プレゼンテーションの知的撮影システム
— 動作認識による映像のタグ付け —

An Intelligent System for Recording Presentations
– Tagging by Recognizing Speaker's Behaviors

尾関 基行 (筑波大学)　中村 裕一 (筑波大学/さきがけ 21)　大田 友一 (筑波大学)
Motoyuki Ozeki†　Yuichi Nakamura†‡　Yuichi Ohta†
† Institute of Engineering Mechanics and Systems,University of Tsukuba
‡ PRESTO, Japan Science and Technology Corporation

概要: 本稿では，映像撮影およびマルチメディアコンテンツの作成を補助する知的撮影・自動タグ付けシステムについて述べる．本研究では対象を手元作業とし，適切な映像を取得してそれを構造化するためのカメラ制御アルゴリズム，話し手の発話情報と動き情報を用いたタグ付け方法，受け手の興味に応じた提示方法等について検討し，実装を行ってきた．本稿では，その中でも話し手の作業動作や指示・提示動作の認識とそれによるタグ付けに重点をおいて説明する．本研究では，いくつかの実験によりこれらの手法の有効性を確かめた．

Abstract: In this paper, we introduce an intelligent system for recording presentations on desktop manipulations. We investigated and implemented a fundamental method for appropriately controlling cameras, a method for video tagging by human behavior understanding, and a method for presenting the recorded data. Among them, we will mainly present the method for recognizing deictic or illustrating movements and for tagging the video. We applied our method to typical presentations, and obtained good results.

1　はじめに

従来の映画や従来のテレビ放送といった映像メディアでは，映像は制作側の意図により撮影・編集され，一方的に視聴者に提供されるものであった．しかし，近年，放送用映像や資料映像などを蓄積し，視聴者の見たい形で再利用する試みが行われるようになってきた．また，映像は非常に有用なコミュニケーション手段でもあるため，遠隔会議や講義，CSCW 等での利用も盛んに行われ始めている．

このような背景から，映像製作会社だけでなく，一般企業や教育機関，さらには個人でも，手軽に映像を撮影し，コミュニケーション手段とすることに関心が高まっている．しかし，映像の取得には多くの技術と労力が必要となるため，非専門家や個人が質の良い映像を撮影することは難しい．また，カメラマンやディレクターを雇うことには多大なコストがかかり，多くの場合には現実的とはいえない．そのため，撮影や編集作業を補助するような自動システムが必要とされている．

この問題に対し本研究では，映像の撮影，映像への情報付加 (タグ付け) を自動化し，それを用いて適切な映像提示を行う知的システムの構築を行っている．対象は手元作業を中心としたプレゼンテーションとし，複数カメラによる自動撮影，話し手の動作認識を基にしたタグ付けなどを行う．また，これらを利用して，視聴者の興味を反映した映像提示の実現を目指す．我々は，このような考え方に基づいて実際にシステムを構築し，簡単なプレゼンテーションに対して実時間で動作させる実験を行った．以下本稿では，本研究の基本的な考え方，システム構成，実験例について述べる．

2　映像取得システムの概要

2.1　映像取得の枠組み

教材映像や説明映像等では，情報を分かりやすく正確に伝達することが最も重要である．まず，映像取得において特に重要な部分，またはその候補となる部分

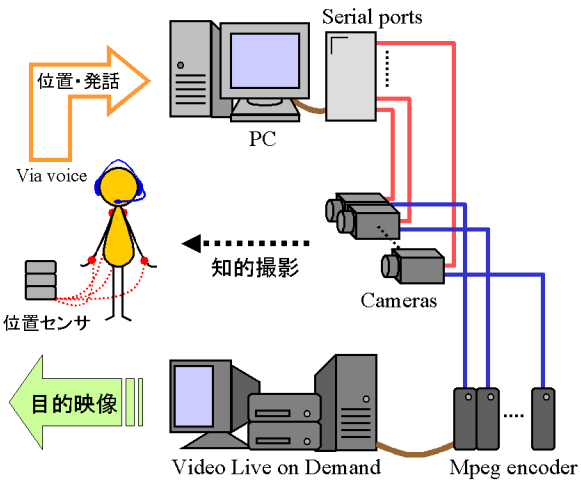

図 1　システムの概要図

を，適切な解像度・視点・カメラ制御で撮影するといったカメラワークが必要となる．さらに話し手および視聴者にとって重要な部分を選択し，目的に応じて加工・提示することが必要となる．これらの，映像メディア取得のポイントを整理すると，以下のようになる．

カメラワーク: 注目対象，視点，解像度，カメラ制御方法

タグ付け: 人物行動（動き・発話），意図，環境の状態の記述と映像への付加

提示方法: 提示部分の選択（時間的選択，視点選択，カメラ選択），要約，メディア変換等

本研究では手元作業を撮影対象とし，それぞれの要素について自動化へのアプローチの検討を行った．

我々の構築しているシステムの概要を図 1 に示す．本システムでは，位置センサにより話し手や特定物などの位置を取得し，首振りカメラを制御することで自動撮影を行う．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る．

ここで，目的に応じた映像提示のためには，複数のカメラからの映像を記録して，各時刻のショットがどのような情報を持っているかを逐次記述することが必要となる．本研究では，人物の発話と動きデータを利用し，注目すべき部分を検出して映像にタグを付与する．さらに得られたタグを基にして，視聴者が見たいと思う部分を効果的に提示することを目指す．

関連する研究としては，講義映像のアーカイブ化 [1][3]，遠隔講義の伝送等 [2][5] の研究がある．遠隔講義の伝送に関する研究では，教師，黒板，教室の様子を教師と生徒の動きや位置，音声などを利用して認識し，伝送映像を自動的に切り替えるシステムが提案されている．これに対し，本研究で対象とする手元作業の撮影では，以下のような点について検討すべき問題が異なる．

- 手作業などの動作をあらかじめシナリオで細かく記述することは難しく，予測できない動作が現れることが多い．しかも，手や物体の動きが視野に対してかなり速く動くため，カメラ制御が難しい．
- 注意を向けるべき対象として，話し手の顔，身振り，手先，物体その他，様々なものが考えられる．そのため，各映像 (ショット) の持つ情報を詳細に記述し，それを基に映像提示を行う必要がある．

このように，カメラワーク (予測しにくい対象の追跡・撮影) と撮影シーンの認識 (注目動作の認識) に関して，多くの新しい問題を含んでいる．これらの内，本稿ではタグ付けの部分，特に話者行動理解によるタグ付け手法について詳しく述べる．

2.2　手元作業映像取得のためのカメラワーク

手元作業映像取得におけるカメラワークの基本は，注目すべき対象を適切な大きさ・位置で画面に捉えることである．しかし，人や物体をただ単純に追跡すれば良いというわけではない．この問題に対し本研究では，注目対象について “**何**” の “**どういう状態**” を捉えるか，という観点から分類を行った．このように，注目対象を “対象物” と “対象とする状態” に分けて考えることにより，注目対象とカメラ設定の関係を簡潔に説明することができる．

現在の段階では，“対象物” として，**話し手・作業空間・注目物体・注目場所**を用意し，‘対象とする状態” については，**＜状況＞・＜操作＞・＜物体＞**を用意している．このような注目対象に対応したカメラワークを行うために，**カルマンフィルタによる平滑化，枠制御アルゴリズム**を提案した．撮影する際には，対象に応じて各処理で用いるパラメータを調節し，目的に応じたカメラワークを設定する．カメラワークの詳細については既に報告しているので，文献 [6] を参照されたい．

2.3　映像のタグ付けと提示

前述したように，映像提示を効果的に行うためには，映像の各部分について，適切な記述を付加データとして与えることが必要となる．このような付加データには，各映像が含む情報，その重要性，また，ある情報を提示するための適切さ等に関するものがあるが，本稿ではこれらをまとめてタグと呼ぶことにする．手元作業映像に対するタグとして，具体的に以下のようなものが挙げられる．

- 各カメラの設定 (位置，注目対象，解像度，制御方法等)
- 各時点での各カメラの状態 (制御値，制御状態等)
- 各時点での話し手の言動，意図
- 各時点で映っている対象とその状態

カメラの設定については，各カメラが狙う対象の種類とその撮影方法に関する記述を付加する．カメラの状態については，各カメラの各時刻の制御値（パン・チルト角）などに加え，追跡状況や枠制御アルゴリズムの状態を付加情報として与える．話し手の言動・意図については，話し手が行っている動作の種類，発話内容，意図，さらに話し手が見せようと思っている対象等が挙げられる．映像に映っている対象とその状態としては，対象の画面上での位置，シーン内での 3 次元位置，状態 (向き)，その他多くの項目が考えられる*．

本稿以下では，この中でも特に重要な項目である，話し手の言動・意図の認識とそれを利用した映像提示について詳しく述べる．

3　話し手の動作検出

説明映像の提示やコミュニケーション支援を考えた場合，話し手が注意を促す部分，また普通の人なら見るであろう部分が重要となる．これに最も深く関わっているのは，話し手の動作と発話である．これらを認識することができれば，映像に含まれる情報やその重要性を判断することができる．そのため，本研究では話し手の動きと発話を利用し，話し手の動作とその意図を推定して映像のタグ付けを行う．

この問題に対して，従来から我々のグループでは，MMID[4] を用いて指示・例示動作抽出の研究を行ってきており，発話の処理と体の動きの処理を併用することにより，かなり良い精度で指示動作を抽出できることがわかっている．

そこで本研究では，この研究を基にして指示動作および例示動作を抽出し，映像のタグ付けを行うことを考える．ただし，本研究では実時間での認識を実現するために，上記の研究とは少し異なる方法をとる．

3.1　発話の処理

指示・例示動作を行う際に最もよく現れるのが指示詞の発話である．そのため，本研究では指示詞とそこから得られる情報を簡単に分類し，動作認識に利用する．指示詞対象の分類，指示詞の形態，視聴者が注目すべき情報についてまとめたものを表 3 に示す．

*ただし，これについてはまだ実装しておらず，今後の課題となっている．

表 3　指示内容とそれに含まれる情報

指示対象の分類	指示詞の形態	視聴者が注目すべき情報
操作・状態	このように， こうして, 等	動き・状態， 操作方法
注目・名付	これ系， この＋名詞, 等	注目物体・部分
場所・部分	ここ系， この＋場所	注目場所や位置関係
時点・区切	ここで, これで	全体の流れの区切り

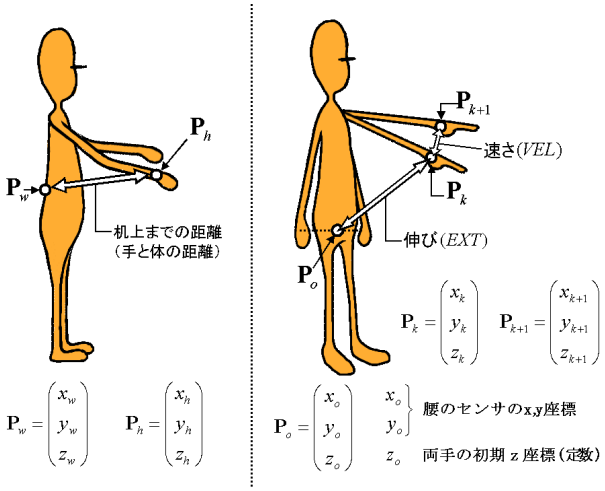

図 2　机上操作と指示・提示を行う動き

ここで “名付” とは，ある物体の名前を紹介するような場合である†．これらの指示詞は，動作の起こりを示すトリガとして利用する．

3.2　話し手の動きの処理

話し手の動きを用いることで，指示詞だけでは得られない “動作の行われた場所” を知ることができる．また，指示詞は文脈指示としても用いられるため，必ずしも指示対象が物理的に存在するとは限らない‡．このような場合に動き情報を併用して認識を行うことにより，不要な動作の検出を防ぐこともできる．

現時点で実装しているのは，以下の 2 つの動きと状態の検出である．

- 机上での操作 (例示動作)
- 指示・提示を行う動き

これらの動きは，手と腰の位置から検出する（図 2 参照）．まず，以下の式が成り立つ間，机上での操作の

†例えば,「これは○○です.」等.

‡例えば,「この」は非常に頻繁に登場し，その多くの場合が注目の要求とは関係がない.

表 1　撮影例で使用したカメラ設定

動作ラベル	指示詞ラベル	話し手の動き
例示動作	操作・状態	手が体から離れている
物体指示動作	注目・名付	提示（指示）を行う動き
場所指示動作	場所・部分	指差しを行う動き
調整子	時点・区切	手が下に降りている

可能性があるとする．

$$|\mathbf{P}_w - \mathbf{P}_h| > Th_{wh}$$

Th_{wh} の値は，実際に机上操作を行ったデータより決定した．

指示・提示を行う動きは，手が大きく伸びて，急停止する動きであると定義して，次の 2 つの値を計測することで検出を行う．

- 手の伸び変化の極大値
- 手の速さ変化の極小値

ここで，伸び変化 $EXTC$ と速さ変化 $VELC$ は，図 2 の記号を用いて以下のように表される．

$$EXT_k = |\mathbf{P}_k - \mathbf{P}_o| \quad (1)$$

$$EXTC_k = EXT_k - EXT_{k-1} \quad (2)$$

$$VEL_{k+1} = |\mathbf{P}_{k+1} - \mathbf{P}_k| \quad (3)$$

$$VELC_k = VEL_k - VEL_{k-1} \quad (4)$$

ここで P_o は，z（高さ）を両手の初期位置，x, y（水平位置）を現在の腰の位置とした基準点である．ただし，細かな動きやノイズの影響を除去するために，計測値にガウシアンフィルタをかけて平滑化している．

伸び変化の極大値が検出されると，その時刻の伸び変化の値を閾値と比較し，閾値より大きければ “伸びの検出” とする．伸びが検出されると，同様にして速さ変化を計算し，その極小値を求める．極小値が検出されると，その時刻の速さ変化の値を閾値と比較し，閾値より小さければ “指示・提示の動きの検出” とする．これらの閾値は，実際に指示・提示の動きを行ったデータより経験的に決定している．

我々のグループがこれまで蓄積してきた MMID[§] を対象に，プレゼンテーションを行った場合の手の伸び変化と速さ変化，および指示・提示の動き検出を行った様子を図 3 に示す．図中の $< a > \sim < c >$ は，実際に指示動作が行われた部分である．手の伸び変化の極大値，速さ変化の極小値が連続して起こる部分と，ほぼ重なっていることが読み取れる．経験的に決定した

[§] プレゼンテーションの映像データと共に，話し手の位置データ，発話内容，指示・例示動作の時刻などが含まれる．

表 2　撮影例で使用したカメラ設定

カメラ	対象ラベル	制御方法	追跡基準点
カメラ 1	話し手・中	＜状況＞	腰の一点
カメラ 2	注目物体・小	＜物体＞	右手の先端
カメラ 3	作業空間・中	＜操作＞	両手の中点

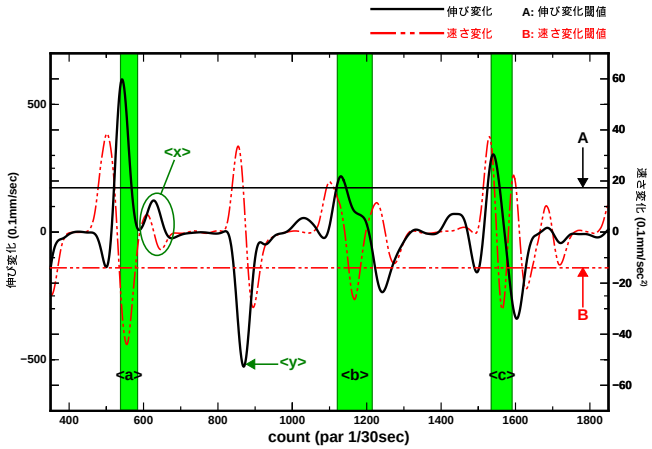

図 3　手の伸び変化と速さ変化

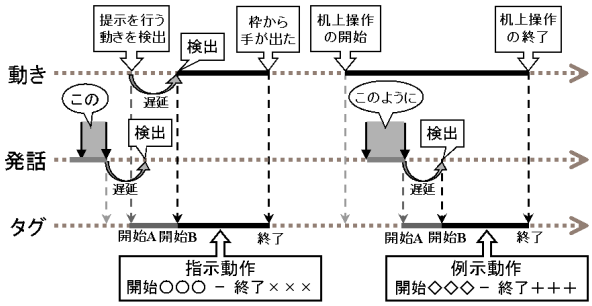

図 4　指示詞と動きによるタグ付けの原理

閾値により，$< x >$ のような小さな動きは検出しない．また，体から手が離れる方向に伸びを定義しているので，$< y >$ のように勢い良く手を降ろす動きなども検出していないことがわかる．

3.3　指示詞と動きによる動作検出

指示詞と話し手の動きの組み合わせを利用することで検出できる動作をまとめる．表 1 に，指示詞と動き，対応する動作ラベルを示す．

図 4 に指示詞と話し手の動きによるタグ付けの原理を示す．指示詞と動きの検出がほぼ同時[¶] に起こった時点で，動作の検出とする．最下段のタグ付けラインでの “開始 A” はオフラインで処理した場合，“開始 B” はオンラインでのタグ付け処理の場合である．現在のシステム構成では，指示・提示の動き検出に約 1.3 秒，指示詞検出に 2〜4 秒の遅れが生じる．

[¶] 3 秒以内に両方が起こる等．

Count	動作[指示詞内容]	注目対象	フレーム:開始 − 終了
0th	操作・状態[このように]	作業空間	225 − 345
1th	注目・名付[この]	注目物体	497 − 706
2th	操作・状態[このよう]	作業空間	858 − 1037
3th	注目・名付[これ]	注目物体	1217 − 1308
4th	操作・状態[こうやって]	作業空間	1488 − 1668
⋮	⋮	⋮	⋮

図 5　タグ付けの例

動作の終了は手が体に近づいたこと，または，画面内に仮想的に考えた枠から出たことにより検出する‖．この枠の大きさは，撮影する対象とその状態に応じて決める [6]．これらの情報をタグとして映像に付与し，それを映像提示に利用する．

4　システム構成

指示・例示動作検出のためのシステム構成は以下のようになっている．

音声認識は IBM 社の ViaVoice を使用し，音声認識用 PC 上で認識された単語から指示詞のみが抽出され，ホスト PC に送られる．話し手の動きは，話し手に装着された 6 自由度磁気センサ Flock of Birds(Ascension Technology Corporation) で計測され，ホスト PC に送られる．

ホスト PC に送られたデータは，並行して動作するカメラ制御プロセスと動作認識処理プロセスの入力となる．カメラ制御プロセスの処理結果はパン・チルト制御値として各カメラに送られる．それと同時に，動作認識処理プロセスで上記の動作認識が行われ，映像にタグが付与される．また，実時間でのカメラ切り替えを行う場合には，動作が検出されると信号が映像スイッチャーに送られ，最も重要度の高いカメラからの映像に切り替わる．この仕組みにより，遠隔会議や協調作業空間でのコミュニケーション支援としても有効なシステムとなっている．

5　実験例

これまで提案してきた手法を実際のデータに適用した実際結果を示す．

動作認識精度の確認

まず，MMID に収録されているデータを用いて，動作認識精度の確認を行った．データは 1〜3 分程度の長さの比較的簡単なプレゼンテーション 7 本であり，そ

‖ただし，時点・区切は時間的幅を持たないため，終了という概念は用いない．

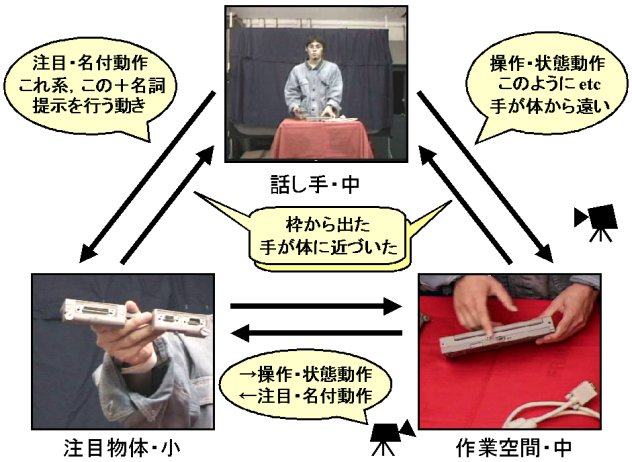

図 6　切り替え条件

の中には，物体の提示，場所の提示，操作の実演などが収録されている．ただし，音声認識については，収録音声の質が悪いため，トランスクリプトを用いた**．

このデータに対し，指示・提示動作を検出した結果，検出率 (再現率) は約 67%(33 例中 22) であった．検出されなかった動作の多くは，動きが小さく，動作検出の閾値の設定以下だったものである．これについては，上記のアルゴリズムで検出することは難しく，発話，物体の位置，指形状等の情報を使っていくことが必要である．ここで，動きの処理だけを行った場合，離れた物体を取る場合などの動作が誤検出されることがある (誤検出率は約 32%となる) が，発話を併用することにより，テストデータに対しては誤検出はなかった．

また，例示動作の検出率は 100%(9 例中 9) であったが，これは例示動作が必ず指示詞を伴っていたからである．実際のプレゼンテーションでは，必ずしもこのような条件が成立するとは限らないため，今後も調査を続ける予定である．

なお，これらは比較的簡単なプレゼンテーションに対する実験例であり，料理等のより複雑なプレゼンテーションに対して適用すると，かなり精度が悪くなる．これは，上記で想定しているよりも多様な動作が行われていたり，体の動きが小さかったり，発話が複雑であったりすることによるものである．現段階ではまだ，このようなプレゼンテーションに対するシステマティックな実験を行っていないが，順次これらの対象も扱っていく予定である．

プレゼンテーションの撮影

4 章で説明したシステム構成により，実際の手元作業プレゼンテーションの撮影を行った．使用したカメ

**音声認識の精度が 100%であると仮定していることになる．

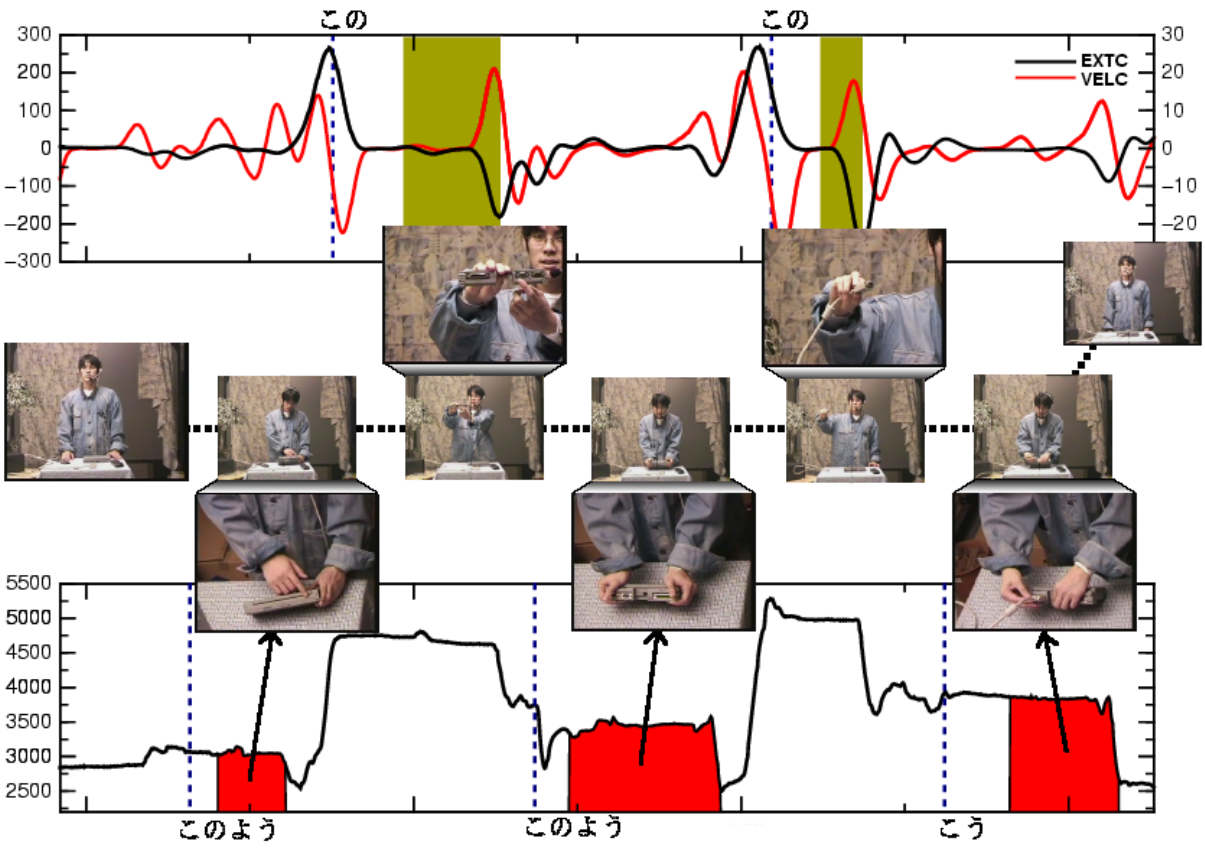

図 7　切り替えの効果

ラは図 6 の 3 台であり，それぞれのカメラ設定は表 2 の通りである．プレゼンテーションはノートパソコンにディスプレイケーブルを取りつける説明とした．

この実験例では，話し手の動作認識により図 5 に示すようなタグを映像に順次付加している．これを利用することで,「注目動作による切り替え」「カメラ制御状態による切り替え」「特定動作部分の切り出し」といったような提示を行うことができる．一例として図 6 の方法で，注目動作によって映像を切り替えた例を図 7 に示す．指示詞と話し手の動きの関係により，適切なタイミングでカメラが切り替わっていることがわかる．また，この切り替え処理はスイッチャーを使用して自動で行っており，遠隔コミュニケーションシステムへの応用も可能である．

なお，ここで挙げた映像切り替え以外にも種々の提示方法が考えられ，応用目的に合わせた効果的な映像提示方法について探っていく予定である．

6　まとめ

マルチメディアコンテンツ取得，コミュニケーション補助の為の知的撮影システムの枠組みを提案した．そのために，複数のカメラによって取得された映像に情報を付加する方法を検討した．手元作業映像で重要な動作について，話し手の発話情報（指示詞）と動き情報を利用した検出手法を提案し，簡単な実験によりその有効性を示した

今後の課題として，まず未実装の部分を完成させることがあげられる．さらに，特定物の位置やシナリオなど，様々なデータを使用した映像のタグ付けとそれを有効に活用した提示について研究を進めていく予定である．

参考文献

[1] L. He, et al. Auto-summarization of audio-video presentations. *Proc.ACM Multimedia*, 1999.

[2] 宮崎英明, 亀田能成. 複数カメラを用いた講義映像の実時間作成法. *MIRU'98*, pp. 123–128, 1998.

[3] S. Mukhopadhyay and B. Smith". Passive capture and structuring of lectures. *Proc.ACM Multimedia*, 1999.

[4] Y. Nakamura, et al. MMID: Multimodal multi-view integrated database for human behavior understanding. *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, 1998.

[5] 大野直樹, 池田克夫ほか. 遠隔講義における講義状況に応じた送信映像選択. 第 5 回知能情報メディアシンポジウム, 1999.

[6] 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—手元作業を対象とした適応的カメラワーク—. *PRMU2000*, 2000.