社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

映像メディア取得のための手と把持物体の追跡と認識

ー 多種類の画像センサによるロバストな実時間追跡ー

伊藤　雅嗣† 　尾関　基行† 　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: †{itom,ozeki,yuichi,ohta}@image.esys.tsukuba.ac.jp

あらまし　プレゼンテーションにおいて話者が物体を持っている場合，その把持物体が話題の中心になっていることが多い．本研究では複数の異なる種類の画像センサ（可視光カメラ・赤外カメラ・ステレオカメラ）を相互補完的に用いることで手と把持物体の分離認識・追跡を行う手法を提案する．また，発話と動作情報を用いて，話者が物体に対する説明を与えている状況を検出し，撮影されている映像，時刻，画像中での物体の位置と関連付けて蓄積することにより，高次利用可能な映像データを取得する手法を提案する．

キーワード　赤外線カメラ，ステレオカメラ，把持物体追跡，実時間追跡，映像インデキシング

Hand and Object Tracking System for Producing Video-based Multimedia

ー Robust and realtime tracking by multiple vision sensors

Masatsugu ITOH†, Motoyuki OZEKI†, Yuichi NAKAMURA†, and Yuichi OHTA†

† IEMS,University of Tsukuba　Tennoudai 1–1–1, Tsukuba, Ibaraki, 305–8573, Japan
E-mail: †{itom,ozeki,yuichi,ohta}@image.esys.tsukuba.ac.jp

Abstract　In presentations, an object held by a hand is often the focus of attention. We propose a novel method for detecting and tracking a hand-held object by integrating three diferent types of vision sensors (RGB camera, infrared camera, stereo camera). Based on this method, we developed a support system for producing integrated multimedia. In this system, the system detects the situation that a speaker is explaining an object, and the system records and relates object's figure and annotations concerning the object. Thus, we obtained a video-based multimedia that is indexed by important objects and annotations related to them.

Key words　IR camera, Stereo camera, hand and object tracking, realtime tracking, video indexing

1. はじめに

近年のディジタル技術の向上により，映像や映像を基にしたマルチメディアコンテンテンツが様々な用途に利用されるようになってきた．映像に様々な情報を付加して記録・蓄積し，キーワードや画像による検索を可能にするための研究が数多く行われている．また，コンテンツ不足の問題が指摘されており，既存の映像を再利用するだけではなく，種々の用途に用いるための映像コンテンツを簡単に取得することも重要な研究課題となっている．

このような背景から，我々は図 1 に示すような，撮影の時点から映像と共にその内容に関する情報を自動的に取得することのできる知的撮影システムの構築を行っている [4]．このシステムは，多大なコストをかけることが難しい組織での映像制作や個々の組織での多様なマルチメディアコンテンツ作成を補助することを目的としている．現在，このシステムでコンテンツとして想定しているのは実験教材や作業マニュアルであり，一人または少数の人間が作業を実演したり，機材の説明を行ったりする場面を対象としている．

このような場面で特に重要な意味を持つのは，その中に出てくる物体に対する説明や，それらの物体になされる操作である．そのため，物体やそれらに関する注釈を基に映像を構造化し，欲しい情報に柔軟にアクセスすることができる映像閲覧や映像検索を可能にすることが望まれる．

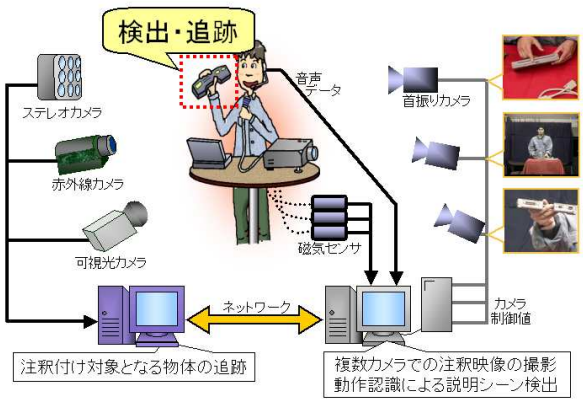

図 1　システム全体の概要図

図 2　プレゼンテーション例

本研究では，このような目的のために，作業中に話者が物体を持っている場合に物体が話題の中心になっていることが多いことに着目し，プレゼンテーションシーンから把持物体を検出・追跡する手法について研究を行ってきた．

本稿では物体に関する事前知識をできるだけ用いずに実時間でロバストな検出と追跡を行うために，複数の画像センサを用いる手法を提案する．また，話者の動作を認識し，物体に関する説明が与えられたことを検出することによって，物体の外観，位置，付加された注釈情報などを関連付けて映像の付加データとする手法を提案する．

2.　把持物体追跡による映像インデキシング

我々が対象とする場面は，図 2 に示すようなプレゼンテーションである．この例は，ノートパソコンに付属のアダプターを装着する手順を想定したものであるが，一人の話者が物体とその操作方法を順を追って説明している．その最中に，しばしば物体に関する説明が与えられるが，シーン中には種々の物体が存在し，物体は把持されたり，組み立てられたり，また，隠蔽によって見えかくれしたり，消滅したりする．また，料理番組などで良く見られるように，プレゼンテーションの協力者や聞き役が近傍にいる場合も想定する．

このようなシーンを撮影し，得られた映像を教材や作業マニュアルとして利用することを考えた場合，物体がどのように動き，その物体にいつ，どのような情報が関連付けられたかといったことを関連付けて記録し，必要に応じてそれを提示することが重要となる．本稿ではそのための手法を提案する．

2.1　把持物体追跡

把持物体追跡の条件として，本研究では以下のような前提条件を設定した．

- 複数の物体，複数の人が存在する．
- 物体の大きさ，色，形状等に関する予備知識はない．上述したように，種々の物体が形を変えながら出現するため，登場するすべての物体に関する知識 (テクスチャや大きさなど) をあらかじめ与えておくことは難しい．
- 作業中は背景が常に変化する．
  作業中は複数の人物が登場し，行き来することもあるので背景は常に変化するものとして考える．

しかし，机上作業シーンを対象としていることから，次のことを前提条件として仮定する．

- 重要な物体の多くは人間によって把持され，移動される．
- 作業に関わる物体の存在する範囲が，作業空間として予めわかっている．

以上の考え方から，本研究では，物体像を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，それを基に把持物体を認識する方法をとる．この際に，通常の可視光 (RGB) カメラから得られる色情報，動き情報だけで手と把持物体を分離認識し追跡するのは難しいため，手領域を検出するために赤外線カメラを利用する．また，様々な物体が把持物体になる可能性があることから，特定の作業空間内にあることを検出するために磁気センサなどの接触型センサを用いることは現実的ではない．そのために，ステレオカメラを利用する．本研究では，このように，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によって，手と把持物体の分離認識・追跡を行う．

2.2　映像インデキシングへの利用

把持物体の検出・追跡を行うことによって，現れた物体やその位置による映像インデキシングが可能となる．さらに，プレゼンテーションを行っている話者の行動を認識し，物体に対して説明が加えられている状況が検出できれば，物体に対する注釈情報を獲得することもできる．これらを関係付けて蓄積しておけば，欲しい情報に簡単にアクセスすることのできる映像メディアが得られる．

例えば，機材の操作を学習している最中に，ある部品がどのように使われるのかを知りたいときがよくある．このような際に，部品の名前を入力したり，部品の外観を持ったアイコンをクリックすることにより，その部品に対する説明や，その組み立て状況を示した映像が再生されれば，マニュアルとして使いやすいものとなる．さらには，部品をカメラの前に差し出すだけで，該当する映像が検索されるようなシステムの要素技術となる(注1)．

このような目的のために，本システムでは，作業に登場する物体について，以下に挙げる情報を自動的に取得し，作業シーン映像と関連づけて保存する．

(注 1): 現在はシステム構築上の問題のため，まだ実装されていない．

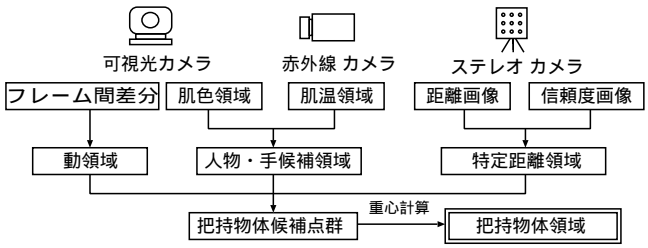

図 3　把持物体領域抽出の概要

（1） 物体の 3 次元位置
（2） 物体への注目度（把持されているか否か）
（3） 物体についての説明

本論文では，まず前者の 2 つについて，その取得システムおよび取得手法を詳細に述べ，3 つ目については，我々のグループで構築してきたプレゼンテーションの知的撮影システムと組み合わせる方法を紹介する．

3. 多種類の画像センサの利用

3.1 処理の概要

可視光カメラ，赤外線カメラ，ステレオカメラの 3 つの画像センサにより，肌色領域，肌温領域，特定距離領域，動物体領域を抽出し，これらを組み合わせて手と把持物体領域の抽出を行う．ここで，肌色領域とは人間の肌の色を持つ領域，肌温領域とは人間の肌の表面温度以上の温度を持つ領域，特定距離領域とは想定している作業空間にある物体の領域である．

まず，前節での議論に基づき，各センサから得られた情報から手領域，把持物体領域として抽出する条件を以下のように定める．

- 手領域は「肌色領域 ∧ 肌温領域 ∧ 特定距離領域 ∧ 動領域」とする．
- 把持物体領域は「特定距離領域 ∧ 動領域 ∧ 手以外の領域」とする．

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．

使用する各画像センサの詳細を以下に示す．

3.2 可視光カメラ（肌色領域，動領域の抽出）

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．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．本研究では複数人物の腕や手などから抽出した多くのサンプルを基に肌色領域の分布を求め，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．

$$D^2(r,g) = (r-0.416)^2 \times 720 + (g-0.339)^2 \times 3255$$

ただし ,$r = R/(R+G+B), g = G/(R+G+B)$

本研究では各画素に対して，上式の値を閾値処理することにより肌色領域を抽出する．対象とするシーンの画像例を図 4 に，

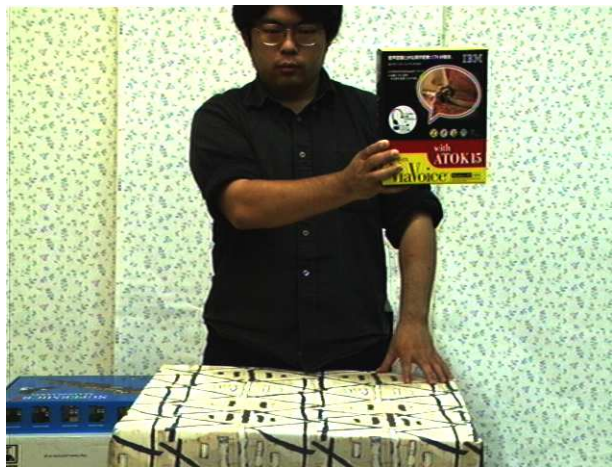

図 4　可視光カメラ画像

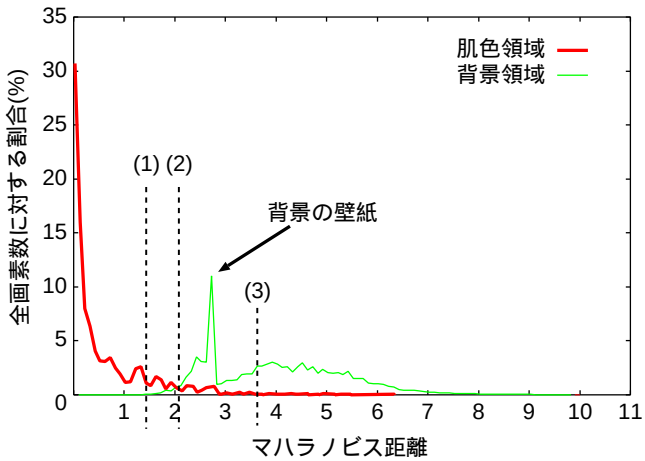

図 5　マハラノビス距離による画素分布

肌色領域と肌色領域以外の背景領域それぞれのマハラノビス距離による画素数の分布を図 5 に示す．ここで，閾値を決める基準として以下の 3 つを検討した．

(1) 背景領域（ノイズ領域）がほとんど抽出されないような閾値（肌色領域検出の精度 P(注2) を重視した閾値）．
(2) 最も誤識別率が小さくなるような閾値
(3) 肌色領域がほとんど抽出されるような閾値（肌色領域検出の再現率 R(注3) を重視した閾値）．

手領域等を抽出する際に論理積をとるため，個々の領域は多めに抽出される方が好ましいと考えられる．そのため，(3) の閾値を試みたが，室内には肌色に近い色を持つ領域が多くあることから，それらが大きな雑音となった．このことから，結果的に (2) の閾値が良い性質を示した．以上の方法によって抽出された肌色領域を図 6 に示す．

動領域の抽出にはフレーム間差分を用いる．得られる動領域画像の例を図 7 に示す. 差分は 1 フレーム単位で行い，閾値は経験的に求めた値を用いている．

図 6　肌色領域画像

図 7　動領域画像

図 8　赤外線カメラ画像

3.3　赤外線カメラ (肌温領域の抽出)

本研究で使用する赤外線カメラ(注4)は検出波長が 7 ~ 14μm であり，人体が出す赤外線の 7 ~ 10μm に対して良い感度を持つ．

(注2): 精度 P= 正検出数/検出数

(注3): 再現率 R=正検出数/ 正解数

(注4): 日本アビオニクス IR-30

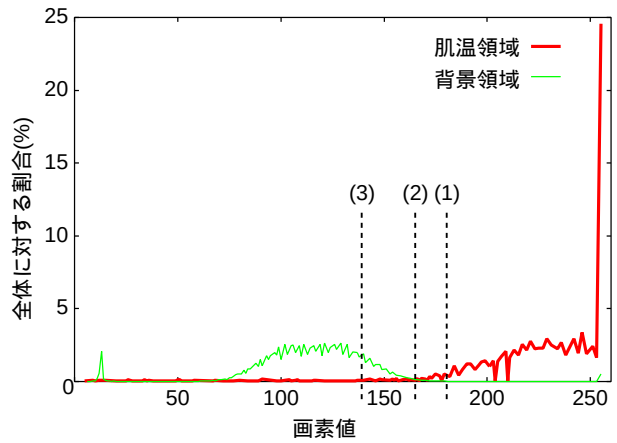

図 9　画素値による画素数分布

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領域の候補とする．

赤外線カメラの画像例を図 8 に示す．図 9 に赤外線カメラから得られた肌温領域とそれ以外の背景領域の画素の分布を示す．閾値としては，肌色領域と同様に以下の (1) ~ (3) が考えられる．

(1)　背景領域 (ノイズ領域) がほとんど抽出されないような閾値 (肌色領域検出の精度 P を重視した閾値)．

(2)　最も誤識別率が小さくなるような閾値

(3)　肌色領域がほとんど抽出されるような閾値 (肌色領域検出の再現率 R を重視した閾値)．

論理積をとることから，候補領域を多めに抽出することが好ましいため，(3) の閾値を選択し，実験でも良い結果を示した．

3.4　ステレオカメラ (特定距離領域の抽出)

前述したように，事前に作業空間の位置がわかっているものとする．また，ステレオカメラは固定されているため，得られる奥行き画像の画素値から，簡単に特定距離領域に入っているかどうかを判定できる．

本研究で使用するステレオカメラ(注5)から得られる距離画像例を図 10 に示す．現在の実験環境では，ステレオカメラが作業空間に正対しているため，特定距離領域を単に閾値処理によって抽出している．

3.5　手領域，把持物体領域の抽出

3.1 節で述べたように，論理積により各領域の抽出をする．ここでは単純に論理積を計算し，その他の処理は行っていない．

可視光カメラからの肌色領域と赤外線カメラからの肌温領域の論理積を取ったものを図 11 に示す．さらに図 11 とステレオカメラからの特定距離領域の論理積を取ることで抽出した手領域を図 12 に示す．

肌色領域，肌温領域，特定距離領域，動領域を用いて抽出した把持物体領域を図 13 に示す．

(注5): コマツ製 高速ステレオビジョン FZ930

図 10　ステレオカメラ画像

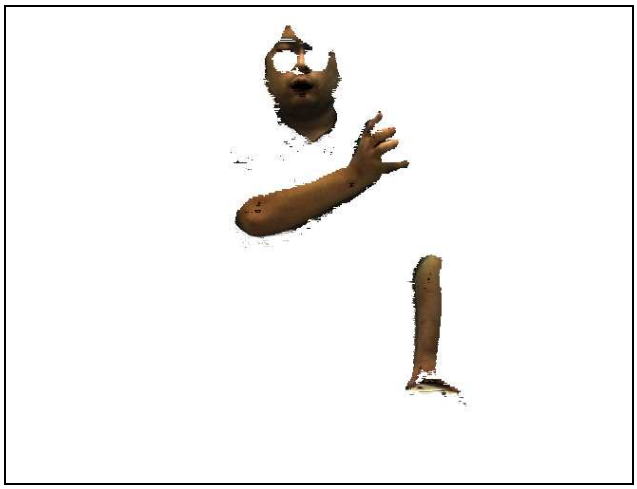

図 11　肌色 ∧ 肌温領域

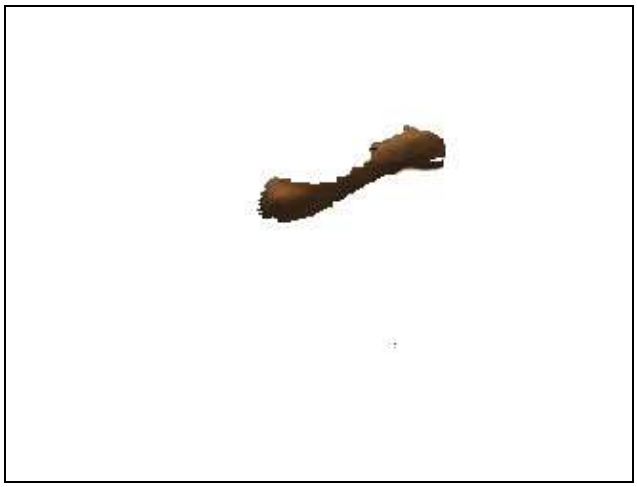

図 12　肌色 ∧ 肌温 ∧ 特定距離領域

4.　把持物体追跡システム

4.1　システム構成

本システムのハードウェア構成を図 14 と表 1 に示す．PC 間の同期のためには NTP を用い，定期的にタイムサーバーに問い合わせることによって時刻合せを行う．また，データ転送の遅延が生じるために，各画像の取得時刻を各画像と合わせて記

図 13　肌色+肌温 ∧ 特定距離 ∧ 動領域

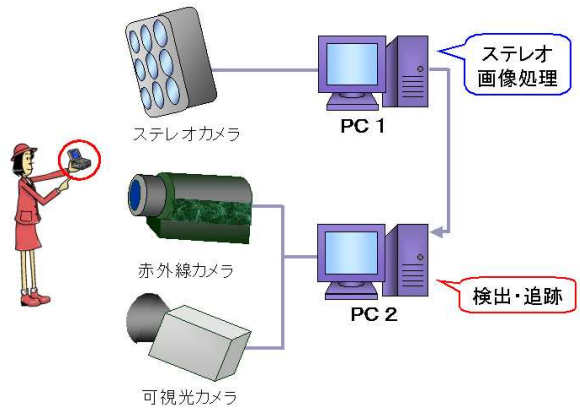

図 14　ハードウェア構成図

録し，同時刻に取得された(注6)3 枚の画像を選択して利用する．

4.2　キャリブレーション

各カメラの位置関係は図 15 のようになっている．手前下が赤外線カメラ，手前上が可視光カメラ，奥がステレオカメラである．それぞれ作業空間が中心に写るように光軸方向を決めた．

各カメラにレンズ歪みがあること，3 つのカメラの光軸（視点）を完全に一致させることが難しいことから，3 枚の画像の間の位置合わせを行う必要がある．本研究では，レンズ歪みを補正するために 2 次の幾何補正，光軸（視点）を補正するために 2 次元の射影変換を用いる．2 次元射影変換を用いる理由は，本研究の設定している作業空間が各カメラから見て比較的奥行きの範囲の狭い設定となっているためである．ただし，実際の計算では，両方を兼ねる 3×5 の行列を変換行列として用いた．

実際のキャリブレーションには以下の方法をとった．全てのカメラから見える位置にキャリブレーションボードを置き，ボード上の 25 点の特徴点を各カメラの画像から抽出する．キャリブレーションボードの大きさは縦 1m× 横 1.5m で，カメラから 2m の距離での可視光カメラと赤外線カメラの視野にほぼ等

(注6): 赤外線カメラには同期信号が入力できないため，厳密には同期をとることができない．本研究の方式では，1/2 フレーム以上のずれが無いことが保証されるだけである．

表 1　各 PC スペック・役割

PC	CPU	RAM	OS	役割
PC1	P4 1.7GHz	RDRAM 512MB	Linux kernel2.4	ステレオカメラからのデータを処理 PC2 へデータをネットワーク経由で送信．
PC2	P3 933MHz	SDRAM 256MB	Linux kernel2.2	PC1 からのデータを受取り 可視光カメラ・赤外線カメラからの映像を取得 ３つのカメラからのデータを統合，手と把持物体の認識・追跡
Intel C++ Compiler for Linux ver5.01				

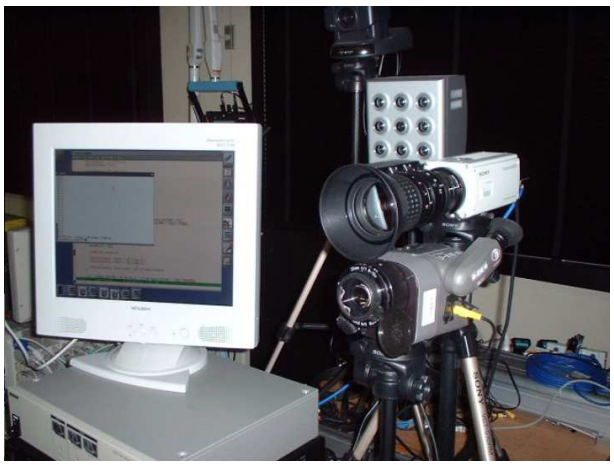

図 15　各カメラ位置関係図

しい．次に，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，ステレオカメラからの画像に対して以下の式で表される $M_{3\times5}$ を求める．実際の計算には最小二乗法を用いる．

$$\begin{pmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_n \\ 1 & \dots & 1 \end{pmatrix} = M_{3\times5} \begin{pmatrix} u_1 & \dots & u_n \\ {u_1}^2 & \dots & {u_n}^2 \\ v_1 & \dots & v_n \\ {v_1}^2 & \dots & {v_n}^2 \\ 1 & \dots & 1 \end{pmatrix} \quad (1)$$

ここで，$(x_1, y_1) \sim (x_n, y_n)$ は基準となる可視光カメラの画像上での特徴点座標，$(u_1, v_1) \sim (u_n, v_n)$ は補正前の赤外線カメラ，又は，ステレオカメラの画像上の特徴点座標である．

5.　映像への注釈付けとインデキシング

これまで述べてきた把持物体追跡システムと我々のグループで構築している図 1 のような知的撮影システムから得られる情報とを統合することにより，映像への注釈付けを行う手法について説明する．

把持物体システムと知的撮影システムそれぞれから得られる情報は以下のようになっている．

把持物体追跡システム:　物体の位置，物体のテクスチャ

知的撮影システム:　動作と発話による注目誘導行動，撮影された映像

これらの情報を用いることによって把持物体がいつどこからどのように移動し，そのときにどのような説明・映像が付加されたかを関連付けて保存する．処理の例を以下に示す（図 16）．

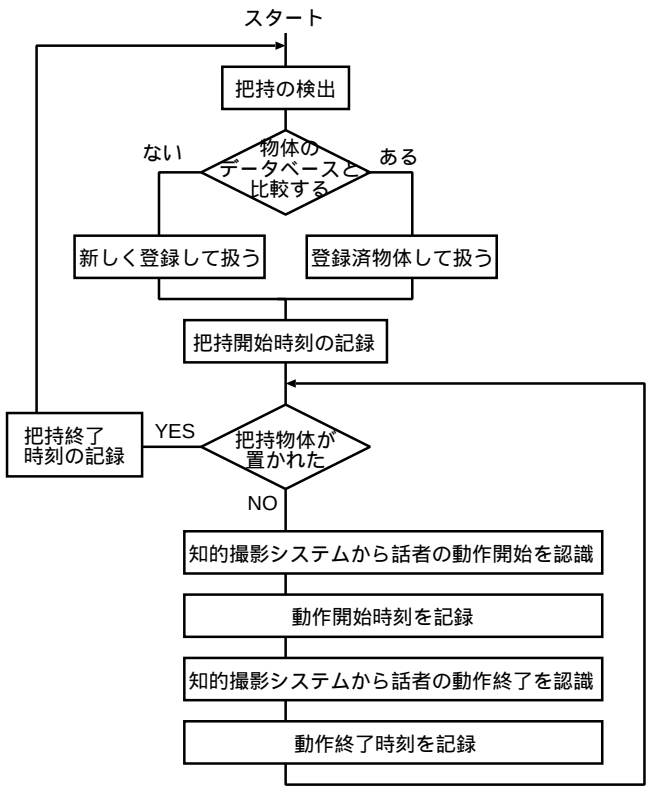

図 16　処　理　例

1　話者が物体を持ち上げた瞬間にテクスチャを取得．テクスチャは以後アイコンとしても使用する．またその時刻を動作の開始時刻として記録．

2　獲得したテクスチャを過去に把持された物体一覧と比較．もしはじめての物体だった場合は新規の物体として登録．過去に把持されたことのある物体の場合は，その物体に対する一連の動作として行動を追加する．

3　物体を把持中に知的撮影システムの「これは」「このように」などの発話情報と動作認識（注目誘導行動）から物体に説明を与えている状況を検出，その動作が起った時刻を取得，記録．

4　話者が物体を置き，把持が終わったらその時刻を記録し，物体に対する説明が終わった時刻とする．知的撮影システムから得られた映像から各物体の動作開始時刻 ~ 終了時刻の映像クリップを切り出し，その物体に対する説明映像とする．

5　再生時に取得したアイコン一覧が表示される．アイコンをクリックすることにより付加された説明映像が再生される．アプリケーションの概要を図 17 に示す．

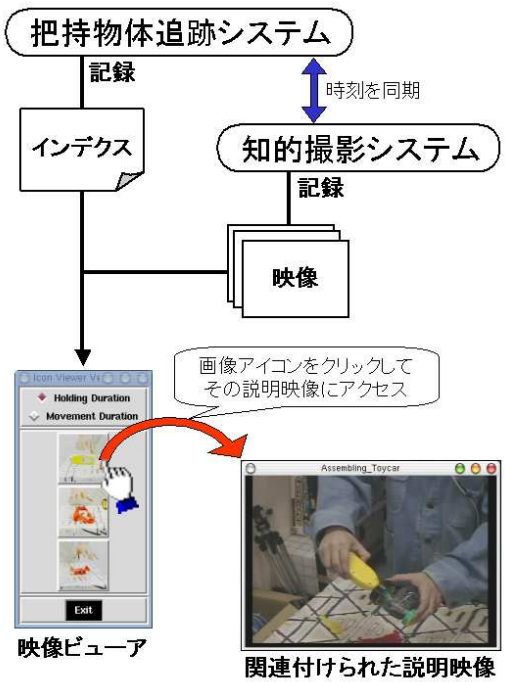

図 17　アプリケーション概要

図 18　物体追跡例

6. 実　　験

把持物体の検出・追跡

以下のような環境で把持物体の追跡実験を行った．システム構成は 4.1 節で説明した通りである．

- 「テーブルの上に外観の似た複数の物体が存在し，背景中で人物が移動する場合」と，「机上にはただ一つの物体だけ存在し，背景中に人物が存在しない場合」の 2 通りの場面を設定
- 被験者 (3 人) に自由に物体を掴んで動かしてもらう

追跡例を以下の図 18 に示す．検出・追跡された把持物体を赤枠で表示した．また，各人，各設定での 15 秒の動画 (各 450 フレーム，計 2700 フレーム) を評価に使用し，各フレームを正解・未検出・誤検出の 3 通りに分けてそれぞれの割合を算出した．ここで「未検出」とは把持物体領域が検出されなかった場合，「誤検出」とは検出されるべき把持物体領域以外の領域が抽出された場合を示す．結果を表 2 に示す．比較のため，2 台のカメラ (可視光と赤外線) のみを用いて追跡した場合の結果も付加した．これらの結果から，複雑な環境でも精度良く把持物体のみを追跡しているのが分かる．本システムでは，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，非常に良い結果であると言える．

表 2　追 跡 結 果

(a) 提案手法 (カメラ 3 台)

状況	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%)
未検出数	30 (2.2%)	11 (0.8%)
誤検出数	4 (0.3%)	80 (5.9%)

(a) 比較 (カメラ 2 台)

状況	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	1005 (74.4%)	370 (27.4%)
未検出数	37 (2.7%)	5 (0.4%)
誤検出数	308 (22.8%)	975 (72.2%)

ただし，現在はまだ以下のような問題点があり，今後の改良が必要である．

- 小さな物体は追跡できないこと．現在は手のひら程度のサイズ以上が必要となっている．
- 作業空間の奥行きの範囲を深く取ることができないこと．3 つのカメラの位置合せのために 2 次元の射影変換を用いているため，奥行きの範囲が大きくなると誤差が大きくなる．

映像のインデキシング

実際に映像へのインデキシングを行った例を示す．内容は一人の人物が机の上にある料理を順に説明していくことを模擬したものである．

得られた映像を図 19 に示す．図は把持物体追跡の結果を示しており，左下に注釈映像の候補として選択されている映像を表示している．「把持物体の検出→動作の検出→映像との関連付け」という流れに対応して，把持物体上に表示されている枠が「破線→太い赤実線→実線」と変化していく．

物体と関連付けられる映像は，それが持ち上げられた時点から，説明が終了し，再び物体が置かれたところまでを 1 クリップとした．この例では，意図した通りに映像への注釈付けが行われている．

7. ま　と　め

複数の画像センサを相互補間的に用いて，プレゼンテーション場面での手と把持物体を追跡，認識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とを組み合わせることにより，映像中の物体に注釈映像を関連付け，閲覧が可能なシステムの構築例を示した．

実験例で示したように，物体に関する事前知識がない状態でも精度良く追跡が行えるという点で，他に例を見ないものと

(a) 図は把持物体の追跡経過を表示した例．左下に上書きされている画像は，知的撮影システムで撮影されている映像．

(b) 把持物体が検出されたところからインデキシングが行われる．検出された把持物体（下中央）の上に枠が表示されている．

(c) (a) の発話情報と (b) の動き情報から指示・例示動作が検出されると，それに関する情報（開始・終了時刻，動作の種類）が把持物体と関連付けて保存される．

(d) 把持物体が机に置かれたところでインデキシングを終了し，把持物体の置かれた位置とテクスチャがインデックス情報と共に記録される．

(e) 前に検出された物体の情報は保持したまま，新たな把持物体が検出され，同様の処理が始まる．

図 19　物体への注釈付けの結果

なっている．また，撮影時に物体に関するインデクスを付加していくことができるため，これからのマルチメディアコンテンツ制作への利用可能性は高い．

しかし，まだ解決しなければならない課題も多く，これからの研究を必要としている．今後，以下のような改良を行っていく予定である．

- より小さな物体の追跡を可能にすること
- 撮影・追跡範囲の拡大
- システムの小型化

文　献

[1] “高速ステレオビジョン FZ930 取扱説明書”，コマツ，2000.
[2] 吉見 修，山口 博義，“膨張確度係数を用いた視差画像における物体輪郭の鮮鋭化”，画像センシングシンポジウム，2000．
[3] 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”，信学総合大会，D-12-190，1999．
[4] 尾関基行, 中村裕一, 大田友一, “プレゼンテーションの知的撮影システム”, 信学総合大会（シンポジウム）, 2001.
[5] Ismail Haritaglu, Ross Cutler, David Harwood and Larry S. Davis, “Detection of People Carrying Objects Using Silhouettes”, ICCV,1999
[6] Motoyuki Ozeki, Yuichi Nakamura, Yuichi Ohta, “Camera-work For Intelligent Video Production –Capturing Desktop Manipulations”, IEEE International Conference on Multimedia and Expo, pp.41–44, 2001.
[7] 尾関基行, 伊藤雅嗣，中村裕一, 大田友一，“複合コミュニティ空間における注目の共有 ~ 人物動作理解による物体への注釈付け ~ ”, 日本 VR 学会 第 6 回全国大会, pp.239-242，2001.