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.

さりげなく作業支援を行うメディア
～ 物体変化の認識と作業過程の同定 ～

津吹　陽介† 　小阪　拓也† 　亀田　能成† 　中村　裕一†† 　大田　友一†

† 筑波大学大学院システム情報工学研究科　〒 305–8573 茨城県つくば市天王台 1–1–1
†† 京都大学学術情報メディアセンター　〒 606-8501 京都市左京区吉田本町
E-mail: †{tsubuku,kosaka,kameda,ohta}@image.esys.tsukuba.ac.jp, ††yuichi@media.kyoto-u.ac.jp

あらまし　本稿では，机上作業を支援・教示するためのマルチメディアシステムを提案する．この枠組みでは，システムはユーザを観察し，必要なときにだけ必要な支援をする．すなわち，ユーザの状況を推定し，状況に応じて教示映像を提示することによって適切なアドバイスを行う．本稿では特に，ユーザの作業状況の認識について説明する．我々の手法の特徴は，複数の画像センサを用いて把持物体の変化検出を行うこと，および，これにより作業状況を推定することである．この手法の可能性を簡単な実験により示す．

キーワード　映像メディア，映像インデキシング，視体積交差法，把持物体追跡，教示システム，対話システム

Video-Based Media for Gently Giving Instructions
– Object change detection and working process identification –

Yosuke TSUBUKU†, Takuya KOSAKA†, Yoshinari KAMEDA†, Yuichi NAKAMURA††, and Yuichi OHTA†

† Systems and Information Engineering, University of Tsukuba, 305–8573 Japan
†† Academic Center for Computing and Media Studies, Kyoto University, 606–8501 Japan
E-mail: †{tsubuku,kosaka,kameda,ohta}@image.esys.tsukuba.ac.jp, ††yuichi@media.kyoto-u.ac.jp

Abstract　This paper proposes a framework of multimedia for gently giving instructions. In this framework, a system observes the user and gives appropriate advices that relevant to the user's status and really necessary for the user. In other words, a system estimates the users' status and gives appropriate advices by giving pre-recorded videos for the task. In this paper, we present a new method for object tracking and measurement by two or more pairs of cameras, and an status identification method based on the object recognition. We also present the potential of this framework by simple experiments.

Key words　Video-Based Multimedia, Video Indexing, Shape-from-Silhouettes, Object Tracking, Instruction System, Interaction System

1. はじめに

組み立て，料理，実験などの作業を行っているとき，正しい手順で作業を進めているのかどうか，次に何をしたらよいのか，またどのくらいの選択肢があるのかを知りたいといったことがしばしば起こる．このような場合，もし人間の先生がいたならば，生徒の様子を見守って，状況に適したわかりやすい説明をしてくれるだろう．我々は，このような機能を持ったメディアが必要であると考えている．

しかし，従来の映像メディアの構成とその利用形態を考えると，このような機能を持たせることは簡単ではない．例えば，料理番組やそれを基にした教材ビデオを考えてみよう（我々はこれを教示映像と呼ぶ）．多くの場合，料理番組はレシピで決められた一定の手順でしか収録されていない．教示映像などでも，示された手順通りに作業を進めることをユーザに強要し，その確認はユーザまかせである．その部分的な解決方法としては，映像処理技術を用いたり，人手を利用してインデックスを付与し，要約を行ったり，再生の順番を変えたりすることが考えられているが，それをユーザの状況に合わせて適切なタイミングで適切な映像を提示するのは難しい．また，従来型の質問

応答システムのように，適切な質問を発するように強いる枠組みでは，作業中のユーザにとって使いやすいシステムとはいえない．

このような問題を踏まえ，我々は「さりげなく作業支援を行なう映像メディア」を実現することを目的とし，作業状況の推定とその結果に基づく適切なメディア提示について研究を行っている． 具体的には，机上作業を対象とした把持物体を追跡し，その物体の変化を検出し，前もって記録されインデックスの付けられた映像の中から教示に直した部分を取り出し，適切なタイミングでその映像部分を提示するという方法をとる．これにより，システムは自発的にユーザの現在の状態を推定し，次の操作説明，将来的な選択肢など，種々の情報をユーザに負担をかけずに提示できるようになる．ただし，この枠組みには総合的なシステムが必要とされているため，現在の段階では，全体を自動化するまでの研究には至っていない．本稿では，重要な要素部品の提案と，その簡単な実験例を示すことによって，将来的な可能性を示す．

2. さりげなく作業支援を行うメディア

人間の先生が作業などを教える場合，先生に期待される役割には多様なものが考えられるが，我々は次のような役割を重要視している．

- 生徒の状況を的確に認識する．ただし，生徒の意図が分からない場合には，確かめるための質問をする．
- 生徒の操作に自由度を持たせる．例えば，依存性のない操作の順番を変えてもかまわない．
- 生徒側に助けの必要がない場合にはただ見守り，無駄なおせっかいをしない．
- 困っている状況に合わせて的確な教示をしてみせる．生徒が質問したときには丁寧に答える．

我々は，このような機能を実現するための総合的な枠組みとして,「さりげなく作業支援を行う映像メディア」を考えている．この枠組は，あらかじめ蓄積された映像を用いて，上記の機能を実現するものである．

その概要を図 1 に示すが，ここで重要となる要素技術としては，以下の 4 つがあげられる．

（1）教示映像へのインデキシング
（2）把持物体とユーザの動作の認識
（3）ユーザの現在の作業状況の推定
（4）ユーザの状況に応じた情報の提示

このうち，教示映像へのインデキシングについては，関連する研究が広く行われており，また既に我々も映像を用いた QA システムの枠組 (QUEVICO) [1] について提案している．ユーザの現在の作業状況については，把持物体と動作の認識を行なう．それに関しては，4. 章で詳しく述べる．

一方，ユーザの作業状況の推定については，検出した物体・動作と，教示映像中のインデックスとを対応付けることにより実現する．

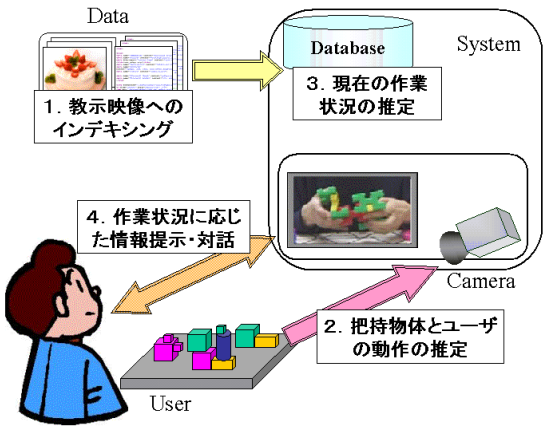

図 1　システムの概要

3. 作業状況の記述と推定

3.1 作業状況の記述

ユーザの行動を見守る機能の実現は簡単ではない．人間の先生を考えた場合でも，ユーザの意図を理解するためには，観察力と細心の注意が必要となる．本研究では，このような機能を直接的に実現しようとするのではなく，物体の識別とユーザの動作の時系列的な認識によって，現在の作業状況を想定される状況と対応させるにとどめる．

その概要を説明する前に，いくつかの用語の整理をしておく．まず，我々は作業をタスクの集まりとして考え，各タスクはユーザの低レベルな動作であるアクションと，関連する物体で構成されるものとする．

各々の説明は以下のようになる．

作業: 作業は教示の目標であり，また図 2 に示すようなタスクの集合で表現される．

タスク: 目的の作業を達成するために行う動作であり，以下に述べる” アクション” と” 関連する物体” を複数組み合わせることで記述する．現在のところ，基本的な動作である「(物体を）移動させる」「(物体を）組み付ける」の 2 種類のみを想定しているが，将来的には,「(物体を）はずす」「(物体を）変形させる」等の様々な動作も導入する予定である．

アクション: 画像処理による認識に対応したユーザのプリミティブな動作であり，“アクション名”，“関連する物体” で記述する．現在のところ,「(物体を）持ち上げる」「(物体同士を）接触させる」「(物体を）置く」の 3 種類を想定しているが，これに関しても将来的な拡張を予定している．後で述べるように，物体の変化が検出できれば，それに関連するようにアクションも導入できるようになる．

物体: 動作の対象となる具体的形状を持つもので，画像特徴 (テクスチャ，色ヒストグラム，形等) で記述する．後述するように，物体の体積を計測することができれば，それも記述のための良い特徴となる．

タスクとアクションの対応付けを可能とするために，各タスクの代表的なアクションパターンを予め設定している．タスク

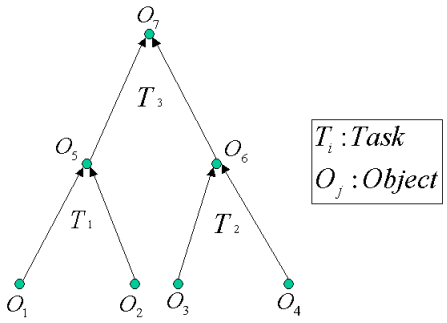

図 2　作業の表現例

- *lift* O_1
- *lift* O_2
- *make* O_1 *and* O_2 *touch (and make* O_3*)*
- *place* O_3

図 3　タスク「物体の組み付け」のパターン

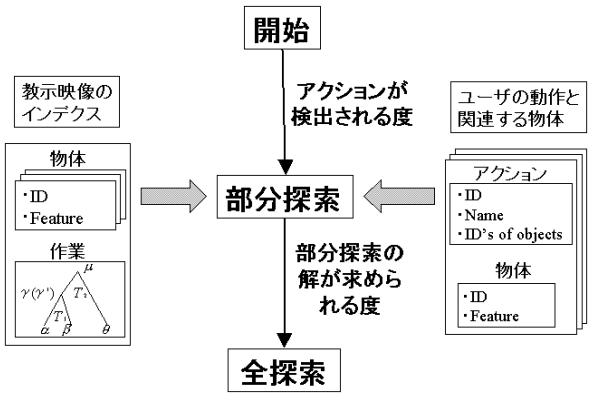

図 4　ユーザの作業状況の認識

「(物体を) 組み付ける」の例を図 3 に示す．その際，図中の O_i は物体を意味する．

3.2　作業状況の推定

以下に作業状況の推定の概要を示す．(図 4 参照)

(a) 準備として教示映像のインデックス情報を読み込み，作業のタスクと物体に関する情報を取得しておく．

(b) ユーザが何らかのアクションを行った場合，そのアクション名と関連する物体の情報 (例えば，テクスチャや色ヒストグラム等) を時系列順に記録していく．

(c) (a) と (b) を照合することにより，ユーザの現在の作業状況を推定する．

ただし，(c) の照合には以下の類似度を用いる．

物体の類似度 $S(O_i, O_j)$:　色や形状等の画像特徴に基づく．現在のところ，画像中の物体領域の色ヒストグラムを比較することにより類似度を求めている．

アクションの類似度 $T(A_i, A_j)$:　アクション名の同一性と関係する物体の類似度との積で計算する．アクション $A_i(N_i, O_{i1}, O_{i2}, ..., O_{in})$ と，アクション $A_j(N_j, O_{j1}, O_{j2}, ..., O_{jn})$ との類似度 $T(A_i, A_j)$ は以下の式で求められる．ここで，N_i はアクション，O_{ik} は関連する物体を意味する．

$$T(A_i, A_j) = \delta(N_i, N_j)(\prod_{k=1}^{n} S(O_{ik}, O_{jk}))^{\frac{1}{n}} \quad (1)$$

$$\delta(X, Y) = \begin{cases} 1 & (X = Y) \\ 0 & (X \neq Y) \end{cases} \quad (2)$$

タスクの類似度:　画像認識を用いても，タスクを直接検出することは困難なため，ユーザが行ったタスクと教示映像中のタスクとを直接対応づけることができない．そこで，現在に至るまでの一連のアクション同士を比較することで対応付ける．作業状況の推定は，以下で述べる部分探索と全探索の組合せで実施される．

部分探索:　システムは，教示映像中のタスクと対応するユーザの一連のアクションを探す．その際，DP マッチングを適用することで，教示映像に記録されたとおりに行動していない場合でも，部分的に対応づける．この処理では，全ての部分的な対応付けが可能な全ての組合せを求める．

全探索:　タスクの一貫性を考慮に入れることで，起こりうる一連のタスクを推定する．本研究では，部分探索の結果を基に深さ優先探索を行うことで，起こりうるタスクの組み合わせを求める．観測された結果から推定されるタスクの組合せの数が爆発的に大きくなる場合には，図 5 のようにユーザに問い合わせるることで曖昧性を削減する．また，各候補に対して，各類似度を基にスコアを付けるということも行っている．

これらの処理は我々の試作システムにすでに実装されているが，細部についてはまだ十分に検討が終わっていないため，これらの報告は他の機会に行いたい．

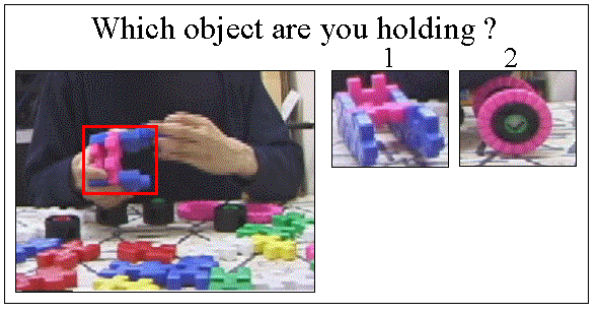

図 5　ユーザの作業状況の認識

4.　把持物体認識・追跡

4.1　物体・動作認識の条件

机上作業の支援を考えた場合，様々な物体を認識することが必要となるが，重要な物体の多くは人間によって把持され，移動されるという知見が得られている．そのため，本研究では，把持物体に焦点をあて，把持物体の種別とその変化の認識によって，上記の枠組みに利用することを提案する．

まず，把持物体認識の前提条件として，本研究では以下のようなものを考えている．

(1) 複数の物体，複数の人が存在しても良い．

（2）把持される物体の大きさ，色，形状等に関する予備知識はない．上述したように，教示映像中に物体のテクスチャなどが蓄積されているが，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見え方やその変化に関する知識をあらかじめ与えておくことは難しい．

（3）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，物体も変化するため，テーブル上の状況や背景は常に変化するものとして考える．

（4）作業に関わる物体の存在する範囲 (以下，立体的な範囲をボリュームと略す) が，作業空間としてあらかじめわかっている．

以上の考察から，本研究では，画像中から直接物体テクスチャを探し出すのではなく，ある特定のボリューム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．

4.2 複数の画像センサの利用

一般に，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，赤外線カメラを利用する [2]．これにより手領域，把持物体領域を次のように求めることができるようになる．

手領域: 可視光カメラから得られた肌色領域と，赤外線カメラから得られた肌温領域の論理積を手領域とする．

把持物体領域: フレーム間差分とテンプレートマッチングの併用によって動領域を求め，手領域と近接している動領域を把持物体領域とする．

なお，両領域の実際抽出にあたっては，条件をみたす領域に対してさらにノイズ除去を行っている．

また，特定のボリューム内にあることを検出するために，視体積交差法 [3] [4] を利用する．このために，可視光カメラ，赤外線カメラのペアを 2 視点に設置する．

手領域の抽出

肌色領域を検出する手法はこれまでに数多く提案されてきている．かなり精密なモデルも提案されているが，本研究では，RGB 空間でのゆるい分布を用いる．その理由は，我々の想定する環境では，上述したように照明の変化，テーブルや物体等の相互反射や，肌色の個人差が予測できない場合があることを考え，厳密なモデルをあらかじめ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漏れが少なくなることを目指す．

そのために，rg 色度平面上に肌色モデルを作る．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．ただし，これは我々の環境 (カメラ，照明，机の色等) での値である．

$$mean(\bar{r}, \bar{g}) = [0.415483, 0.330955]$$

$$\Sigma = \begin{bmatrix} 0.002025 & -0.000557 \\ -0.000557 & 0.000387 \end{bmatrix}$$

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．

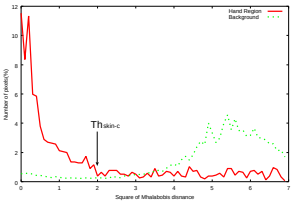

図 6　マハラノビス距離

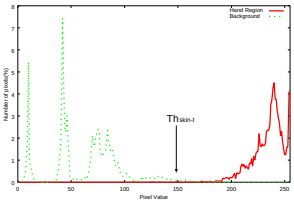

図 7　肌温領域の設定

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

図 6 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，及び両領域を分離するために用いた閾値を示している．他の要素領域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が，我々の実験環境では背景に肌色に近い領域が存在するため，図 6 のような閾値とした．

次に，肌温領域の検出について説明する．赤外線カメラから得られた画像は，温度の高い領域ほど高い画素値を示す．本研究では，閾値処理によって画素の高い部分のみを抽出し，得られた領域を肌温領域とする．図 7 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．そのため，肌色領域がほとんど抽出されるように閾値を設定した．

4.3 把持物体領域と手領域の分離

把持物体領域と手領域を分離するために，フレーム間差分と背景差分の 2 種類の手法を併用する．一般的には，動物体の検出には背景差分が有効であり，多くのアプリケーションで利用されているが，我々の環境では背景に変化が生じる可能性があるため，フレーム間差分とテンプレートマッチングを併用する．その方法は以下のようになる．

まず，フレーム間差分を動領域の検出に用いる．上記で想定した状況下では，机上の空間で検出された動領域は，手と把持された物体に属する領域であると考えられる．したがって，図 9 のように検出された領域が十分な大きさを持っている場合には，それを動領域とする．しかし，物体の動きが少ないときなどには，フレーム間差分が十分に検出されない．そのため，本研究ではテンプレートマッチングを併用して，物体の推定位置を求める．

テンプレート画像は，フレーム間差分が検出されたときの差分の重心を中心とした 50×50 の画像を使用し，テンプレート画像は差分が検出されたときのみ，新しいものに更新する．

4.4 ユーザの動作の認識

各アクションについて，現在の認識方法を以下に示す．

持ち上げる:　物体領域の中心座標が一定の高さよりも高くなった場合に持ち上げたと認識する．

接触させる:　二つの物体領域の中心座標が一定の高さよりも高く，さらにそれらの距離が接近している場合に接触したと認

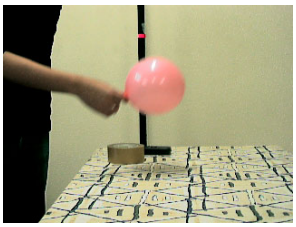

図 8　原　画　像

図 9　フレーム間差分法

識する．

置く:　物体領域の中心座標が一定の高さよりも低くなった場合に置いたと認識する．

5.　体積の計測と変化の認識

背景差分が安定に求められるタイミングでは，視体積交差法により物体の体積を計算する．前節の方法では，動領域の存在は確実に検出できても，その領域を正確に背景から分離することができないためである．

まず，各視点における把持物体領域を検出する．しかし，フレーム間差分では物体の形状を正確に得ることができない．そこで背景差分を用い，物体検出を行う．背景画像はフレーム間差分で動領域が検出されない時刻に更新する．

これらの画像を入力とし，ボクセルカービングによって物体の体積を得る．試作システムでは，50×50×50 (=125000) のボクセルを空間上に設置し，1 個の大きさは 20 mm×20 mm×20 mm とした．このような設定で，我々の計算機 (dual Xeon 3.06GHz) ではで約 15 frame/sec で動作が可能である．このようにして得られた体積は物体認識に用いることができるだけではなく，その変化の認識にも用いることができる．

しかし，体積の計測結果はノイズに弱く，そのままではノイズの影響と物体の実際の変化を区別することは難しい．そこで，ノイズの影響を軽減するため，一定幅のウィンドウを設定し，移動平均と同様にそのウィンドウをスライドさせながら 2 番目に小さい値を採用することにした．その理由は，視体積交差法では，計測値は原理的に真の体積以上の値をとるが，視点の数が少ない場合には，その増分が顕著に現れるからである．したがって，ノイズが無い場合には，最も小さい値が真の値に最も近いと考えられるが，ノイズにより極端に小さくなった場合を省くために，2 番目としている．将来的にはロバスト統計などの利用も考えられる．

これらの計測値を利用することによって，各アクションに対応する物体変化の認識を行なう．

6.　実　　験

6.1　把持物体認識

把持物体検出を行なったときの把持物体の追跡精度を調べるために，三種類の状況で，把持物体の検出を行なった．

一つ目は，図 10 の上段の花を用いて，机上領域で任意に把持物体を動かした．　二つ目は，図 10 中段の車のおもちゃを用いて，両手が自由に動くという条件設定のもと，机上領域で任意に把持物体を動かした．　最後は図 10 下段の小さな球を用いて，一つめと同じ条件で追跡を行なった．

（ここで，ボクセルの最小単位は 2 cm とした．この場合の計算速度は，約 15 frame/sec である.）

表 1 は各物体に対する追跡結果である．ここで，”準正解 ”とは物体枠から 5 ピクセル分外側が検出された場合，”誤検出 ”とは把持物体以外の領域が検出されたか，検出が起こらなかった場合を表す.

車を用いた追跡の例では，両手を使う作業の問題，つまり把持物体が片方の手の陰になると，検出精度が落ちること，また小さなボールを用いた追跡の例では，微小領域を精度良く追跡することは難しいこと，小さな物体であるため，ノイズと見なされ，検出されにくいことなどが問題点としてあがっている．しかし，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，従来の研究に比べて十分に良い結果であるといえる．

表 1　検 出 結 果

	花	車の模型	小さな球
正解	950 (95.0%)	920 (92.0%)	524 (52.4%)
準正解	50 (5.0%)	54 (5.4%)	297 (29.7%)
誤検出	0 (0.0%)	26 (2.6%)	179 (17.9%)
total	1000(100%)	1000(100%)	1000(100%)

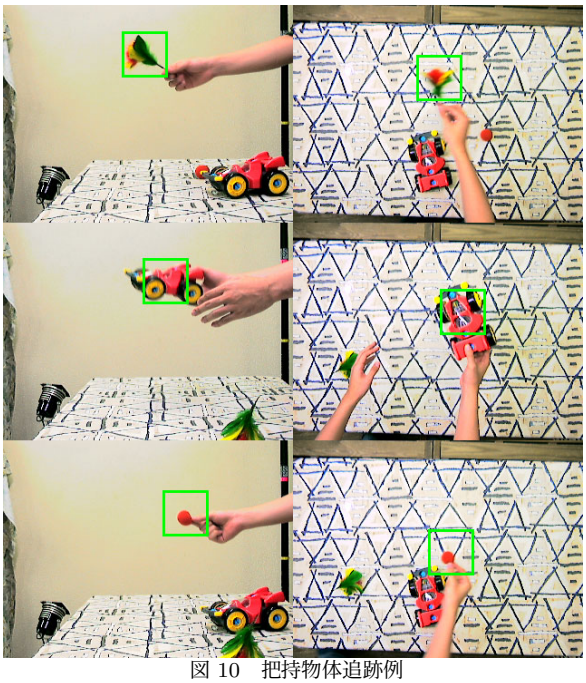

図 10　把持物体追跡例

6.2　体積変化の検出

体積計測の精度を計るため，ニ種類の実験を行なった．ひとつは，２次元画像取得時の，フレーム間差分と背景差分の精度を比較する実験，もうひとつは体積計測の精度を確認する実験である．

一つ目の実験では，一辺が 20 cm の立方体を用い，机上空間を動かした．

図 11 は，背景差分法による手法で体積を測定した結果である．グラフは作業空間中で検出された体積の計測と，空間中の

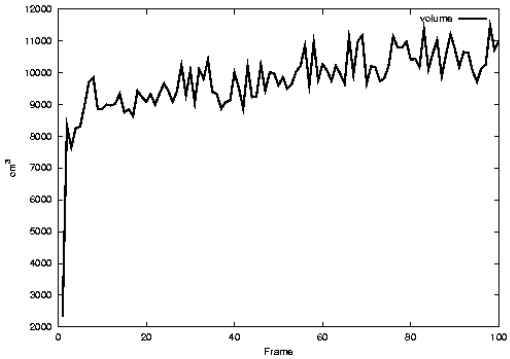

図 11　体積検出結果

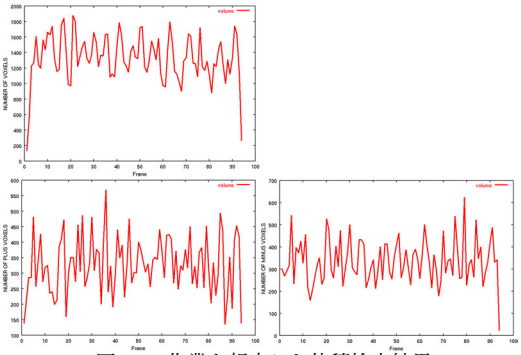
図 12　作業を想定した体積検出結果

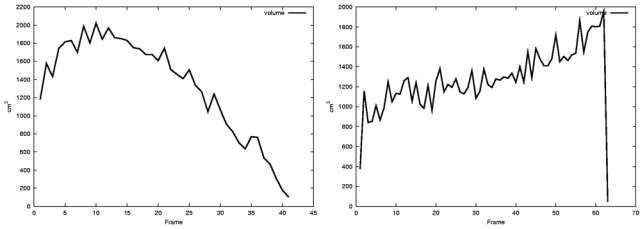

図 13　風船を用いた体積検出結果

移動量を表している．ただし，立方体の体積が 8000 cm^3 であるのに対して，手法の性質上，真の体積以上の値が算出されている．

図 12 物体を机上で動かし，計測されるノイズの量について見積もったものである．図 12 上段は物体のボリュームを計測したもの，下段前フレームとの差分を比較することにより，ノイズ分を抽出したものである．これを見ると，物体を大きく動かすと大きなノイズが検出されてしまうことが分かる．今後，作業を行なって行く上でこれらノイズの対策を検討する必要がある．

次に，風船を使って体積変化を観測した実験例を示す．この例では，風船を膨らまして，空気を抜きながら作業空間中を移動させ，体積の変化を観測する．実際に観測された結果を図 13 に示す．図左側のグラフは，風船のしぼむ様を観測したもの，右側のグラフはある時刻で風船を割ったものである．グラフを見ると，体積の変化によってそれぞれの現象が捉えられていることが分かる．

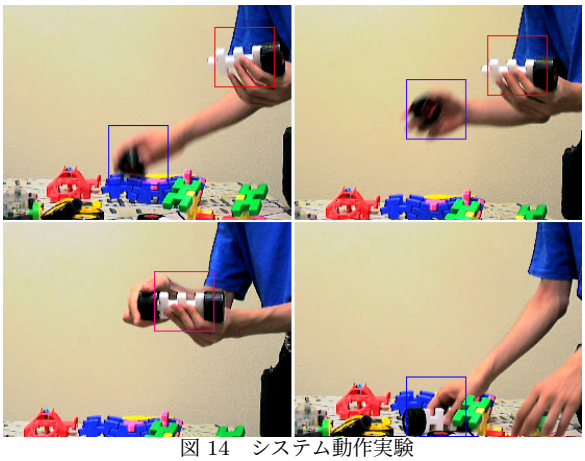
図 14　システム動作実験

6.3　作業状況の認識

物体追跡や，体積検出を用いて，各アクションの認識を行うことも我々のシステムでは可能である．ここでは，ブロックの組み付け作業を題材とし，物体の結合・分離，物体を持ったか，置いたかの 3 種類のアクションについて図 14 に認識された例を示す．

7.　結　　論

本稿では，さりげなく作業支援を行う映像メディアの枠組みを提案し，その際に必要となる，ユーザの作業状況を推定する手法について述べた．この手法では，2 種類の画像センサを相互補完的に用い，複数の視点に設置することにより，把持物体の追跡と体積の変化を検出する．本稿ではこの要素技術について説明した．また，ブロックを組み付ける作業状況の推定を行った簡単な例を示した．

本研究のシステムはまだ構築し始めた段階であり，その要素技術も十分に検討が終わっていない．これらを少しずつ検証しながら全体としての可能性を探ることが今後の課題となっている．また，この枠組みで重要な部分を占めるユーザインタフェースについても，その設計と実装を進め，さりげなく作業支援を行うことの可能性と意義を明らかにしていく予定である．

文　　献

[1] H.Izuno, Y.Nakamura, Y.Ohta ”QUEVICO QA Model for Video-based Interactive Media” ,Proc. Third International Workshop on Content-Based Multimedia Indexing,pp.413-420, 2003.

[2] M.Itoh , M.Ozeki, Y.Nakamura, and Y.Ohta ”Simple and Robust Tracking of Hands and Objects for Video-based Multimedia Production”, IEEE Conf. on Multisensor Fusion and Integration for Intelligent Systems, pp.252-257, 2003.

[3] S.M.Seitz and C.R.Dyer, ”Photorealistic Scene Reconstruction by Voxel Coloring”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp.1067-1073, 1998.

[4] A. Laurentini, ”The Visual Hull Concept for Silhouette-Based Image Understanding”, Transactions on Pattern Analysis and Machine Intelligence, 16(2), pp.150-162, 1994.