画像・映像の撮影・編集・提示から対話的映像メディアまで

中村 裕一 (筑波大学 機能工学系)
〒 305-8573 つくば市 天王台 1–1–1
(yuichi@image.esys.tsukuba.ac.jp)

1 はじめに

マルチメディアという言葉がすっかり定着したが，複合メディアの本質的な部分には，まだ手付かずの問題が多い．短時間でわかりやすく情報や知識を伝えるメディアの形態，また，その構築方法について，まだまだ試行錯誤で探っていく必要がある．

このような問題意識から，筆者は複合メディアの一つである映像をテーマに関して，画像処理，動作認識，自然言語処理を用いていくつかの研究を行ってきた．過去，現在の研究スコープを広くまとめると図 1 のようになる．種々の蓄積型メディアや実時間型情報源と人間との柔軟なインタラクションめざし，その要素的な研究を進めている段階である．

本稿では，その中から，ニュースや料理番組などのインデキシングにおける画像情報，言語情報の利用，プレゼンテーション，作業，個人行動等を伝えるための映像撮影・編集と動作，発話情報の利用，対話的に映像内容を提示するためのデータ構成やその QA 手法 (QUEVICO)，そのインデックスを自動的に取得する試み等について紹介する．

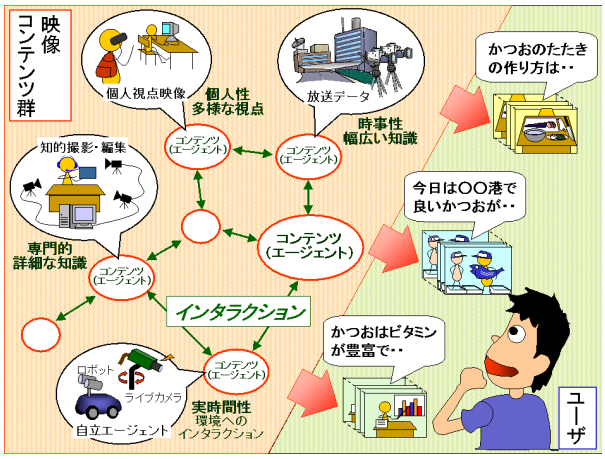

図 1: 映像・マルチメディアで何ができるだろうか？(ユーザが「かつおのたたきが食べたいなあ」と言った場面を想定)

2 インデキシングと検索: まずは映像の整理

映像処理の分野でのこれまでの研究や実用化を分かりやすく整理すると図 2 のようになる．上段が従来の映像コンテンツの流れであるが，一方向的であり，マルチメディアとしての有効利用は難しい．そのため，90 年代の初めから研究されてきたのが図 2 の右下部分，放映番組や資料映像等の解析である．スポーツ番組や娯楽番組のハイライトだけを見たい，ニュース番組のダイジェストを作って欲しい，長期間蓄積された映像から自分に興味のある事柄に関するものだけを集めたい等の要求に応えることによって，映像データのそれまでになかった利用方法を可能にするものである．

これまで試みられてきた手法で，最も簡単で効果的な方法は，映像中の言語情報を用いる方法である．例えば，Informedia プロジェクト [6] の News on Demand では，ニュース映像のクローズトキャプションを電子的に取得し，それを用いる．ユーザの要求があった場合には，蓄積されたデータを，文書検索でよく用いられる TFIDF を用いて検索することにより，該当するニュース映像を提示する．また，言語情報を音声認識によって得る手法，映像中のキャプションを抽出して文字認識を行う手法も提案されている [7](図 2(e)), [8].

しかし，映像内の重要な情報を画像が担っている場合も多い．その最も一般的な例としては，人物の顔があげられる．例えば，登場人物によるインデックスをつけるためには，顔認識を用いることが必要となる [9]．さらに，言語情報との統合によって顔画像データベースを構築し，それを検索に用いる試みも行われて

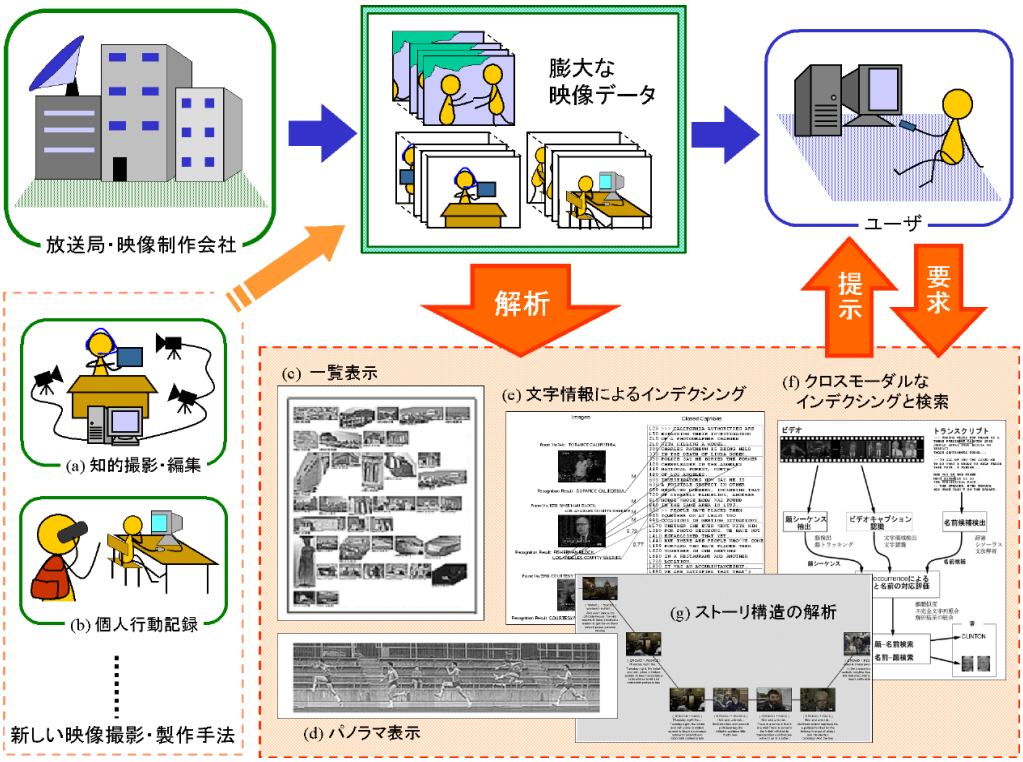

図 2: 映像・マルチメディア処理とその目的

いる [10](図 2(f)). 映像中の顔から名前，名前から顔が検索できれば，映像中の人物情報を検索，提示するための有効な手段となる．これらの研究が可能になってきたのは，顔検出，顔認識が実用化のレベルに達してきたことに負うことが大きい．

映像のストーリー構造の解析・要約には，画像と言語の両方の特徴を統合的に処理することが有効である．例えば，ニュース映像ではスピーチ，会議，訪問等，特定の状況を説明する部分が重要な役割を果たしているため，画像，言語両方からこれらの手がかりを取り出し，関係付けて整理し直す手法が提案されている [11]．その結果を，例えば図 2(g) のように時間を追って並べれば，ニューストピックの要約ができあがる．

さらに，映像の要約を映像として再構成するのも有効な手法である．つまり，映像中の重要な部分を選び出し，それをつなぎ合わせることができれば，簡単な要約ができあがる．Video Skimming と呼ばれる手法は，画像，音声，発話 (トランスクリプト) から特徴を抽出し，統合的に用いることによって，映像の長さを数分の一から二十分の一程度まで縮めることを可能にした [12].

本節で述べたような処理以外にも，映像の高度利用を目指す研究が盛んに行われている．また，最近では，映像解析結果を用いて個人の好みに合ったサービスを提供するサービスも模索されており，将来的には大きな市場を生むと予想されている[1].

3 映像メディアの取得: 簡単にコンテンツを作る

放送番組や既存の映像だけではコンテンツの量が不足すること，著作権のしばりを受けずに自由に使えるコンテンツが欲しいこと等から，一般企業や教育機関，さらには個人のレベルでも，手軽に映像を製作したいという需要は大きい．しかし，映像の撮影は，世界で起こっている出来事の一部分 (時間，空間的な一部分) を知的に切り出し，編集する行為であり，真面目に取り組むとかなり難しい問題でもある．単純に撮り流したホームビデオが，他人にとって見るに耐えない代物となることからも，それがよくわかる．このように，映像を誰でも手軽に使えるコミュニケーショ

[1]例えば，video portal という言葉が生まれている．これは，WWW ページに対するポインタを提供するポータルサイトのように，映像コンテンツへのポインタを提供するサービスをさす.

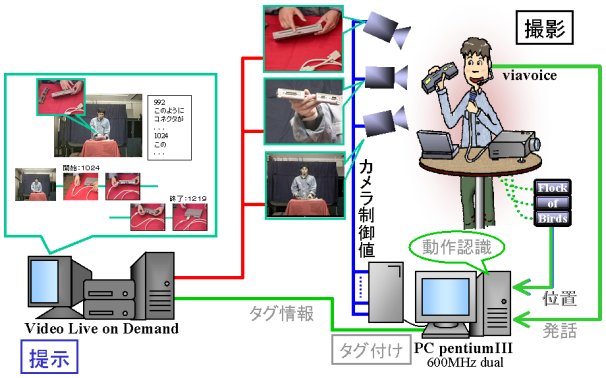

図 3: 自動的に「撮って編集する」システム

図 4: カメラ設定の選択表 (一部分)

ン手段とするためには，映像撮影の問題を見直し，それをサポートするシステムを用意することが必要である．

我々はその一つのアプローチとして，料理や組み立て等の解説 (プレゼンテーション) 場面を題材として，図 3 のようなシステムを構築している [17][20]．このシステムに，**カメラマンの機能** (人間の行動を知的に撮影する)，**ディレクターの機能** (人間の行動を認識して映像を知的に編集する) の 2 つの機能を持たせることによって，手軽に映像メディアを製作する環境を実現する．

人間の行動を知的に撮影する: 顔や手先など，撮影の主対象となる部分を複数のカメラで常に追跡して，いつでも映像として利用できる状態にする自動化撮影機能． **何をどのように伝えるか**という目的と**カメラの自動制御アルゴリズムやそのパラメータ**との関係を探り，わかりやすく不快感がない映像を取得する．

人間の行動を認識して映像を知的に編集する: 人間の行動 (ここではプレゼンテーションを対象) において，重要な意味を持ち，注目する必要がある場面や部分を検出する機能．注目すべき部分は，時間的・空間的に常に変化するため，**人間の行動** (**体の動きや発話等**) **を利用して，もっとも見せたい部分を検出する**ことが重要なポイントである．

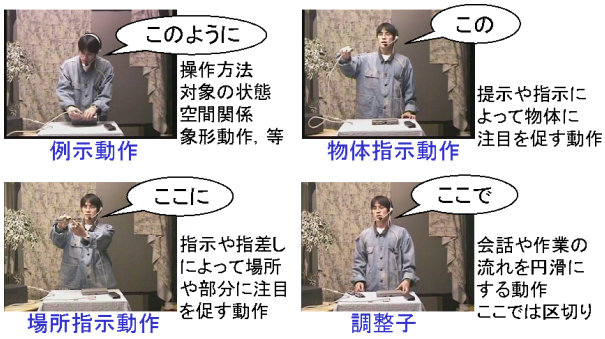

図 5: 注目を要求する動作

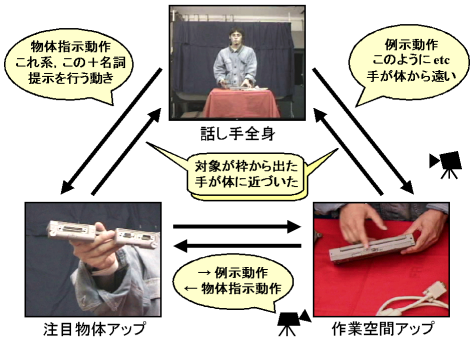

図 6: 使用したショットとその切り替え条件

我々の構築したシステムでは，位置センサや画像処理により話し手や特定物などの位置を取得し，複数台の首振りカメラを制御することで自動撮影を行う．図 4 のように，撮影の対象と目的を簡単に指定することで，その目的にあったカメラワークで首振りカメラが動作する．各々のカメラで撮影された映像は，MPEG エンコーダを通して保存され，ランダムアクセスが可能になる．また同時に，位置センサと音声認識を併用して話し手の動作認識を行い，映像へのタグ付けを行う．

このデータを用いて，図 5 のように，話し手が注意を促している動作を検出し，それを基にして視聴者が見たいと思う部分を効果的に提示することができる．一連の映像として提示する場合の編集規則例を図 6 に示す．これらのしくみを使って実際に撮影されて編集された映像の例を図 7 にあげる．静止画ではわか

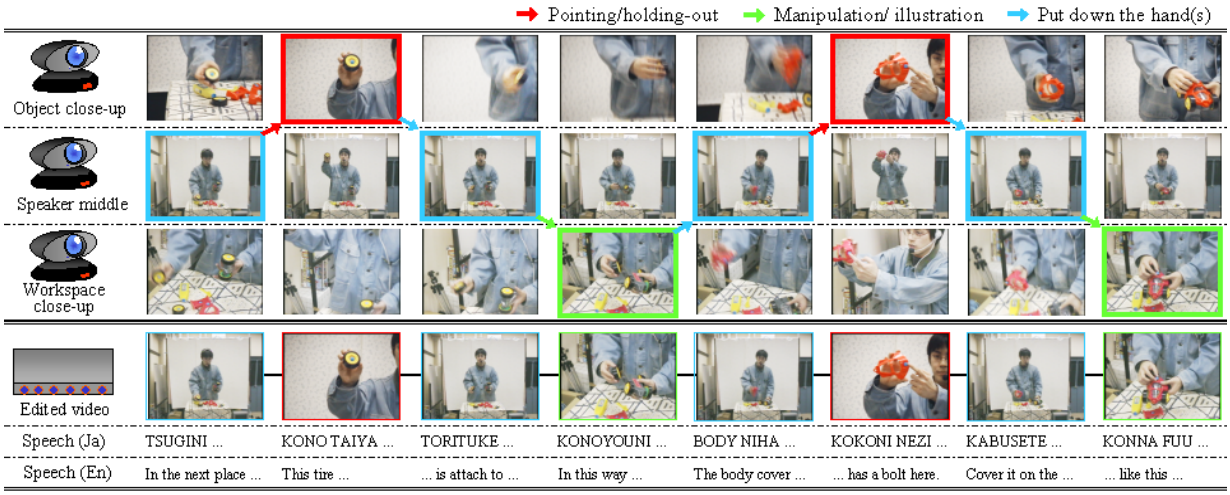

図 7: 自動撮影・編集で結構良い映像が得られる

図 8: まずどうすれば良いの？

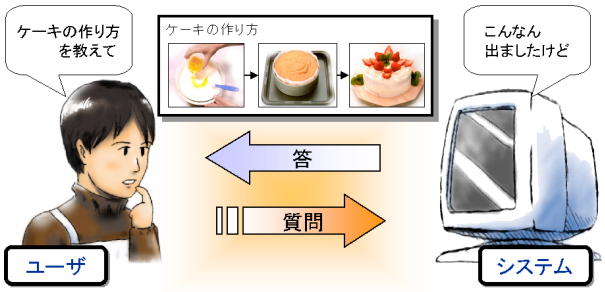

図 9: 質問に答えてくれる映像メディア

りにくいが，カメラの切り替えを含め，かなり自然な映像が得られている．これからの応用的な展開を探っている段階である．

4 対話的映像メディア: 教えてくれるメディア

料理や機械の組み立てのような作業を行うことを考えてみよう．図 8 のような状況で人間の先生に質問したならば，言葉だけではなく写真を見せる，図を描く，実演を行うなどしてわかりやすく教えてくれるはずである．映像のように複数のモダリティを持つメディアを駆使して，例えば，図 9 に示すように，できるだけわかりやすい方法でユーザに答や実例を提示することはできないだろうか．

関連する研究としては，自然言語による知的ヘルプシステムや質問応答システムに関する研究が数多く報告されてきた．しかし，マルチモーダルデータの扱いには特有の問題があり，それらの手法を単純に適用することはできない．そこで，我々は新しい枠組み *QUEVICO*[2] を提案している．

この枠組みのポイントは以下のようになっている．

「質問と答」からのインデキシング: 作業などの映像に対して様々な質問を想定し，答となる部分をマークアップするために必要となるタグセットを設定した．これによって，質問が「要求した情報」を含むデータ断片を選択する．

複数モダリティの効果的な利用: 複数のモダリティを用いて，質問の答として最も適切な提示方法を選択する手法，及び，十分なインデックス (タグ)

[2]QUEstion-based Video COmposiotion: この名は，古事記に現れる久延毘古神を典拠とする．久延毘古神は，案山子の姿をしており世の様々な出来事に熟知しているとされる．

が与えられていない場合でもそれなりに答える手法を提案した．

2 番目の項目は以下のような考え方に基づく．質問が行われた場合，人間はまず，質問のタイプ (Q) からその質問が要求する情報 (A) が何であるのかを推測し，何が (F) その要求される情報を提供するのかを考え，それが実際に含まれているデータ断片 (D) を探し出すという三段階の経路を経ることで，答となるデータ断片を求める．これをモデル化したのが図 10 の経路モデルである．十分なインデックスが (タグ) 与えられれば，これで質問に答えることができる．しかし，多くの場合には完全なインデックスを付与することは難しい．そのため，複数の要求される情報や答の形態，データ断片の関連性を考え，図 11 に示されるように多対多のリンクにより経路モデルを考えることにより，「質問」と「答となるデータ断片」をつなぐ．

QUEVICO システムの概要は以下のようになっている．図 12 に示したような多視点の映像データが複数台のカメラによって撮影され，QUEVICO で定めたタグセットによりマークアップされて，未編集のまま蓄えられる．システムは，ユーザとの対話を通じて提供すべき情報を推定し，図 13 に示すように返答する．例えば，ユーザがかつおを切り身にする作業において「どの程度切るのですか」と質問した場合，システムは「程度」を説明する映像断片と「1cm 程度の厚さにスライスする」という言語的な説明をユーザに提示する．現在の仕様では，表 1 を含む 30 種類程度の質問に対して，複数のモダリティを有効に利用してユーザに答えることができる．

5 マルチモーダルな認識: インデキシングと対話の道具

2 節，3 節で述べたように，画像，言語，音声の統合的な処理が様々な形で映像やマルチメディア処理に使われている．その基本的な考え方は，各モダリティに固有の特徴を集めること，各モダリティでの認識結果の曖昧性を解消すること，他モダリティの助けにより重要部分を判定すること等である．以下では，それに関する我々の新しい取り組みをあげ，議論の種としたい．

4 節で少し述べたように，我々は，工作，料理，電気製品の使い方，スポーツ習得などの映像に対して，

表 1: 質問タイプと要求される情報の例 (抜粋)

質問タイプ	要求された情報
～とはどのような作業ですか	説明，方法
誰が～しているのですか	動作主，方法，程度
何を～するのですか	対象，入力
～するには何が必要ですか	入力，道具
～したらどうなりますか	出力，終点
何を使えばいいのですか	入力，道具，方法，程度，量
どこで～しているのですか	場所，始点，終点，方法
～で使うものはどこにありますか	始点，場所
どこに～すればいいですか	終点
いつ～しますか	時間

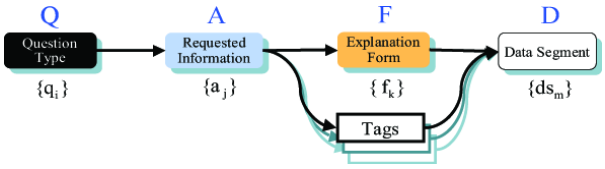

図 10: 経路モデルの概要

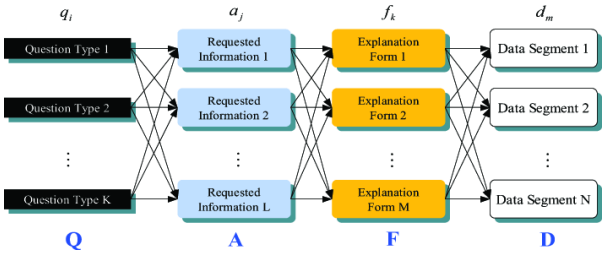

図 11: 多対多による各要素間の関係

ただ単に検索のためのインデクスを付けるだけではなく，それを基に指導をしてくれるメディアの実現をめざしている．そこで重要となるのが，教示者やユーザ

図 12: 料理映像「かつおのたたきの調理」

図 13: QUEVICO はこんな答を返す

の状況や話題の対象を推定する機能である．例えば，現在の説明がどの作業/物体に対するものか，ユーザがどの作業/物体について質問しているのかを的確に

こんにちは。
二十分で晩ご飯です。
今日は（えー）初夏にふさわしいお献立を御紹介したいと思います。
では、今日のお料理です。
（え）まず、豚肉の唐揚げ香味ソース添え。
さっぱりとしたお味になります。

…… 中略 ……

そして、戻し汁はスープに使いますからとっておいて下さい。
それからキュウリですね。
スープに入れるキュウリですけれども、これは皮を剥いて使っていきますので、上下ですね（えー）剥いていきましょう。
全部、上下を通して（えー）皮剥きで全部皮を剥いていきます。
（えーと）夏になってね、（あのー）たくさんキュウリが出回るようになりましたらね、（あの）こういつも生で食べるだけじゃなくて、このようにちょっとスープに入れてね、煮たりするとおいしいと思います。
そしてキュウリは（ま）大体椎茸に合わせて縦半分に切ります。

……

図 14: 料理教示発話の例（NHK「きょうの料理」）

判断する必要がある．

まず，そのために有効なのが，専門家，教師などの発話を高度に解析することである．黒橋 (東大) らの研究[3]により，図 14 の発話文が構造化された例を図 15 に示す [21]．発話文に多く現れる「省略」が補われるとともに，各々の発話文の役割と文間の結束性が求まっている．このような構造を基に，タスク[4]を基に質問応答や物体の認識を行うことによって，より適切な指導やアドバイスが可能になる．

そのための画像認識側の処理として，我々は 3 節で述べたシステムを拡張し，図 16 に示すような物体追跡の機能を持たせた．このシステムでは，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．その結果，図 17 の右図のように，複数の人物が動いているような場合でも，良い精度で目的の物体を検出・追跡することが可能になった．

さらに，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

[3]学術創成研究プロジェクト [22] として我々と共同研究を行っており，これまで述べてきた我々のシステムとの統合を予定している．

[4]ここでは詳しく述べないが，小作業から大目的まで粒度に階層を持つ

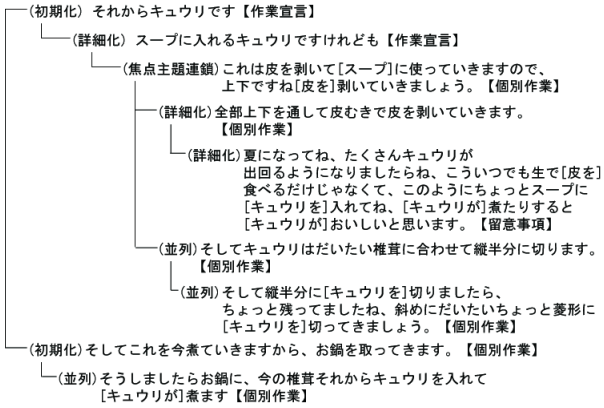

図 15: 料理教示発話の構造解析結果

観，位置，付加された注釈情報などを関連付けて映像へのインデキシングを行うことができる．得られた映像の例を図 18 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→注釈映像との関連付け」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．また，注釈情報として記録される映像は，動作認識をトリガとして適宜最適と思われるカメラからの映像に切り替えられている．注釈映像の区間は，物体が持ち上げられた時点から，説明が終了して再び物体が置かれたところまでを 1 クリップとした．

将来的に，ユーザが行っている作業や，ユーザの状態などを認識することによって，**ユーザの状態に合わせて説明してくれる**メディアを実現するのが目標である．

6 おわりに

本稿では映像処理におけるメディア/モダリティ統合に関して，我々の研究を紹介してきたが，方向性と事例を中心に紹介し，モダリティ統合に関するつっこんだ議論は省かせて頂いた．最初にも少し述べたが，これらの本質についてはまだ十分に整理されていないのが現状であり，これからの事例蓄積，問題の整理等が望まれる．これらの研究には多様な技術が必要であり，横断的な研究協力が必要であることもその特徴である．既に述べたように，筆者らは種々の蓄積型メディアや実時間型情報源とのインタラクションめざし，その要素的な研究を進めている段階であるが，広く外部の研究者との交流や研究協力を希望している．

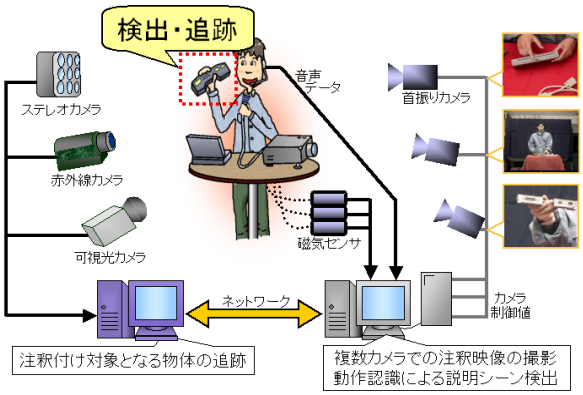

図 16: 話題の対象となっている物体を追跡する

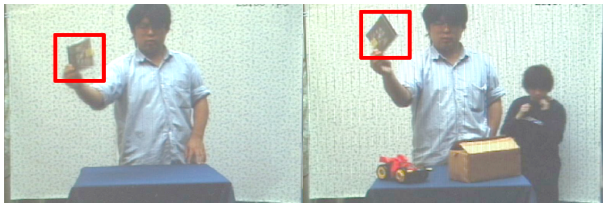

図 17: 把持物体検出結果 (左: 簡単な背景，右: 複雑な背景)

また，ここでは紹介しなかったモダリティ変換 (例えば，文章や映像の図的表現 [23][24]) や，個人行動記録 (例えば，[15])，遠隔通信，映像編集等の問題にも種々のモダリティ統合が必要であり，多くの興味深い研究テーマが存在する．これからの研究の進展が期待される．

参考文献

[1] 中村, 向川: 画像・映像の知的生成と編集 - CV 技術を用いた新しい画像・映像処理. 松山, 久野, 井宮編:『コンピュータビジョン：技術評論と将来展望』第 17 章, 新日本コミュニケーションズ, 1998

[2] 有木: メディア解析から見たパターン認識. 信学技報 PRMU99-171, 1999.

[3] 中村, 外村: 見たい部分を簡単に短時間で— 気の利いた映像メディア技術を目指して—. 信学誌, Vol. 82, No. 4, 1999.

[4] 外村, 谷口, 阿久津: PaperVideo:紙を用いた新しい映像インタフェース. IEICE,IE94-59, 1994.

[5] Taniguchi, Y., Akutsu, A., Tonomura, Y.: "PanoramaExcerpts:Extracting and Packing Panoramas for Video Browsing," ACM Multimedia97, 1997.

[6] Wactlar, H., Kanade, T., Smith, M., and Stevens, S.: Intelligent Access to Digital Video:

図 18: 物体像と説明を関連づける

The Informedia Project. *IEEE Computer*, Vol. 29, No. 5, 1996.

[7] 佐藤，金出: 文字認識と異種情報の対応関係に基づいたニュース放送からの情報抽出．情処論，Vol. 49, No. 12, 1999.

[8] 有木ほか: ニュース映像中の記事に対する音声・文字・映像を用いた索引付けと分類. 信学技報, PRMU96-97, 1996.

[9] 佐藤: ドラマ映像における登場人物のアノテーションシステム．第 5 回知能情報メディアシンポジウム, 1999.

[10] Sato, S., Nakamura, Y., and Kanade, T.: Name-it: Naming and detecting faces in video by the integration of image and natural language processing. *IJCAI*, 1997.

[11] Nakamura, Y. and Kanade, T.: Semantic analysis for video contents extraction — spotting by association in news video. *ACM Multimedia*, 1997.

[12] Smith, M. and Kanade, T.: Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques. *IEEE CVPR*, 1997.

[13] 上田: コンピュータを駆使した最新の放送番組製作技術. 情報処理, Vol. 40, No. 11, 1999.

[14] Informedia Experience on Demand. “http://www.informedia.cs.cmu.edu/eod/”

[15] S.Kubota, Y.Nakamura, Y.Ohta: Detecting Scenes of Attention from Personal View Records – Motion estimation improvements and cooperative use of a surveillance camera, Proc. IAPR Workshop on Machine Vision and Applications, pp.209-213, 2002

[16] 中村: コミュニケーションのための画像・映像処理. 信学技報, PRMU99-252, 2000

[17] M.Ozeki, Y.Nakamura, Y.Ohta: Camerawork for Intelligent Video Production —-Capturing Desktop Manipulations, Proc. Int. Conf. on Multimedia and Expo, pp.41–44, CD–ROM TA1.5, 2001

[18] M. Ozeki, M. Itoh, Y. Nakamura, and Y. Ohta: “Tracking hands and objects for an intelligent video production system,” Proc. Int. Conf. on Pattern Recognition, pp.1011–1014, 2002.

[19] H.Izuno, Y.Nakamura, Y.Ohta: QUEVICO: A Framework for Video-based Interactive Media, Int’l Workshop on Intelligent Media Technology for Communicative Reality, pp.6-11, 2002

[20] M. Ozeki, Y. Nakamura, and Y. Ohta. “Human behavior recognition for an intelligent video production system,” IEEE Proc. Pacific-Rim Conference on Multimedia, pp.1153–1160, 2002.

[21] 西田ほか: 料理教示発話の構造解析, 言語処理学会 第 9 回年次大会, (2003.3 発表予定)

[22] 西田豊明 (研究代表): 「人間同士の自然なコミュニケーションを支援する知能メディア技術」科学研究補助金，学術創成研究，研究成果報告書，2002 年 3 月

[23] 村山，中村，大田: DocScape: 文章の概覧性を高めるための概念図の生成と利用情処論, Vol.44, No.4, 2003 (掲載予定)

[24] 村山，伊津野，中村，大田: ビデオアイコンダイアグラムによる映像内容の構造表現，信学技報 PRMU2001-45, pp.47-54, 2001