把持物体追跡システムについて詳しく説明する。

2.2 物体追跡の条件

本システムでは、以下のような机上作業プレゼンテーションを対象とする。

1. 図1にあるように、話者（作業を説明する人物）が物体を前に掲げたり指さしたりする。

2. 物体を参照しながら、その名前や使い方などについて説明する。
3. 紹介した物体（部品や道具）を使って、組み立てや分解などの作業を行う。

このような場面では、物体は回転や変形などによって常にその外観を変化させる。また料理番組などでよく見られるように、シーン内には話者以外にもアシスタントが近傍にることが多い。よって本研究では、以下のような環境条件の下での物体追跡に取り組む。

- 物体の大きさ、色、形状などに関する予備知識はない。種々の物体が形を変えながら出現するため、すべての物体に関する知識をあらかじめ与えておくことは難しい。
- 作業中に複数の人物が登場したり、物体の位置や姿勢が変化するため、背景は常に変化する。

ただし机上作業シーンであることから、次の前提条件も仮定できる。

- 重要な物体¹は、話者の手によって動かされたり操作されたりする。
- 机の上が作業空間であるため、物体が存在する空間がわかっている。

これら2つの前提条件を加えても、上に述べた2つの環境条件下における物体追跡は簡単ではない。この問題に対し本研究では、3種類の画像センサ（可視光カメラ・赤外線カメラ・ステレオカメラ）を相互補完的に用いることで、このように厳しい条件の下でもロバストな把持物体追跡を行う手法を提案する。

3 複数の画像センサの利用

本研究では、物体像を物体のモデルと直接照合するのではなく、ある特定の作業空間中で手と共に移動する領域を検出し、手の領域と分離することで把持物体を認識する。そのために、まず3種類の画像センサから、それぞれ以下の要素領域を抽出する。

肌色領域：肌色の領域。可視光カメラから得られる画像より抽出する。

動領域：動いている領域。可視光カメラから得られる画像より抽出する。

肌温領域：人間の肌の表面温度以上の温度を持つ領域。赤外線カメラから得られる画像より抽出する。

特定距離領域：作業が行われると想定される空間（机の上の空間）にある物体の領域。ステレオカメラから得られる距離画像より抽出する。

¹インデックスが付けられるべき物体。

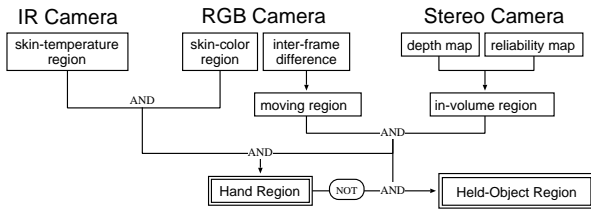


図3 各領域の抽出と論理和の流れ

要素領域の論理積を以下のようにとることで、手領域と把持物体領域を検出することができる。

$$\begin{aligned} \text{手領域} = & \text{特定距離領域} \wedge \text{動領域} \\ & \wedge \text{肌温領域} \wedge \text{肌色領域} \quad (1) \end{aligned}$$

$$\begin{aligned} \text{把持物体領域} = & \text{特定距離領域} \wedge \text{動領域} \\ & \wedge \neg \text{手領域} \quad (2) \end{aligned}$$

ただし、一部の“ \wedge ”は厳密な論理積を表さない。これについては4.3節で説明する。

処理の概要を図3に示す。この手法を用いることにより、理想的には、背景中に他の人物が存在する場合や物体が動いている場合でも、誤検出することなく把持物体のみを追跡することが可能となる。

以下、それぞれの要素領域の抽出方法について説明する。

肌色領域の抽出

肌色はRGB色空間において rg 平面にまとまりのある分布を示すことが知られている。そこで rg 色度平面上に肌色モデルを作り、肌色領域を抽出するのに用いる[4]。ただし近接物体からの反射光や照明の当たり具合により、実際に観測される色はかなり変化する。そのため本研究では厳密なモデルを構築するのではなく、粗いモデルで近似し、誤検出は多少許しながら検出もれが少なくなることを目指す。

腕と手から抽出した多数のサンプルを基に肌色領域の分布を求め、そこから計算した平均値と共分散行列 Σ を以下に示す。

$$\begin{aligned} \text{mean}(\bar{r}, \bar{g}) &= (0.437773, 0.334845) \\ \Sigma &= \begin{pmatrix} 0.003915 & -0.000230 \\ -0.000230 & 0.000935 \end{pmatrix} \end{aligned}$$

これを基に、肌色領域を識別するためのマハラノビス距離 $D^2(r, g)$ を以下のように定めた。

$$D^2(r, g) = \begin{pmatrix} r - \bar{r} \\ g - \bar{g} \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} r - \bar{r} \\ g - \bar{g} \end{pmatrix}$$

図4の左側のグラフは、可視光カメラから得られた画像のマハラノビス距離 $D^2(r, g)$ の各値における手領域と背景領域の画素分布を示している。他の要素領域

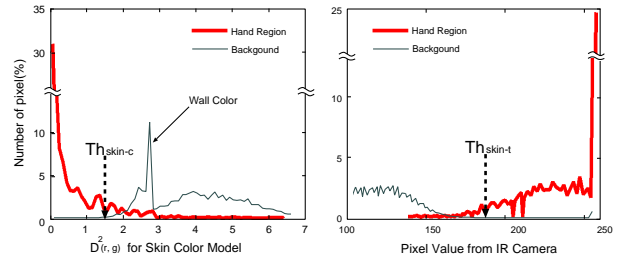


図4 典型的な机上作業プレゼンテーションにおける、肌色領域と背景領域のマハラノビス距離による画素の分布（左）と肌温領域と背景領域の温度による画素の分布（右）

との論理積をとることから、肌色領域ができるだけ多く検出される閾値を選ぶことが有効であると予想される。ただし我々の実験環境では肌色に近い色合いの壁紙を用いているため、その影響を避ける必要があり、図4のような閾値 $Th_{\text{skin-c}}$ を決定した。

動領域の抽出

動領域は、可視光カメラから得られた画像のフレーム間差分を用いて抽出する。差分は4フレーム間隔で行い、閾値は経験的に求めた。

肌温領域の抽出

本研究で使用する赤外線カメラ²は検出波長が $8 \sim 14 \mu\text{m}$ であり、人体が出す赤外線の $7 \sim 10 \mu\text{m}$ に対して良い感度を持つ。そこで赤外線カメラ画像に閾値処理を行うことによって、人体の肌表面温度に相当する画素を抽出し、得られた部分を手や顔の候補領域とする。

図4の右側のグラフは、赤外線カメラから得られた肌温領域と背景領域の画素分布を示す。他の要素領域との論理積をとることから候補領域は多めに抽出することが好ましいため、肌色の領域がほとんど抽出されるよう閾値 $Th_{\text{skin-t}}$ を決定した。

特定距離領域の抽出

机上作業ではその作業のほとんどが作業机の上の空間で行われるため、事前に手や物体が現れる空間はわかっているものと仮定する。本稿の実験では、特定距離領域の幅・奥行きをそれぞれ作業機の幅と奥行き³に、特定距離領域の高さを作業機の天板上から話者の頭までの高さとした。ステレオカメラは作業空間に正対して配置されているため、特定距離領域は単純な閾値処理によって抽出できる。

ステレオカメラから得られる奥行き画像には、一様な背景や繰り返しパターンに起因するノイズが含まれる。これに対し、本システムで使用しているステレオ

²日本アビオニクス IR-30。

³作業台の手前の端に意味なく手を置くことが多いため、厳密には、作業機の奥行き幅よりもステレオカメラからみてやや浅めに特定距離領域の奥行きを設定している。

カメラ⁴では、奥行き画像がどのくらい信頼できるかを画素単位で示した信頼度画像 [5] が得られるため、信頼度の高い画素のみ使用することでこれらのノイズの問題を解決する。この処理は図 2 の PC2 で行っている。

4 手と把持物体領域の抽出

4.1 位置の補正

各画像センサにレンズ歪みがあることと、3つのセンサの光軸（視点）を完全に一致させるのは難しいことから、各センサから得られる画像の間で位置の補正を行う必要がある。本研究では、可視光カメラから得られた画像を基準として、赤外線カメラとステレオカメラから得られた画像を変換する。補正には、レンズ歪みを補正するために 2 次の幾何補正を、光軸の不一致を補正するために 2 次元射影変換を用いる。ここで、位置合わせに 2 次元射影変換を用いるのは、各画像センサから見て作業空間の奥行き範囲が比較的狭いためである。実際の計算には、2 次幾何補正と 2 次元射影変換の両方の機能を兼ねる 3×5 の変換行列を用いる。

キャリブレーションにはキャリブレーションボードを用いるが、作業空間の中心（作業機の中心）に置いたときに各画像センサの視野全体がボードで埋まる大きさのものを用意し、その上に 25 点の特徴点を配置した。これらの特徴点を各センサから得られる画像からそれぞれ抽出し、赤外線カメラとステレオカメラについて、以下の補正処理の計算式における行列 $M_{3 \times 5}$ を求める。

$$\begin{pmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_n \\ 1 & \dots & 1 \end{pmatrix} = M_{3 \times 5} \begin{pmatrix} u_1 & \dots & u_n \\ u_1^2 & \dots & u_n^2 \\ v_1 & \dots & v_n \\ v_1^2 & \dots & v_n^2 \\ 1 & \dots & 1 \end{pmatrix} \quad (3)$$

ここで、 $(x_1, y_1) \sim (x_n, y_n)$ は可視光カメラから得られる画像上の特徴点座標、 $(u_1, v_1) \sim (u_n, v_n)$ は補正前の赤外線カメラまたはステレオカメラから得られる画像上の特徴点座標である。行列は最小二乗法を用いて計算する。

4.2 同期

図 2 に示した通り、可視光カメラと赤外線カメラからそれぞれ得られる画像は PC1 で、ステレオカメラから得られる奥行き画像は PC2 でキャプチャする。PC2 でキャプチャされた奥行き画像はネットワークを介して PC1 に伝送され、領域の抽出および手と把持物体の検出・追跡は PC1 ですべて行う。

この際、奥行き画像の取得や PC2 から PC1 へのデータ伝送による遅延が生じるため、画像間で同期をとる

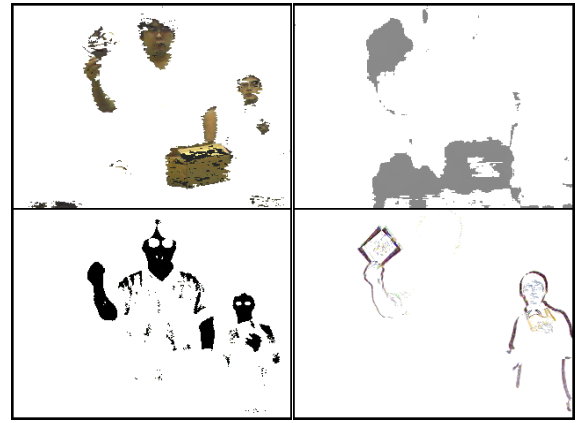


図 5 各画像センサから得られる領域（左上：肌色領域，左下：肌温領域，右上：特定距離領域，右下：動領域）

必要がある。このため本システムでは、各画像のキャプチャ時刻を画像と共に記録し、最も近い時刻に取得された画像の組を用いて領域抽出処理を行う⁵。

4.3 手・把持物体の検出

手領域の検出については、まず 3 章で挙げた式 (1) に示した論理積をとることによって候補領域を抽出する。次に、候補領域に対して膨張収縮処理を行い、細かいノイズを削減する。最後に、残った候補領域の面積をラベリングにより計算し、閾値 Th_{hand} よりも大きい領域を 2 つまで手領域として検出する。ここで閾値 Th_{hand} は、我々の撮影環境における手の大きさを基に計算した。

把持物体領域についても、まず式 (2) に示した論理積をとることによって候補領域を抽出する。ただし式 (2) の最初の “^” は単なる論理積処理ではなく、まず特定距離領域をラベリングによって各領域に分離し、各領域の中の動領域と判定された画素の割合（画素数の比）を計算する。手領域ではなく、かつ、この割合が閾値よりも大きくなる小領域を把持物体の候補領域とする。抽出された領域候補は膨張収縮処理によりノイズを削減し、残った候補領域の面積が閾値 Th_{obj} よりも大きなものを 2 つまで把持物体領域として検出する。ここで閾値 Th_{obj} は、我々の撮影環境において、検出した物体の中で最小のものを基に計算した。

手や把持物体の位置は、検出された領域の重心を計算することによって求める。この重心計算をフレーム毎に繰り返すことで、手と把持物体を追跡する。さらに求めた位置をカルマンフィルタを用いて平滑化することで、滑らかな追跡を行う。

各画像センサから抽出された領域を図 5 に、追跡結果の例を図 6 に示す。図 5 からわかるように、肌色の領域を検出するだけでは、肌色の箱や背後の人物など手以外の領域も検出されてしまう。温度情報と奥行き

⁴コマツ高速ステレオビジョン FZ930。

⁵PC1 と PC2 の時刻は NTP を用いて合わせている。



図 6 追跡例（左：話者のみ，右：話者に加え，机上に静止物体があり背景に人物が動いている）

表 1 作業内容の分類

作業内容	数の変化	位置関係	話者の動作
提示	1	-	指示・提示
分離	1 → 2	離れていく	例示・操作
組み付け	2 → 1	一緒になる	例示・操作

情報をさらに組み合わせることで，誤検出された領域を候補から外し，手と把持物体のみを正確に検出できることがわかる．

5 把持物体追跡による作業内容の検出

2.1 節で述べたように，我々の映像コンテンツ取得システムは，把持物体追跡システムと知的撮影システムにより構成されている．知的撮影システムでは指示・提示動作と操作・例示動作を認識を行っているが，把持物体の有無がわからないと「指示と提示」「操作と例示⁶」を区別することができない．また，把持物体の数や位置関係がわからないと操作の内容を知ることができない．そこで，これら 2 つのシステムから得られる情報を組み合わせることにより，机上作業プレゼンテーションにおける典型的な作業を検出する．

各システムから得られる情報を以下に示す．

把持物体追跡システム：物体の位置，物体のテクスチャ，物体が把持されている期間，把持物体間の距離

知的撮影システム：作業を説明する動作（指示・提示，例示・操作）の検出結果，発話内容，多視点映像

上記の情報の中から「（把持物体の）数の変化」「（把持物体の）位置関係」「話者の動作」の状態を常に監視する．これら 3 つの状態の組み合わせによって，3 つの作業内容「提示」「分離」「組み付け」の 3 つの作業内容を検出・識別する．

作業内容の分類と 3 つの状態の関係を表 1 に示す．話者の動作の「指示・提示」とは「手を前に伸ばして，指示もしくは物体の提示を行う動作」，「例示・操作」とは「机の上に手を延ばして，例示もしくは物体の操作を行う動作」である．知的撮影システムで検出された

⁶ ここでいう例示とは，たとえば「このような形の」と言いながら両手で形を作るなど，実際の物は使わずに例で示すような動作のこと．

表 2 追跡結果（単位はフレーム数）

	把持物体のみ	複数物体・背景人物
全フレーム数	1350	1350
正解数	1316 (97.5%)	1259 (93.3%)
検出もれ数	30 (2.2%)	11 (0.8%)
誤検出数	4 (0.3%)	80 (5.9%)

動作認識の結果はネットワークを通して把持物体追跡システムに送信され，把持物体追跡システムで作業内容の検出処理を行っている．

6 実験

6.1 把持物体の追跡精度

把持物体の追跡精度を調べるために，(a) 机の上に 1 つだけ物体が存在し，背景中に人物が存在しないシーンと，(b) 机の上に複数の物体が存在し，背景中で人物が動いているシーンを用意した．シーンの例を図 6 に示す．

実験では，3 人の被験者に同時に 1 つもしくは 2 つの物体を掴んで自由に動かしてもらった．各シーンについて 15 秒の映像（450 フレーム）を評価に使用し，フレーム毎に正解・検出もれ・誤検出の 3 通りに分けてそれぞれの数を計算した．ここで，「検出もれ数」とは把持物体が検出されなかった場合，「誤検出数」とは把持物体以外の領域が抽出された場合を表す．

結果を表 2 に示す．結果より，背景で人物が動いており，作業空間中に似たような物体が存在する複雑な環境でも，精度良く把持物体のみを追跡できていることがわかる．個々の物体に関する事前知識を全く与えていないことを考慮すると，十分に良い結果であるといえる．

6.2 作業内容の検出精度

5 章で述べた作業内容の検出について，その検出率を調べた．今回の実験では，机の上にある 4 つの物体の中から適当に 1 つもしくは 2 つの物体を選び，それらを用いて「提示」「分離」「組み合わせ」の作業を各 80 回ずつ行った．各作業について，正解数・検出もれ数・誤検出数を数え，作業総数に対する割合を計算した．ここで，「検出もれ数」とは作業が検出されなかった場合，「誤検出数」とは検出された作業の種類が実際に行われたものと違う場合を表す．

結果を表 3 に示す．どの作業についても 7～8 割の正解率となった．把持物体の検出もれに起因する作業の検出もれは 1～2 割ほどあるものの，作業の誤検出はほとんどなかった．実際に検出された組み付け作業の例を図 7 に示す．左の図から順に，組み付け前の物体が 2 つある状態，組み付けている途中の状態，組み付けが終了し物体が 1 つとなった状態を表す．

表 3 作業内容検出結果

	提示	分離	組み付け
作業総数	80	80	80
正解数	70 (87.5%)	61 (76.3%)	68 (85.0%)
検出もれ数	10 (12.5%)	19 (23.8%)	11 (13.8%)
誤検出数	0 (0.0%)	0 (0.0%)	1 (1.3%)

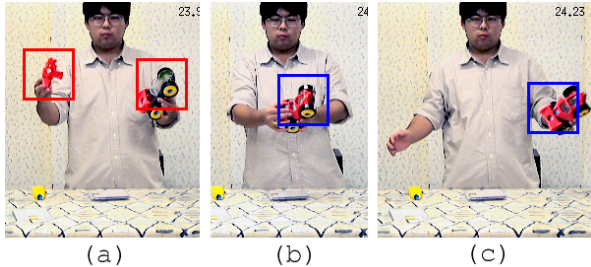


図 7 組み付け作業の例

6.3 アプリケーション例

ある物体が説明されている部分を作業マニュアル映像から検索したいとき、その物体像のアイコンをクリックすることで該当部分が再生されれば便利である。これを実現するためには、話者が作業を行っている間の映像とその物体のテクスチャを関連づけて記録すればよい [6]。

処理の流れを図 8 示す。把持物体追跡システムでは、物体を把持した時刻、置いた時刻とその時の物体テクスチャを獲得し、同時に話者の動作開始時刻と終了時刻を知的撮影システムから受け取ってインデックスとして記録する。撮影終了後、それらのインデックスと撮影された映像を用いて、図 9 のようなビューアを作成する。右のウィンドウに机上作業に登場した物体像のアイコンが並んでおり、これらのどれか一つを選んでクリックすると、それに対する説明が簡単に閲覧できる。

7 まとめ

複数の画像センサを相互補間的に用いることにより、机上作業プレゼンテーションにおける手と把持物体を検出・追跡する手法を提案した。また、把持物体追跡システムと知的撮影システムを組み合わせることによって、作業内容（提示・分離・組み立て）を検出する手法を示した。さらに映像インデキシングの応用例として、物体のアイコンを用いた映像ビューアを作成した。実験で示したように、物体に関する事前知識がなく、背景中に人物が動き回っているという環境条件の下でも、把持物体を追跡できる手法であることを示した。

しかし、手のひらより小さな物体の追跡や、撮影・追跡範囲の拡大、システムの小型化など解決すべき課題も多く残っており、それらは今後の課題となっている。

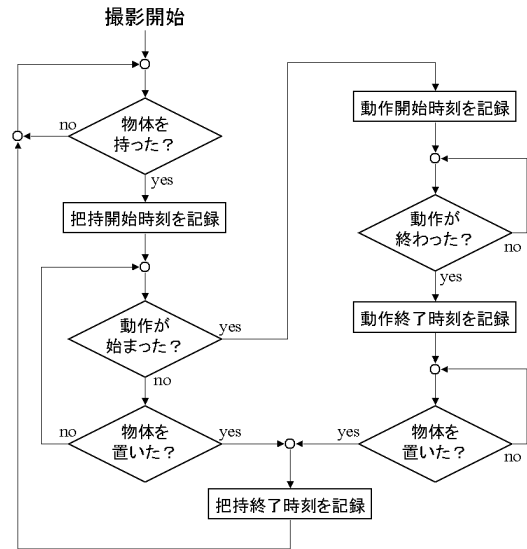


図 8 インデキシング処理の流れ（動作検出の結果は知的撮影システムより得られる）

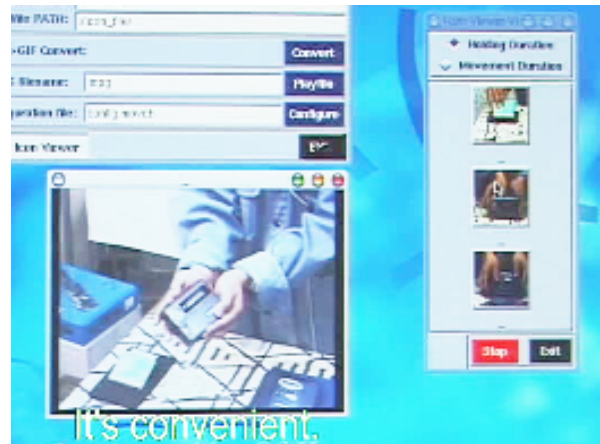


図 9 物体像のアイコンによる映像ビューア

参考文献

- [1] Izuno, H., Nakamura, Y. and Ohta, Y.: QUEVICO: A Framework for Video-based Interactive Media, *Proc. Int'l Workshop on Intelligent Media Technology for Communicative Reality*, pp. 6–11 (2002).
- [2] Ozeki, M., Nakamura, Y. and Ohta, Y.: Camerawork for Intelligent Video Production – Capturing Desktop Manipulations, *Proc. ICME*, pp. 41–44 (2001).
- [3] Ozeki, M., Nakamura, Y. and Ohta, Y.: Human Behavior Recognition for an Intelligent Video Production System, *Proc. PCM*, pp. 1153–1160 (2002).
- [4] 近藤博仁, 孟洋, 佐藤真一, 坂内正夫: テロップ認識と顔照合を統合したニュース映像中人物の自動索引付けシステム, 電子情報通信学会 総合大会, Vol. D-12-190 (1999).
- [5] <http://www7.airnet.ne.jp/komatsu/Stereo/stereo.j/>
- [6] Ozeki, M., Itoh, M., Nakamura, Y. and Ohta, Y.: Tracking Hands and Objects for an Intelligent Video Production System, *Proc. ICPR*, pp. 1011–1014 (2002).