会話ができる映像コンテンツを撮る・つくる

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)

概 要

本稿では，映像の撮影から編集までを一貫して扱う自動化システムの実現，及び，得られた映像を用いた会話のできる映像コンテンツの実現に向けた我々の取り組みを紹介する．対象としては，ビデオマニュアルなどを想定した，教示映像の取得と対話的提示，会話のできる議事録を想定した，会話シーンの自動撮影とその自動編集などを手がけている．その際に，単に映像を取得する自動カメラシステムを構築するだけでなく,「効果的に情報を提示するために，どのような映像部品 (単語) を用意し，どのような映像編集規則 (文法) にしたがって提示すればよいか」という問題を実証的に探っている．これらの研究を基に，人間どうし，人間対コンピュータ，種々の形態で会話のできる映像コンテンツを撮る，つくることをめざしている．

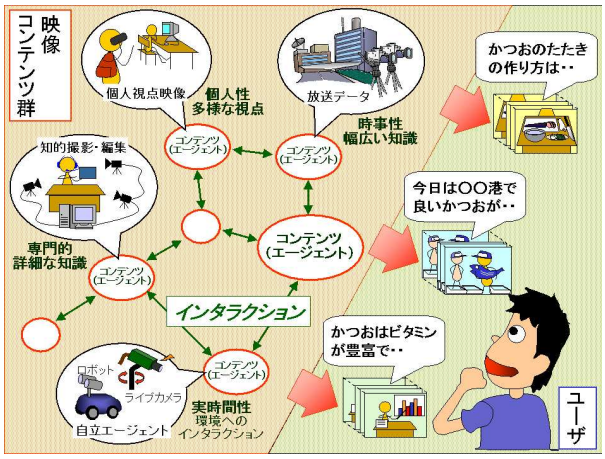


図 1: 映像・マルチメディアで何ができるだろうか？(ユーザが「かつおのたたきが食べたいなあ」と言った場面を想定)

1 はじめに

映像メディアを用いると，多くの情報をわかりやすく提示できる．その反面，映像を撮影すること，編集して他人にわかりやすく見せることには多くの労力を必要とする．また，非専門家にとっては，その方法すら十分にわからない場合が多い．この問題を解決するために，我々のグループでは，自動映像撮影・編集を行う知的システムを構築しながら，映像の新しい利用形態を探っている．

その一つの将来像に，ユーザの質問や行動に合わせて映像を提示する「会話のできる映像メディア」がある．図 1 は，このような映像メディアに関する，我々の過去，現在の研究スコープを広くまとめたものである．これは，ユーザの問いかけに対して，様々な映像コンテンツや実世界の映像ソースから適切なものを選んで返すものである．これは質問応答の発展形であるが，このような機能をさらに発展させることにより，幅広い対話・会話のできる映像メディアを実現することをめざしている．

映像の自動取得，自動インデキシングなどの要素技術が揃えば，このようなメディアは様々な分野，例えば，教材，宣伝・商品説明，作業マニュアル，遠隔コミュニケーション，個人の映像日記で用いることのできるものとなる．

本稿では，このような目的に対して我々が行ってきた要素研究をいくつか紹介する．

教示映像の取得: 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，自動撮影と自動編集を行うシステムを構築している．これにより，映像マニュアルや，遠隔通信による教示を行う環境を実現する．

会話シーンの撮影と編集: 複数人が部屋で自由に会話する場合の自動撮影とその記録が行えるように，カメラシステム，編集システム等を構築している．現在，部品となる要素技術がほぼできあがり，全体として一貫した動作をすることを確認している段階である．

対話型映像メディア: 我々が提案してきた対話型映像メディアは，質問やユーザの行動に応じて適切

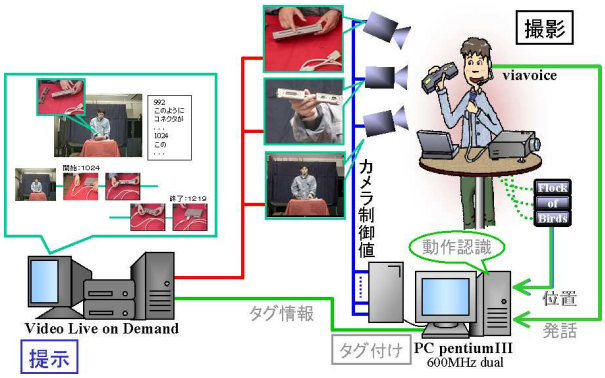


図 2: 自動的に「撮って編集する」システム

図 3: カメラ設定の選択表 (一部分)

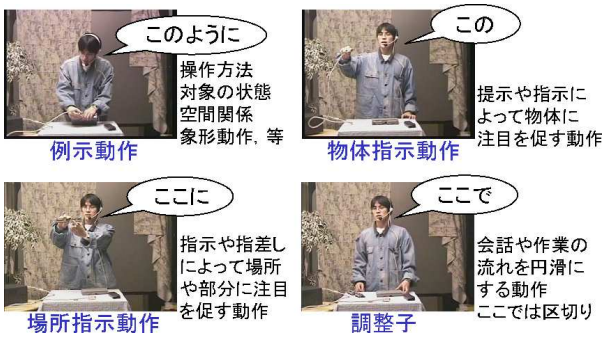


図 4: 注目を要求する動作

な映像データを編集しながら提示する．そのために，上記 2 つの知見と技術を用いてコンテンツを取得し，要求に応じた編集を行ってユーザに提示する．

2 教示映像の取得: 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，著作権や肖像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 2 のようなシステムを構築している [4][5]．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してきた．

カメラマンの機能: 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．

ディレクターの機能: 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．

我々の構築したシステムでは，位置センサや画像処

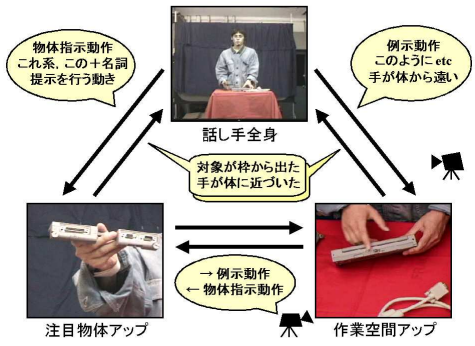


図 5: 使用したショットとその切り替え条件

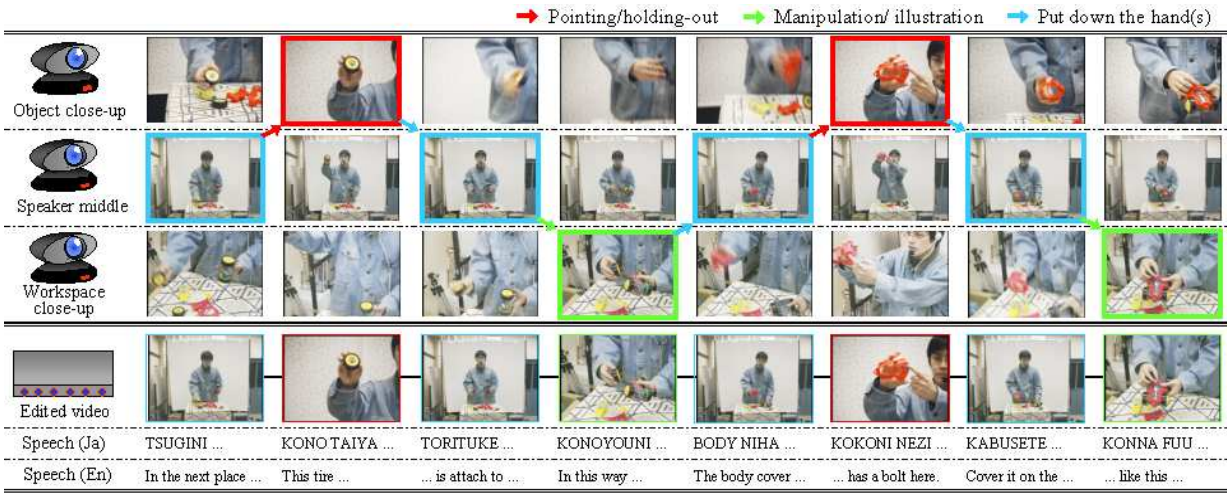


図 6: 自動撮影・編集で結構良い映像が得られる

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 3 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 4 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 5 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 6 にあげる．静止画ではわかりにくいが，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．

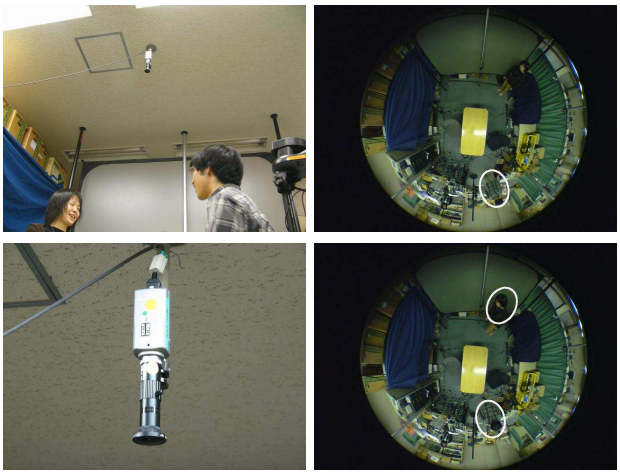
3　会話シーンの自動撮影と編集

2 人から数人の人が会話するシーンは映画や TV 番組に盛んに出てくるが，それらはドラマやドキュメンタリーの内容を表現するために重要な役割を果たしているため，様々なテクニックを用いて撮影，編集されている．このような会話シーンにおける映像取得・提示の問題を扱うことは，映像文化的に興味深い知見を得られるだけでなく，映像メディアの応用範囲を大きく広げることになる．例えば，会議やミーティングの議事録を映像として残したいという需要は大きいが，これらは身の回りで日常的に発生するため，専門家を雇って撮影，編集するほどコストはかけられない．しかし，単純な固定カメラからの記録・提示では，見辛いだけでなく，コミュニケーションの様子もわかりにくい．そのため，質の良い映像を撮影・提示できる自動化システムが必要となる．また，会話形式を用いて情報提示を行うことの有効性も指摘されており [8]，このような意味でも，会話を詳細に撮影，蓄積，提示する技術が必要とされている．

3.1　会話シーンの自動撮影

我々は図 7 のような撮影システムを構築している．このシステムは，人物の検出・追跡と大まかな向き推定を行う「環境カメラ (群)」と，構図を検証しながら個々の映像を取得する「コンテンツ撮影カメラ (群)」からなる．環境カメラ群，コンテンツ撮影カメラ群の説明を表表 1 に示す．このような構成になった理由は，人物の室内移動を可能にするため広い範囲をカバーしなければならないこと，映像の質を保つために，コンテンツ撮影カメラを頻繁かつ速く動かすことができないことから，観測とコンテンツ撮影が両立しないためである．

動作の概略は以下のようになる．まず，環境カメラ群により，2 人から数人の人間が部屋に入って，適当な場所で会話を始めるまでを図 8 に示すように大まかに追跡する．着座などにより人間の居場所がほぼ固定された段階でコンテンツ撮影カメラ群に撮影要求を出す．カメラと人間の位置関係，撮影目的などが与えられれば，どのコンテンツ撮影カメラがどのようなショットを撮るべきかがわかるため，各々の指示を与えることができる．この指示により，個々のコンテンツ撮影カメラが被写体を要求された構図で捉え，



環境カメラ　　追跡例

図 8: 環境カメラによる人物追跡

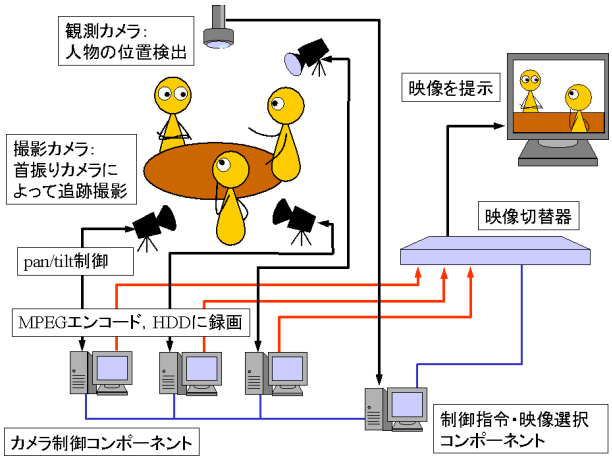


図 7: 複数カメラによる会話シーンの撮影

表 1: 撮影カメラ群の機能

	ハードウェア	機能等
環境カメラ	広視野・高解像度 (SXGA, IEEE1394, 魚眼レンズ), 天井に設置	適応的差分と EM アルゴリズムによる人物追跡，構図の選択，その他
コンテンツ撮影カメラ	pan/tilt カメラ (NTSC), MPEG1/2 エンコーダ	顔検出，顔の向き推定，枠制御アルゴリズムによる構図維持，その他

図 9 に示すように，構図を保ちながら撮影する．その際に，我々が提案している枠制御アルゴリズム [4] を用いることにより，視野を調整しながら質の良い映像を得る．ただし，これらは質の良い映像を撮影するために必要最小限の構成になっており，複雑な分散協調

撮影された映像　　カメラの首振り角

図 9: コンテンツ撮影カメラによる構図の維持 (登場人物が動いたため，カメラの首振り角が変わっているが (右)，構図は維持されている (左))

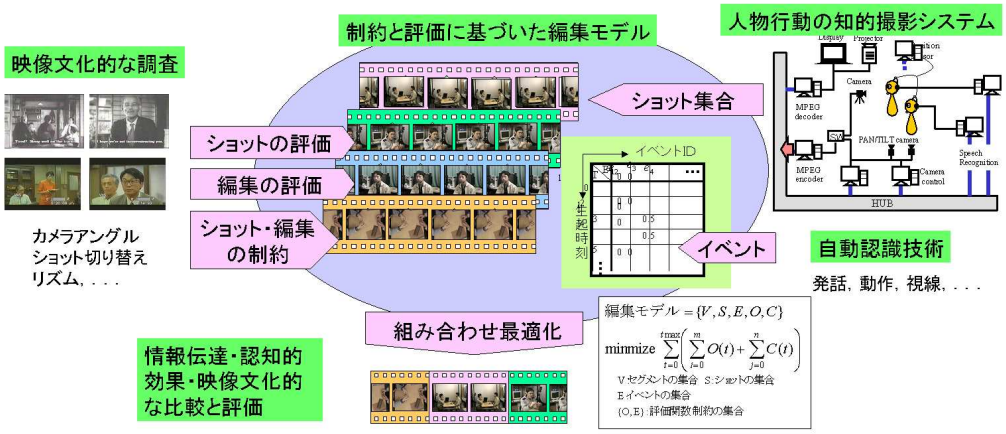
処理などは行わない．

以上のように，個々の要素は従来の研究でも見られるものであるが，従来のシステムには，複数人が部屋に入って座り，会話を始めるまでを連続的にカバーできるものはなく，我々が新しいシステムを構築する必要があった．このようなシステムが全体として動作して始めて，会話シーンを十分に撮影することが可能になる．

3.2　会話シーンの自動編集

この研究での編集における目的は 2 つある．映画や TV 番組に見られるような編集を自動化すること，及び，映像を構成するための映像文法を探ることである．ただし，普遍的なシステムや文法を構築することをはじめから目指すのではなく，簡単なモデルから始め，徐々にそのモデルを精密化していくこととする．そのためのモデルとして我々は，図 10 に示したような，制約充足・最適化に基づく編集モデルを提案してきた [9, 10]．この計算モデルは 5 つの要素で定義されるが，評価関数や制約が編集規則に対応する．これらの規則を種々設定し，パラメータを変えながら編集実験を行うことにより，編集の問題を探る．

このような編集規則を探るために，我々は映画の典型的なシーンを対象とした編集実験を行っている．映画から選んだシーンを模擬して撮影を行い，編集規則やパラメータを変えながら編集し，その効果を確認するものである．その概略を図 11 に示す．ここでは，映画の一シーンに似せるための編集規則とパラメータを考えた場合，異なった目的を考えた場合，違う設定のカメラ群が存在する場合，等の設定を試した結果である．図 11(c) に示したように，多くの場合，



$$\text{Editing} = \{V, S, E, O, C\}$$

ショット集合 (S)	ショット (s_i) の集合
ビデオシーケンス (V)	単位時間の長さを持つビデオセグメント $v(t)$, $(t = 0 \sim t_{max})$ の並びからなる，ビデオシーケンス
イベント集合 (E)	対象シーン中で起きているイベント (e_i) の集合
評価関数 (O)	各ショット (またはショット列) の良さを計算するための評価関数 (o_i) の集合
制約 (C)	ショットの並び方やつなぎ方に関する制約の集合

図 10: 制約充足と最適化に基づく編集モデルの概要

単純な編集規則では映画と同じ編集パターンにならないことがわかる．これは我々が見落としている要素が存在するかもしれないこと，また，監督の目的 (評価の重み付け) が我々の予想とかなり違うかもしれないことを示している．

また，映像として強調したいことと編集のための評価関数や制約には深い関係があり，Film Studies の分野では様々な指摘がなされてきた．例えば，クロースアップショットには，注目対象を示す，力関係を示す，人物の内面を示す，注目対象以外のものを隠す，グラフィック上のリズムを生む，等の効果がある．編集目的に応じてこれらのショットやショット切り替えの評価関数を変えることにより,「伝えたい情報や効果」によって編集結果が変わる「自動編集」を実現できる．

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 12 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

図 12: まずどうすれば良いの？

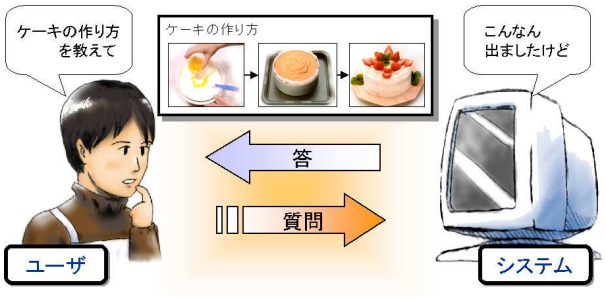


図 13: 質問に答えてくれる映像メディア

ショット 0	全体ショット
ショット 1	全体ショット (180 度逆からの)
ショット 2	婆のクロースアップ
ショット 3	婆の肩越しショット
ショット 4	女のクロースアップ
ショット 5	女の肩越しショット

(a) 利用したショット

編集例 1	使うショットは 0,2,3,4. 発話に 1 秒先行して発話者のクロースアップにスコアをつける (小), 表情が変化したらその人の肩越しショットスコアをつける (中), シーンの最初は全体ショットを使う, 聞き手の様子をクロースアップでとる, 不必要, イベントのない時のショットの切り替えを少なくする, 3 秒以下での間隔での切り替えを禁止, 切り替えを減らす (切り替え 1 回につき負スコア)
編集例 2	編集例 1 に “二人を撮すショットの割合を同じくらいにする” という評価関数を追加した例
編集例 3	編集例 1 にショット 5(女の肩越しショット) を追加, ジャンプカットの禁止という評価関数を追加

(b) 編集規則

```
-------------------------------------------------------------------------------------------------
Time      |0sec                          |30sec                          |60sec
-------------------------------------------------------------------------------------------------
編集例 0    000000000000000033333333333333333334444222222222222222222222222222244422222222222222
編集例 1    000000000000000444333333333333333444422222222222222222222222222222244422222222222222
編集例 2    000000000000000333333333444444444222222222222222222444222222222222444444444422222222
編集例 3    000000000000000333333333333333333000222222222222222225555552222222555555222222222222
-------------------------------------------------------------------------------------------------
婆の発話, 婆の表情の変化     <-->          <->     <-->     <----------> <-----> <----------->
女の発話, お辞儀       <-->                  <->
-------------------------------------------------------------------------------------------------
```

(c) 100 秒間 (表示は 90 秒間) を編集した結果. 編集例 0 が作品 (映画監督によるもの) であり, 編集例 1〜3 が我々のシステムで編集した結果. 数字列はショットの番号を表す. 下は重要なイベントの生起時刻

(d) 映画の一シーンと模擬した撮影から画像を抜き出した例

図 11: 撮影データと編集例 (サンダカン八番娼館の一場面 (編集例 0) を模擬)

ディアを駆使して, 例えば, 図 13 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*[1] 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

[1] 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.

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) 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 14 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 15 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 16 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 17 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 2 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　おわりに

本稿では，映像の撮影から編集までを一貫して扱う自動化システムの実現，及び，得られた映像を用いて会話のできる映像コンテンツの実現に向けた我々の取り組みを紹介した．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の実現をめざし，その要素的な研究を進めている段階であるが，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．広く外部の研究者との交流や研究協力を希望している．

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 [12][13]) や，個人行動記録 (例えば，[14])，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

表 2: 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

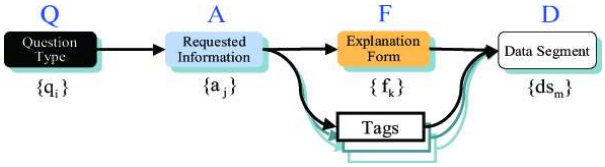


図 14: 経路モデルの概要

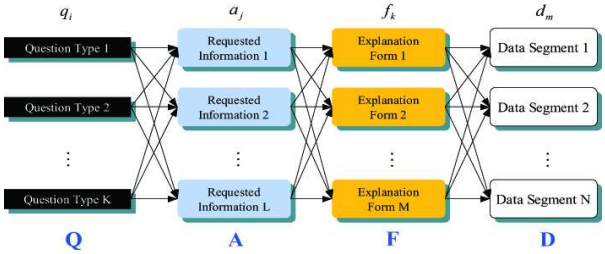


図 15: 多対多による各要素間の関係

図 16: 料理映像「かつおのたたきの調理」

図 17: QUEVICO はこんな答を返す

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, 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.

[2] 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

[3] 中村: 画像・映像の撮影・編集・提示から対話的映像メディアまで, 言語処理学会第 9 回年次大会併設ワークショップ「メディア/モダリティ統合における言語処理」, 2003

[4] 尾関基行, 中村裕一, 大田友一, ”机上作業シーンの自動撮影のためのカメラワーク”, 電子情報通信学会論文誌, Vol.DII-J86, No.11, pp.1606-1617, 2003

[5] 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.

[6] 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.

[7] 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

[8] 西田豊明 (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”, 平成 14 年度科学研究費補助金 (学術創成研究) 研究成果報告書, 2002, 同 15 年度報告書, 2003

[9] 尾形涼, 尾関基行, 中村裕一, 大田友一, ”制約と評価関数に基づいた映像編集モデル”, 信学技報, PRMU-2003-46, pp.13-18, 2003

[10] R.Ogata, Y.Nakamura, Y.Ohta, ”Computational Video Editing Model based on Optimization with Constraint-Satisfaction”, Proc. Fourth Pacific-Rim Conference on Multimedia, CD-ROM 出版, 2003

[11] 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

[12] 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, pp.1150-1162, 2003

[13] 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

[14] 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