

# 机上作業映像のためのイベント駆動型自動編集手法

## 注目喚起行動による映像編集の有効性とその検証

尾関 基行<sup>†</sup> 中村 裕一<sup>†</sup> 大田 友一<sup>†</sup>

<sup>†</sup> 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: †{ozeke,yuichi,ohta}@image.esys.tsukuba.ac.jp

あらまし 料理や組み立て作業などの机上作業プレゼンテーションでは、重要な箇所や作業を聞き手（視聴者）に見てもらうために、話者は視聴者の注目を誘導するような行動（注目喚起行動）を行う。本研究では、このような話者の注目喚起行動に基づいて映像を編集することについて、その適切さや有効性をテレビ映像と比較することで検証する。本稿では、まず机上作業プレゼンテーションにおける典型的な注目喚起行動についてまとめ、テレビ番組におけるショット切り替えとの関連性を調べた。また注目喚起行動とテレビ番組のそれぞれに基づいた編集映像を用いて主観評価実験を行うことにより、注目喚起行動をショット切り替えのトリガとすることの利点と欠点を示した。最後に注目喚起行動の検出方法とその結果を用いた自動編集ルールを提案し、ユーザ評価実験によってこれらの有効性を示す。キーワード 自動編集、動作認識、発話処理、映像制作、e-Learning

## Event-driven Automatic Editing for Videos on Desktop Manipulations

### — Effectiveness of Editing based on Behaviors of Drawing Attention —

Motoyuki OZEKI<sup>†</sup>, Yuichi NAKAMURA<sup>†</sup>, and Yuichi OHTA<sup>†</sup>

<sup>†</sup> IEMS, University of Tsukuba Tenriodai 1-1-1, Tsukuba, Ibaraki, 305-8573, Japan

E-mail: †{ozeke,yuichi,ohta}@image.esys.tsukuba.ac.jp

**Abstract** In desktop manipulations such as cooking or assembly works, speakers usually draw, by typical behaviors, the audiences' attention to important portions. We verified the appropriateness and effectiveness of the video editing based on those behaviors in terms of statistics and comparisons with TV programs. In this paper, we first discuss significant types of human behaviors that commonly appear in presentations, and their usefulness as the triggers for camera (view) switching in TV programs. We show the advantages and drawbacks of this scheme by subjective evaluations. Then, we propose a simple and robust method for recognizing those behaviors, and demonstrate the efficacy by actual automatic editing on our video production system.

**Key words** Automatic editing, behavior recognition, speech processing, video production, e-Learning

### 1. はじめに

マルチメディア技術やネットワークの発達により、大学や予備校の講義を撮影して遠隔地に伝送したり、データベースに蓄積することが一般的になってきた。そのための撮影にはカメラマンやディレクターを雇う必要があるなどコストが大きくなるため、撮影を自動化する研究が盛んに行われている。

一方で、料理や科学実験のような机の上で行う作業（机上作業シーン）の撮影に対しても自動化へのニーズは高い。たとえば机上作業シーンを撮影して伝送することによって、実験設備の揃わない場所でも最新の環境で講義することができる。また

料理や器材の組み立て方法などを映像に記録することによって、映像マニュアルとして利用することもできる。このような背景から、我々は机上作業シーンを対象として、カメラマンとディレクターの機能を代替する自動撮影・編集システムを構築してきた。

システムの概要を図1に示す。人物の手や上半身、物体などを複数台のPan/Tiltカメラで追跡しながら撮影する。撮影対象の位置は、人物の手や腰に装着した磁気センサによって計測されたデータを元に計算する。各カメラで撮影された映像はすべてMPEGエンコーダを通してPCのハードディスクに蓄積する。撮影中、話者はすべての映像（各カメラで撮影された映

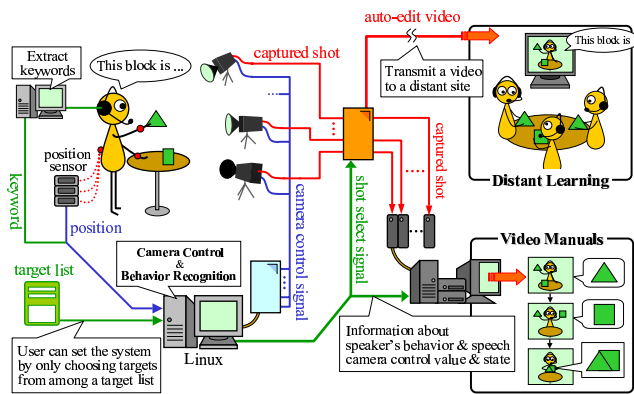


図 1 自動撮影システムの概要

像およびオンライン編集された映像)をディスプレイで確認しながらプレゼンテーションすることができる。映像や位置データの他、音声認識ソフト<sup>(注1)</sup>を用いて抽出した話者の発話やカメラの制御状態(固定中/追跡中)などを同期して記録する。これらの映像と付加情報は、そのまま映像マニュアルとして利用したり、オンラインで編集して遠隔講義に利用することを想定している。

関連研究としては、スライドを用いたプレゼンテーションや教室での講義を対象とした研究[1][2][3][4]が行われているが、机上作業シーンを対象としたものはこれまでほとんど研究されていない。料理シーンを対象としたものでは Bobick らの Intelligent Studios [5] があるが、彼らは実際に複数のカメラを用いたシステムを構築しておらず、また自動編集については言及していない。それに対して我々の撮影・編集システムは、これまで実際に行った多くの撮影・デモンストレーションの経験に基づいて繰り返し再設計してきたものである。

この撮影・編集システムをベースとして、撮影対象に応じたカメラワークを実現するためのカメラ制御手法や、人物の動きと発話を組み合わせた自動編集手法などを我々はこれまでに提案してきた[6][7][8]。しかし自動編集については、人物の行動に基づいた編集の有効性やその適切さについて十分に検証できていなかった。

本稿ではこの点について、まず視聴者の注目を集めるために話者が行う典型的な行動について述べ、人物行動に基づいて映像編集を行う効果について考察する。次にこれらの人物行動を動きと発話のイベントの組み合わせで検出し、その結果を用いてショットを切り替える自動編集ルールを提案する。テレビ番組映像との比較およびユーザ評価実験を通して、人物行動に基づいて映像を編集することの有効性を示す。

## 2. 机上作業映像の編集

### 2.1 映像編集と注目喚起行動

机上作業プレゼンテーションでは、重要な物体や作業を聞き手に見ってもらうために、話者は聞き手の注意をその注目箇所へ誘導するような行動をとる。これは机上作業シーンを映像に記録する場合にも基本的には同じである。しかし話者と聞き手が

(注1): 現在は IBM ViaVoice を使用。



(a) 指示 (b) 提示 (c) 例示 (d) 実演

図 2 机上作業にみられる典型的な注目喚起行動

直接対峙するのに比べ、映像では視聴者と目線を合わせたり注目箇所を視聴者のいる方向へ提示して見せることができないため、注目箇所のクローズアップショットに切り替えることでこれを補っていると考えられる。本研究の基本的なアイデアは、このような「話者が視聴者の注目を集めようとする行動(注目喚起行動)」をトリガとしてショットを切り替え、重要な情報をうまく含んだ編集を実現することである。

従来のカメラ切り替えによる自動編集の研究では、教室における講義シーンのアーカイブ・遠隔講義システムにおいて盛んに行われている。宮崎らの研究[9]では、講師の位置・講師の向き・講師の音声・生徒の動きの組み合わせで講義状況を定義し、各時点での講義状況に応じてショットを切り替える<sup>(注2)</sup>。先山らの研究[10]では、講師の位置・向き・音声などに板書や OHP に対する指示動作などを加えた 10 つの状態記述要因を定義し、これらを入力とした有限オートマトンによる送信映像選択ルールを提案している。

これらの研究に対して、本研究では注目喚起行動を編集のトリガとして用いることの良さを評価した上で、実現可能な範囲でそれを実装しているところに特徴がある。これは注目箇所 1 つ 1 つが小さいためにクローズアップで撮影される上、話者や作業機の付近に集中して注目箇所が存在するため、話者の立ち位置や動作があったか否かだけでなく、話者がどこに注目を集めようとしているのかをその行動から判断しなければならないからである。これは教室での講義シーン撮影でも、講師が教壇の上で何かを説明する場合などに必要となる考え方である。

### 2.2 注目喚起行動の種類

机上作業プレゼンテーションにみられる典型的な注目喚起行動として、以下のものが挙げられる。

指示: 図 2(a) に示すような、注目して欲しい物体や場所を手や指で差し示す動作。

提示: 図 2(b) に示すような、注目して欲しい物体を掲げる動作。

例示: 図 2(c) に示すような、重要な形や大きさ、動きなどを身振りで表現する動作。

実演: 図 2(d) に示すような、作業の一部や一連の作業について特に重要であることを強調し、実際に行ってみせる動作。

(注2): 彼らのシステムでは「カメラワーク移行期間」もショット切り替えの重要な要因としているが、本システムでは各カメラが 1 つの対象を 1 つのカメラワークで追跡撮影するため、すべてのカメラの映像は常に使える状態になっているものとする。

表 1 注目喚起行動とショット切り替えの共起数と誤り数

	指示	提示	例示	実演	呼びかけ	指示詞	動作終了
共起数	6	14	1	19	4	37	29
非共起数	26	23	3	35	1	127	-

注目喚起行動と共起したショット数/全ショット数 = 74/147

呼びかけ：「見てください」、「いかがでしょう」など、注目すべき状態であることを明示的に表す発言。

現場指示詞：「これが～です」、「この～に…」、「このように…」など、話題に挙がっている物体や状態を強調する発言。

これらの行動を明確に区別することは難しい場合もあり、また動作と発言は同時に起こることが多い。特に実演については動作のみから直接判断することは難しく、注目を喚起する発言と併せることで客観的に判断できる。

### 3. 注目喚起行動に基づいた映像編集

#### 3.1 テレビ番組におけるショット切り替えとの共起性

注目喚起行動がショットの切り替えのためのトリガとなることを確認するために、テレビ番組映像における注目喚起行動とショット切り替えの共起数を調べた。調べたテレビ番組は次の3種類である。

- 料理番組（2種類 約25分）
- 科学実験番組（1種類 約5分）
- 工作番組（1種類 約10分）

前述したように注目喚起行動を完全に区別することは難しいが、ここではこれらを客観的に数えるため、次のように注目喚起行動を定義した。

指示：手・指が明らかに注目対象を差したとき、もしくは手・指が注目対象に向いているか触れている状態で現場指示詞・呼びかけ・物体の名前のいずれかを発話したときに行動が開始されたとする。

提示：手に持った注目対象を明らかにカメラ側に掲げたとき、もしくは手で注目対象を軽く持ち上げている状態で現場指示詞・呼びかけ・物体の名前のいずれかを発話したときに行動が開始されたとする。

例示：話題の物体の大きさや形状について手振りで明らかに説明しているとき、もしくは手振りで形などを示している状態で現場指示詞を発話したときに行動が開始されたとする。

実演：作業を行っている状態で、現場指示詞・呼びかけのいずれかを発話したときに行動が開始されたとする。

また全ての注目喚起行動について、以下のいずれかに当てはまる場合は行動が終了したとみなす。

- 手の動きが止まった
- 説明が途切れた
- 手が下りた（作業領域から外れた）
- 手もしくは注目対象が画面から出た

共起数を表1に示す。結果より、注目喚起行動をすべて検出することができれば約50%のショット切り替えを扱うことがで



ショット A

ショット B

ショット C

ショット D

図 3 机上作業映像の典型的なショット

きる。一方でショット切り替えと共起しなかった数も多いが、これは一連の作業の中で注目喚起行動が連続して行われることが大きな原因であり、これらを1つの注目喚起行動群とみなすことによって有効な注目喚起行動を効率よく取り出すことが考えられる。

注目喚起行動に関係のなかったショット切り替えについては、以下のような要因がトリガとなっていたと考えられる。

- 注目を喚起するような言葉が含まれない単なる説明の始まりや終わり
- 注目を喚起するような動きが含まれない単なる動作の始まりや終わり
- 物体が単独で（特に話者の説明なく）変化する
- 話者が作業しない状態で話し続けている
- 画面内にしばらく変化がない（動きがない、動きが単調）

#### 3.2 主観評価実験

注目喚起行動のみによって編集した映像が視聴者にどのような影響を与えるかを調べるために、注目喚起行動に基づいた編集映像とテレビ番組に基づいた編集映像を主観評価実験と比較する。テレビ番組の映像は編集前の映像がないため、テレビ番組の1シーンを模擬したプレゼンテーションを我々のシステムで撮影し、テレビ番組を参考にして手動で編集したものをを用いた。

その際に用いたショットは以下のようなものである。これらは机上作業映像でよく用いられ、題材としたテレビ番組でも多用されている。図3に例を示す。

ショット A：人物と作業領域を含むミディアムショット

ショット B：人物の顔や上半身のクローズショット

ショット C：手元/物体/場所のクローズアップ

ショット A で机上作業の全体的な流れや様子を伝え、ショット C に切り替えることで注目すべき箇所に焦点をあてるのが最も基本的な編集となる。これに加えて、視聴者を飽きさせないように、また情報をより分かりやすく伝えるためにショット B が使われる。また実際には、テロップや CG で作られたショット（ショット D）もよく用いられるが、ショット切り替えと直接的



図4 テレビ番組に基づいた編集映像と注目喚起行動に基づいた編集映像の一部

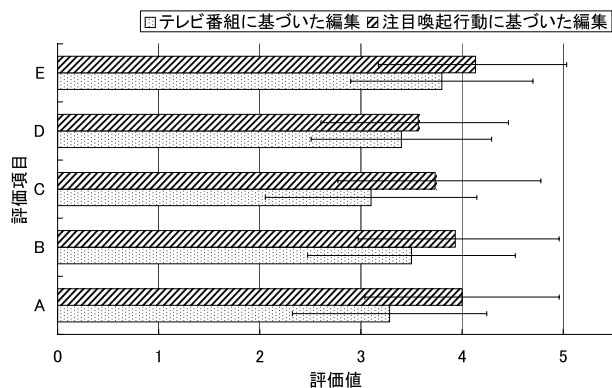


図5 テレビ番組に基づいた編集映像と注目喚起行動に基づいた編集映像の主観評価実験の結果（棒線は標準偏差を表す）

に関係しないため今回は対象外とした。

注目喚起行動に基づいた編集映像としては、前節で述べた定義に従って手で編集したものを用いた。題材として、比較的簡単に模擬しやすいと思われた「炭とアルミホイルで電池を作る科学実験<sup>(注3)</sup>」を選んだ。2つの編集映像では約70%のショットが重複した。編集映像の一部を図4に示す。

主観評価実験では、10人の大学院生に以下の5つの項目について5段階（5が最良）で評価してもらい、評価値の平均を計算した。

- あなたの見たいと思ったところが強調されていましたか？
- あなたが必要ないと思ったところが強調されていませんか？
- ショットの切り替わるタイミングは適切でしたか？
- 編集のテンポについて見ていて飽きませんでしたか？
- 編集のテンポについて見ていてめまぐるしくなかったですか？

結果を図5に示す。元のテレビ映像に基づいた編集よりも、注目喚起行動に基づいた編集のほうがすべての評価項目において良い結果となった。このような結果となったのは、以下のような理由があると考えられる。

- 今回選んだ題材では、客観的にみて重要と思われるほとんどの場面で話者が注目喚起行動を行っていた。そのため重要な注目箇所が現れたところでは両映像とも同じようにクロスショットに切り替わっており、2つの編集映像の間に

表2 注目喚起行動に関係するイベント（ゴシック体で示したイベントは動作検出に用いるもの）

イベント	内容（一例）
動きの状態	動き始めた、止まった、動き続けている
動きの内容	手を伸ばした、手を下ろした、物を持ち上げた
発話の状態	始まった、終わった、続いている、話しかけた
発話の内容	動詞、名詞、形容詞、指示詞、呼びかけ
物体の変化	発生した、外観が変化した、分割した、動いた
画面の内容	注目対象が外に出た、注目対象が隠れた <sup>(注4)</sup>

印象的な違いがなかった。

- 評価後のアンケートでは、ショット切り替えと人物行動の始まりが同期していることに視聴者が良い印象を受ける傾向がみられた。注目喚起行動に基づいた編集ではショットが話者の行動と完全に同期して切り替わるため、テレビ番組の編集に比べて評価が良くなったと考えられる。
- またアンケートより、テレビ番組の編集にみられる「視聴者が飽きない」ことを目的として挿入されたカットが、多くの被験者にとって反って「不必要な部分が強調されている」「見たい部分が途中で切り替わってしまった」と認識されていた。このようなカットの挿入はプレゼンテーションの“微妙な間”に深く関係しており、この微妙な間も含めて模擬したショットを用いなければ元映像とは印象が変わってしまう可能性がある。

この結果は、システムが注目喚起行動の開始と終了を正確に検出することができれば、場合によってはテレビ番組の編集よりも視聴者にとって評価の高くなる可能性を示しているといえる。今後はより幅広い題材を選び、テレビ番組をできるだけ正確に模擬した映像を用いて評価実験することで、注目喚起行動のみをトリガとして編集した映像の問題点を明らかにしていく予定である。

#### 4. 人物行動検出による映像編集の自動化

##### 4.1 注目喚起行動の検出と自動編集ルール

注目喚起行動に基づいた映像編集を図1の撮影システムで実際に自動化する。本研究では、1つのショットA（机上を含む範囲で話者を追跡撮影したもの）と3つのショットC（右手・左手・両手の中点を追跡撮影したもの）を用いた自動編集を考える。話者の注目喚起行動に基づいて以下を決定することにより自動編集を行う。

- ショットAからショットCに「いつ」切り替えるか
- 3つのショットCの「どれ」に切り替えるか
- ショットCからショットAに「いつ」切り替えるか

話者が制約なく自由に振る舞うことを前提とした場合、その動きのみから注目喚起行動を精度良く検出することは一般に難しい。そこで本研究では、スタジオ内で起こる比較的単純なイベントを組み合わせることで、簡単な処理で精度良く注目喚起行動を検出することを考える。

注目喚起行動に関係すると考えられるイベントを表2にまと

(注3): やってみようなんでも実験 (NHK 教育)

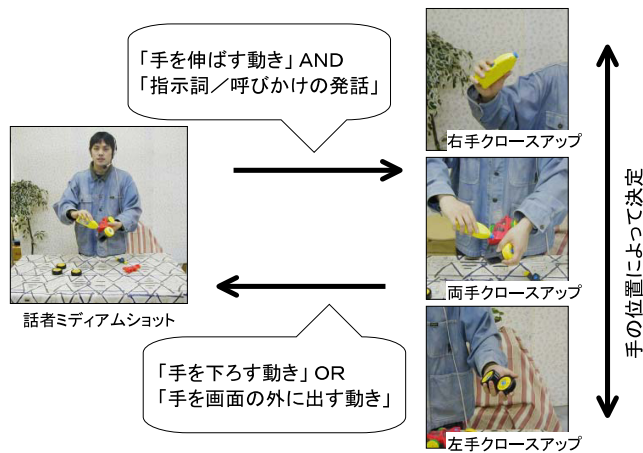


図 6 自動編集ルール

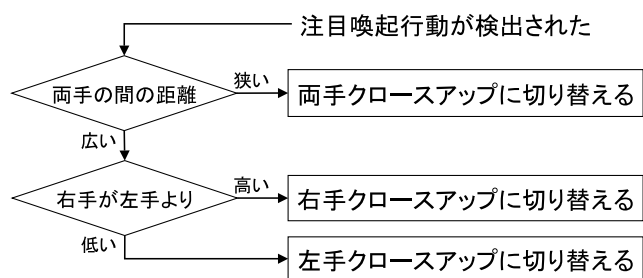


図 7 手の位置関係によるショット C の選択

めた。これらの中より、自動編集に用いるイベントを以下の 4 つの基準で選んだ。

- 机上作業の内容に依存せず、できるだけ一般的に使えること。
- 話者の自然な動作・発話をできるだけ妨げないこと。
- 話者が注目を集めたいと思ったときに意図的に使えること。
- 前後関係（コンテキスト）の解釈を必要としないこと。

上記の基準を満たすものとして、本研究では「手を伸ばす動き」・「手を下ろす動き」・「指示詞の発話」・「呼びかけの発話」・「手を画面の外に出す動き<sup>(注5)</sup>」の 5 つのイベントを選択した。「手を伸ばす動き」は注目喚起行動以外の場合にも多く出現するが、発話に比べて動きの制約は話者に大きな違和感を与えるため [7]、話者が意図的に使用でき、かつ単純なイベントとしてこれを選択した。

これらのイベントを次のように組み合わせることによって、注目喚起行動の開始と終了を検出する。

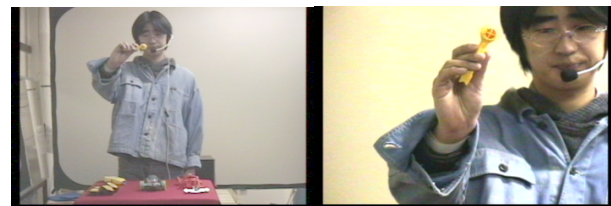
行動開始：「手を伸ばす動き」と「指示詞／呼びかけの発話」が同時に起きた

行動終了：「手を下ろす動き」か「手を画面の外に出す動き」のどちらかが起きた

話者がどこに注目を集めようとしているかは、動作開始時の

(注4)：スタジオでは、オンライン編集の結果を話者は確認しながらプレゼンテーションすることができるため、「画面の内容」もイベントとして含めることができる。

(注5)：多くの場合、注目部分は話者の手によって移動・操作されているため、本研究では「手を画面の外に出す動き」をいうイベントを扱う。



このレンチを組み立てに使用します。



このように上にして...



こうやってカチカチと鳴るまで...

図 8 車の模型を組み立てる作業のプレゼンテーション

手の位置関係によって推測する。自動編集ルールを図 6 に、手の位置関係によるショット C の選択処理を図 7 に示す。

#### 4.2 撮影実験

提案手法による注目喚起行動の検出率と自動編集ルールの使用感を調べるために、6 人の大学院生に対して「車の模型を組み立てる作業（2 分程度）」についてプレゼンテーションしてもらった。被験者は普段から人前で説明する機会を持つわけではなく、机上作業のプレゼンテーションについては初心者であるといえる。

各被験者にはプレゼンテーションを行う前にシステムの動作原理と自動編集ルールについて説明し、注目喚起行動の検出方法を口頭で教えた。机上作業の内容については、「車の模型を組み立てる作業」の手順を図と文字で簡単に記述した模造紙を用意し、スタジオの前の被験者から見える位置に掲示した。スタジオにはオンライン編集の結果を表示するディスプレイが設置されており、被験者（話者）は自分の注目喚起行動が正しくシステムに検出されたか否かをその場で知ることができる。撮影の様子を図 8 に示す。

実験の結果、注目喚起行動の検出率が約 75%、検出されたもののうち約 95% で正しくショット C が選択された（誤検出はすべて音声認識の失敗によるものである）。このように自動編集ルールについて一度説明しただけで、全被験者とも注目を集めたい箇所をショットで強調するように注目喚起行動を行いつつプレゼンテーションすることができていた。評価実験の後にを行ったアンケートの結果でも、自動編集ルールの使用感（システムの使用感）について概ね良い評価を受けた。ほとんどの被験者は自動編集ルールについて拘束感はあまり感じられなかったと答え、編集結果にも満足しているとの感想を得た。



図 9 「紙とアルミホイルで電池を作る科学実験」のプレゼンテーションを撮影した結果

参考のために、3章で題材とした「炭とアルミホイルで電池を作る科学実験」を撮影した結果を図9に示す。これはシステムに慣れた筆者の一人が自由にプレゼンテーションしたものであるが、注目すべき箇所が強調されるように編集されていることがわかる。

## 5. ま と め

視聴者の注目を喚起するために話者が行う典型的な行動について述べ、テレビ番組の編集と比較することで注目喚起行動に基づいた映像編集の有効性を調べた。また注目喚起行動を動作と発話のイベントの組み合わせで検出し、その結果を用いてショットを選択する自動編集ルールを提案した。さらに主観評価実験と撮影実験を通して以下のことを明らかにした。

- 注目喚起行動をすべて検出することができれば、調査対象としたテレビ番組映像におけるショット切り替えの約50%を扱うことができる。
- 注目喚起行動とショット切り替えを同期させることにより、視聴者の評価の高い映像を得ることができる。
- 単純なイベントを組み合わせで注目喚起行動を検出することにより、初めて撮影システムを使用する人でも意図した通りの編集映像を簡単に得ることができる。

今後の課題として、さらに多くのテレビ番組と比較し、注目喚起行動に基づく編集の利点と欠点を明らかにする必要がある。また3.1節で述べた注目喚起行動以外の要因をトリガとした編集について取り組むことにより、視聴者にとってより面白味のある映像を自動的に取得できると考えられる。面白味のある映像という点に関しては、映像の内容にしばらく変化がない場合に視点を変えたショットや人物の表情を映したショットを挿入するというパターンがテレビ映像で多くみられた。今後はこのような要素も取り入れた研究を行っていく予定である。

謝辞 本稿では、評価用映像メディアデータベース検討部会 (VDBWG) により作成された「評価用映像メディアDB」の一部を使用している [11]。

## 文 献

- [1] L. He et al. Auto-summarization of audio-video presentations. Proc.ACM Multimedia, pages 489–498, 1999.
- [2] S. Mukhopadhyay and B. Smith. Passive capture and structuring of lectures. Proc.ACM Multimedia, pages 477–487, 1999.
- [3] Y. Kameda, K. Ishizuka, and M. Mihoh. A live video imaging method for capturing presentation information in distance learning. Proc.International Conference on Multimedia and Expo, pages 1237–1240, 2000.
- [4] M. Gleicher and J. Masanz. Towards virtual videography. Proc. ACM Multimedia, pages 375–378, 2000.
- [5] A. Bobick and C. Pinhanez. Controlling view-based algorithms using approximate world models and action information. Proc. CVPR, pages 955–961, 1997.
- [6] M. Ozeki, Y. Nakamura, and Y. Ohta. Camerawork for intelligent video production – capturing desktop manipulations. Proc. ICME, pages 41–44, aug 2001.
- [7] M. Ozeki, Y. Nakamura, and Y. Ohta. Human behavior recognition for an intelligent video production system. Proc. PCM, pages 1153–1160, dec 2002.
- [8] M. Ozeki, M. Itoh, Y. Nakamura, and Y. Ohta. Tracking hands and objects for an intelligent video production system. Proc. ICPR, pages 1011–1014, aug 2002.
- [9] 宮崎 英明, 亀田 能成, and 美濃 導彦. 複数のカメラを用いた複数ユーザに対する講義の実時間映像化法. 電子情報通信学会論文誌, J82-D-II(10):1598–1605, 1999.
- [10] 先山 卓郎, 大野 直樹, and 椋木 雅之 and 池田 克夫. 遠隔講義における講義状況に応じた送信映像選択. 電子情報通信学会論文誌, J84-D-II(2):248–257, 2001.
- [11] 馬場口登 et al. 映像処理評価用映像データベースについて. 信学技報, PRMU2002-30, 2002.