机上作業シーン映像の
自動撮影・編集手法に関する研究

システム情報工学研究科
筑 波 大 学

2005 年 3 月

尾 関　基 行

論文要旨

学習資源としての映像コンテンツの充実は高度情報化社会を形成し発展させるための重要な課題であり，特に，各地域の教育機関や小さなコミュニティの知識がコンテンツとしてやりとりされることは，社会全体の情報交流の活性化に繋がる大きなメリットがある．これに対して本研究では，料理や科学実験などの机の上で行う作業や説明（机上作業シーン）を題材として，誰でも手軽に映像コンテンツを制作するための自動撮影・編集手法について研究を行った．本論文では，パン／チルト制御による追跡制御とカメラ選択によるショット切替え編集に焦点を絞り，見やすく分かりやすい映像を取得するための自動撮影・編集手法を提案し，試作システムでの評価実験を通してそれら有効性を明らかにする．

まず，自動撮影手法について，机上作業シーン映像では撮影する対象が何であるかによってではなく対象のどのような状態に注目するかによってカメラワークが決まるという考え方を提案し，その注目すべき状態の典型として＜外観＞・＜動き＞・＜周辺関係＞の三つを定義する．そして，これらのカメラワークを自動制御で実現するために，対象の動き（反復・停留）に応じて追跡モードと固定モードを切り替える「可変枠制御」を提案する．

次に，自動編集について，机上作業シーン映像の 90%以上が話者と作業領域の両方を含んだショットと作業領域のみのクロースアップの繰り返しで構成されていることをテレビ番組の分析より示し，その切替えのトリガとして話者がある箇所に注目を集めようとする行動（注目喚起行動）に着目する．そして，注目喚起行動に基づいた編集モデルを考案し，話者の発話と手の位置関係から注目喚起行動を自動検出する手法を提案する．

更に，机上作業シーン映像の自動撮影・編集システムと物体追跡システムを組み合わせることにより，物体に関する情報（位置・テクスチャ・把持状態・作業内容）を映像と同期して記録する映像インデキシングを実現する．可視光カメラ・赤外線カメラ・ステレオカメラから肌色領域・動領域・肌温領域・特定距離領域を抽出して論理積をとることで，物体の大きさ・色・形状などの予備知識がなく背景が常に変化するという条件の下でも，精度良く物体を検出できる手法を提案する．また，このようにして得られたインデックス付き映像コンテンツの利用例として，物体のアイコンをキーとした映像検索，対話型映像メディアのコンテンツとしての利用，複合現実空間のコンテンツとしての利用に取り組んだ結果をそれぞれ紹介する．

目 次

図 目 次

第1章　序論

本論文では，料理映像や科学実験映像などの机上作業シーン映像を自動取得するための撮影・編集手法を提案し，試作システムを用いて行った評価実験によってそれらの有効性を示す．

本研究の背景として次の二つがある．

1. 高度情報通信ネットワーク社会の形成のためには，インフラやソフトウェアだけでなく，そのコンテンツ（情報の内容）の充実が不可欠である．これに対して，各種学習資源のデジタルアーカイブ化による教育用コンテンツの充実，及びコンテンツ制作支援技術の確立は国家規模で重点計画として掲げられている課題の一つであり [1]，盛んに研究・開発が行われている．
2. 各種情報端末の高速化・大容量化に伴い，やりとりされるコンテンツの形態も文字から画像，映像へと，より情報量の多いメディアへと変化してきた．また，MPEG7 や TVML など，映像を効率的に利用するためのメタデータ（付加情報）の国際標準化も進んでいる．このような映像とメタデータを併せ持った高度な映像コンテンツ（インデックス付き映像）は高い表現力と扱いやすさを兼ね備えており，これからのコンテンツ形態として中心的な役割を果たすと考えられる．

このように，インデックス付き映像を用いた学習資源の充実は，来る高度情報化社会を形成し発展させるための重要な課題である．近い将来には，映像制作会社だけでなく一部の高等教育機関や公共施設，企業などでも，ビデオ教材や映像マニュアルなどが制作され，皆にとって有用な情報（= 知識）としてインターネット等を通じて共有されるだろう [2]．これは映像メディアを用いた巨大な「知識のアーカイブ」であり，普及しつつある e-Learning や，テキストベースのものでは Wikipedia[3] など，既に実用化されているもの存在する．

このような知識のアーカイブにおいて，一部の大規模な機関からだけでなく，各地域の教育機関や小さなコミュニティの知識がコンテンツとしてやりとりされることは，社会全体の情報交流の活性化に繋がる大きなメリットがある．しかし映像コンテンツの制作には，撮影・編集・インデキシング[1]など，多大な手間とコストがかかるため，個人・地域レベルで独自に制作することは難しい．上述したような社会を実現するには，例えば，簡単に使えるコンテンツ制作スタジオが各地の初等中等教育機関やコミュニティ施設に設置され，講義やプレゼンテーションを行えば，すぐに映像コンテンツができあがるような仕組みが望まれる．そのためには，誰でも手軽に自分の持っている知識や技術を映像コンテンツとして取得するための技術的・学術的な土台作りが不可欠である．

[1]インデックス（索引）として，メタデータを映像に付加すること．一般に，MPEG7 など XML ベースの記述方式が用いられる．

これに対して筆者は，料理や科学実験などの机の上で行う作業や説明（**机上作業シーン**）を題材として，誰でも手軽に映像コンテンツを制作するための自動撮影・編集手法について研究を行ってきた．机上作業シーンを撮影した映像（**机上作業シーン映像**）は，個人や地域から発信される代表的な学習資源の一つであり，また，文字や画像だけでは伝えづらい知識である「作業」は映像メディアの特性を活かせるコンテンツである．机上作業に関する知識を手軽にコンテンツ化できれば，学校独自の科学実験や工作，地域特有の料理や伝統技術など，局地に滞っている知識を広く共有することができ，社会的な意義も大きい．

映像コンテンツ制作を補助する研究は 1990 年代から盛んに行われるようになり，2005 年現在では実用化に至っているものもいくつかある．それらは，首振りカメラや映像切替器などの機器を自動制御するシステムと，既に撮影された映像素材から編集・管理・インデキシングなどを行うシステムに分けられるが，本論文では機器を自動制御するシステムについて述べる．また，映像制作の専門家に向けたものと専門家以外に向けたものに分けられるが，既に述べたように，ここでは専門家以外の人々が使用することを想定する．

従来研究として，教室での講義シーンを対象とした研究がこれまで数多く行われてきた．これらの研究で得られた知見は，その他の教示シーン（何かについての説明を行っているシーン）に対しても重要な示唆を含んでいる．このように，個々のシステムはあるシーンに特化して設計されたものであっても，それぞれで提案されている技術は教示シーン全般に共通して利用できる汎用的な技術であるべきである．これに対して本研究の取り組みは，これまでになかった机上作業シーン専用の自動撮影・編集システムを実現するというだけでなく，講義シーンを対象とする研究とは違った観点から一般的な教示シーンの撮影・編集に利用できる新たな知見を提供できると考える．

本研究の目的は以下の二つである．

目的 1: 誰でも手軽に自分の持っている知識・技術を映像コンテンツとして記録するために，カメラマンとディレクタの機能を代替する自動化システムを実現する．

目的 2: 知識・技術の映像化という観点において，その教示シーンから何をどのように切り取り（撮影）それらをどう再構成するか（編集）について，その基本となる技法を明らかにする．

これらについて，本研究を以下の三つの課題の集まりとして考え，それぞれについて本研究独自のアプローチで取り組んできた．

カメラマンの機能の実現： コンテンツとして記録すべき箇所はシーン中の各所に遍在するが，本研究では，自動制御された首振りカメラを複数台用いて重要箇所となり得る対象を常に撮影しておく．この際，人間のカメラマンに熟練が必要であるように，コンピュータによる自動カメラ制御にも種々のテクニックが必要である．本研究では，テレビ番組の観察・分析に基づいてショットとカメラワークの関係を定義し，それらのカメラワークを自動化システムで実現するためのカメラ制御手法を提案する．

ディレクタの機能の実現： 撮影された複数の映像（ショット）から，各時点で最も重要である箇所を映したショットに切り替えることにより視聴者の注意を重要箇所に誘導し，説明内容のポイントを分かりやすく伝える．どこが重要であるかは見る側の考え方によって変わるが，本研究では，説明を行っている人物（話者）が視聴者に注目して欲しいと考えている箇所を重要箇所であるとする．指示動作や提示動作など，話者が重要箇所を明示的に強調する行動に焦点を当て，話者の手の位置関係と発話を利用して行動を検出する手法を提案する．

映像インデキシングとその利用： 机上作業シーン映像の自動取得と同時に，シーンに関するメタデータを自動取得し，インデックス（索引）として映像と共に記録する（映像インデキシング）．本研究では，インデックスとして机上作業で重要な意味を持つ「物体」に着目し，机上作業シーンにおける物体の自動検出・追跡手法を提案する．このようにして得られた高度な映像コンテンツの利用例として，物体のアイコンをキーとした映像検索，対話型映像メディアのコンテンツ，複合現実空間におけるコンテンツを紹介する．なお，この部分は伊藤，伊津野，里との共同研究であり，筆者は映像コンテンツの取得，システムの統合，共同研究の取りまとめを担当した．

本論文では，まず 2 章で，教示映像の定義と目的，及び本研究の狙いについて述べる．それらに，関連研究の紹介とシステム構成の解説を加えて，「机上作業シーンの撮影・編集」としてまとめる．その後，上述した三つの課題について，それぞれ 3 章〜5 章として解説する．3 章ではカメラマンの機能の実現について，4 章ではディレクタの機能の実現について説明し，それぞれについて評価実験を行った結果を述べる．5 章では，物体に関するメタデータの自動取得手法の提案とその利用を共同研究として行った内容について述べる．

第2章　机上作業シーン映像の撮影と編集

2.1　はじめに

人が人へ何かを伝えるための媒体（メディア）という観点から映像を定義すると,「空間をいくつかの視点から動画として切り取り（撮影），それを一本の時間軸上に並べ直すことで（編集），元の空間の情報を人へ伝える」ためのメディアであるといえる．そして，言葉や絵画に文法や技法があるように，送り手と受け手がメディアに乗せられた情報に共通の概念を持つための，また受け手に上手く情報を伝えるための，映像化の技法（映像の文法）が存在する．これら映像の文法は，映画制作の技術書などに体系づけてまとめられている [4]–[5]．しかし実際には，映像化する空間（シーン）やその映像の目的・用途によって，必要とされる技法は違ってくる．撮影・編集の自動化に関する研究において，この点が明確にされないまま，映画制作／テレビ制作で用いられる技法のみを正解として議論されることは問題である．

本章では，まず，教示シーン映像及び机上作業シーン映像の定義と目的についてまとめる（2.2 節）．次に，従来研究では見過ごされてきた机上作業シーン映像で顕在化する問題を挙げ，それに対する本研究の狙いを示す（2.3 節）．また同時に，映像化の観点からみた本研究の位置づけについても議論する（2.4 節）．最後に，本研究で提案した手法を実装して評価するためのベースシステムについて，その設計概念・構成・キャリブレーションについてまとめる（2.5 節）．

2.2　教示シーン映像

まず，机上作業シーン映像を含む教示シーン映像全般について，その定義と目的をまとめる．

2.2.1　教示シーン映像の定義

人物が何かについての説明や解説を行なっている場面をここでは**教示シーン**と呼ぶ．教示シーンには，例えば，講義や料理，科学実験，工作，手芸，園芸，コンピュータ，各種ゲーム，スポーツなどがある．この教示シーンを撮影した映像として，教育映像や教養映像，ハウツービデオ，映像マニュアルなどと呼ばれる映像がある．テレビ放送や市販ビデオにみられるこれらの映像は専門家が制作したものであり，純粋な教示シーンのカット（映像の断片）に加え，それを補強する様々なカットを織り交ぜて構成されている．このような映像を作るには，撮影・編集共に高度な知識と設備を要する．

表 2.1: Arijon による映画の分類（例は筆者の考えに基づく）

分類	説明	例
ニュース映画	繰り返しのきかない行為やできごとを捉えようとした映像. 映像制作者は自分が記録しようとする出来事に対して最小限度の支配力しか持たず，最も完全な記録は出来事の一部始終を数台のカメラで撮影しておくことで得られる．ある状況を完全に，且つ偏りなく捉える理想的なカメラ・ポジションのようなものはなく，カメラマンはカメラ位置・視点・解像度などで妥協を余儀なくされる.	ニュース映像 スポーツ中継 料理映像 トーク番組 各種生放送 監視映像
ドキュメンタリー映画	純粋な現実と注意深く再構成された虚構を混ぜ合わせてできた映像．一般に，単一の出来事ではなく共通した誘因によって起こる一連の出来事を扱い，それらの事実を最上の状態で示すために何らかの整理が行われる．この際，出来事はしばしば繰り返し行われて何度も撮影される，つまり演出が入る.	ドキュメンタリー番組 バラエティ番組 音楽番組 教育番組 紀行番組
劇映画	最終的な映画形式である完全なる虚構で構成された映像. 出来事は現実のものであるが，それは必要なだけ何度でも自由に繰り返すことができるので，行動や演技の正確なニュアンスが多くの角度からフィルムに捉えられる．個々の状況は，撮影のために注意深く練り上げられ実演される.	映画 ドラマ

これらの映像と純粋に教示シーンだけを撮影した映像を区別するために，ここでは純粋に教示シーンを撮影した映像を**教示シーン映像**と呼ぶことにする．本研究も従来研究もこの教示シーン映像のみを対象としているが，教示シーン映像を自動取得できるだけでも，日常の教示シーンの記録や教示シーンを遠隔地へ伝送するための撮影など様々な用途に利用できる．むしろ，こういった用途での撮影・編集は，映像制作会社に発注するほどのコストがかけられないため，自動化システムの担うべき領域であるといえる.

Daniel Arijon による映画（映像）の分類 [6][7] を表 2.1 に挙げる．我々が日常テレビで見ている映像は，各番組の主となる映像部分で判断すると，表の右列のように分類できると考えられる．これらは常に明確に区別できるわけではないが，各種映像がこのような傾向にあることを念頭においておくことは重要である．これらの内，教示シーンを含むものは，教育番組や料理番組，教養番組，科学番組，情報番組などであり，これらはニュース映画〜ドキュメンタリー映画に分類される．純粋な教示シーン映像は，主にニュース映画に位置付けられる.

本研究で対象とする**机上作業シーン映像**とは，このような教示シーン映像の中でも，机の上で行う作業を説明する人物を撮影した映像を指す．上に挙げた例の中では，料理，科学実験，工作，手芸などを撮影・編集したものがこの映像にあたる.

2.2.2 教示シーン映像の目的

本章の冒頭で述べたように，必要とされる映像化の技法は，制作しようとする映像の目的・用途によって変化する．例えば，映画やドキュメンタリー映像の目的はストーリーや人物の感情などを臨場的に伝えることであり，映画制作の技術書にあるような，空間的感覚・時間的感覚・人物の感情・シーンの雰囲気などを表現するための多くの技法が必要となる．ニュース映像の目的は話題となっている出来事を正確に伝えることであり，5W1H[1]の要素を漏れなく含めることが必要とされる．また，バラエティ映像の目的はそのシーンで展開されている内容と人々の様子を娯楽的に伝えることであり，体系的にはまとめられていないものの，適度な認知的刺激や負荷を与えることによって視聴者を飽きさせないといった工夫が随所にみられる．

これらに対し，教示シーン映像の用途は「映像を介して知識を伝えること」であり，その第一の目的は「教示内容を見やすく分かりやすく伝えること」である．娯楽性の強い教育番組や教養番組ではバラエティ映像に近いものもあるが，人物の様子よりも教示内容を優先したショット構成と編集パターンに特徴がある．本研究では，教示シーン映像の第一の目的に焦点を絞り，臨場感を与えるための種々の技法や娯楽性を与えるための工夫については考えない．しかし，3 章及び 4 章で述べる実験結果は，この第一の目的を満たすだけでも十分満足できる映像となることを示唆している．

教示シーン映像の目的をこのように定義すると，必要となる技術は,「見やすく映像を撮影すること」及び「分かりやすく映像を編集すること」となる．見やすく撮影するためには，注目している対象が画面内において明確であること，伝えるべき情報が鮮明に映っていること，安定した視野で映像酔いしないことが要される．また，分かりやすく編集するためには，ショット切替え前後の因果関係が明確であること，教示のポイントが強調されていること，伝達する教示内容に不足がないことが要される．

2.3 従来研究と本研究の狙い

2.3.1 従来研究

撮影システムに関する従来研究

料理シーンを対象とした自動撮影システムとしては，Bobick らの Intelligent Studios がある [8]．この研究では，話者の顔・胴・手元などを簡易な人体モデル当てはめを用いた画像処理で検出し,「手元のクロースアップショット」といったディレクタの指示によってカメラを制御する．ただし，彼らは実際に複数のカメラを用いたシステムは構築しておらず，机上作業シーンで問題となる素早く不規則に動く対象を撮影するためのカメラ制御についても取り組んでいない．

講義シーンを対象とした自動撮影システムとして，椋木らは，ある対象を人手で撮影した映像をテンプレートとして用意しておき，自動撮影時に対象の類似した移動パターンが予測された時に記

[1]いつ (When)・どこで (Where)・だれが (Who)・なにを (What)・なぜ (Why)・どのように (How) の英単語の頭文字をとった言葉.

録しておいたカメラワークで撮影する手法を提案している [9]．また Gleicher らは，高解像度のカメラでシーンを撮影し，その映像から一部を切り出すことで擬似的にクロースアップしたショットを得る手法を提案している [10]．撮影後に処理するため，板書に被った人間を半透明にするなどの効果を加えることもできる．

NHK 放送技術研究所では，放送用映像の制作補助を目的とした知的ロボットカメラシステムの開発を行っており，放送品質の映像を取得するために，プロカメラマンによるカメラ制御の特性を自動化システムに組み込む研究を行っている [11][12][13]．この研究が高価な機材と専門的な知識を用いて放送品質の映像を作ろうとしているのに対し，本研究では比較的廉価な機材と分かりやすい設定ＧＵＩを用いた誰でも簡単に使えるシステムを構築することを目的としている．

その他，グループウェアや作業支援ロボットなどの研究でも作業中の人物を撮影するものがあるが，人が映像という形で鑑賞するための結果を得ようとする研究とこれらの研究ではその観点が大きく違う．

撮影・編集システムに関する従来研究

黒板やスライドを用いた教示シーンを対象とした自動撮影・編集システムについては，これまでに数多くの研究が行われてきた [14][15][16][17][18]．その中には，TIDE プロジェクト [19] のように実際の授業で使われているシステムもある．これらは教示シーンを対象とした撮影〜編集の自動化に関する研究であり，最も直接的に本研究と関連している．

また，会議シーンや対話シーンなどを対象とした自動撮影・編集システムの開発も行われている．例えば，井上らは会議シーンを対象として，テレビの討論番組などの編集パターンの統計に基づいて自動編集を行う手法を提案している [20]．また，尾形らは対話シーンを対象として，人物の頭部を適切な構図で捉えるようカメラを制御し，制約充足を用いて様々な目的に応じた編集を実現するシステムを構築している [21][22]．

更に，仮想シーンにおける撮影・編集の研究は CG の分野で多く行われており，有限オートマトンによるカメラワークの適応的決定を行う Virtual Cinematographer[23] やカメラ制御記述言語 DCCL を提案して実装した CamDroid[24] などがある．しかし，仮想シーンではカメラ制御やシーン内の情報取得などに制限が少なく，仮想シーンで有効な手法がそのまま実際の撮影に適用できるとは限らないため，実際の撮影システムでは実現されていない（実現できない）部分が多い．

2.3.2 本研究の狙い

以上の従来研究に対し，机上作業シーンを対象とすることから顕在化する問題に本研究では取り組む．これらの問題を解決するための自動化技術は，一般的な教示シーンの撮影・編集でも重要であるが，従来の研究では保留されてきた部分である．本節では，カメラマンの機能とディレクタの機能における本研究の狙いをそれぞれ述べる．なお，以下の議論はテレビ番組の机上作業シーン映像（料理番組の例：図 2.1）を参考にしているが，その詳細な分析の結果については 3 章と 4 章で再

図 2.1: 机上作業シーン映像（料理番組）の一例

度述べる．

カメラマンの機能

本研究ではカメラマンの機能として，「対象に応じたカメラワークを用いて追跡撮影することによって，ブレや振動のない安定した視野で対象を捉え，見やすい映像を撮影すること」に焦点をおく（図 2.2 上段）．以下にその理由を述べる．

机上作業シーン映像撮影の最大の特徴として，手元や物体などの素早く不規則に動く対象をクロースアップで捉えたショットが多用されていることが挙げられる．このようなショットを単純追跡で撮影すると，映像酔いするような見苦しいショットになってしまう．見やすい映像とするためには，安定した視野で対象を捉えるよう追跡しなければならない．また，映像酔いを避けるだけでなく，対象に応じてその追跡方法を変えることも重要である．これに対して本研究では，パン・チルト制御で固定と追跡を適切に切り替えることにより，映像酔いするような視野のブレや振動を避けると同時に，撮影対象に応じた追跡撮影を実現するためのカメラワークを提案する．

なお本研究では，三脚に固定したカメラのパン・チルト制御による追跡撮影を考え，ドリー（カメラ本体の移動）による撮影は考えない．ドリーは導入にコストがかかるため，非専門家のためのシステムという目的に沿わない．一方，ズーム制御は，対象を適切な大きさで画面内に収めたり注目すべき対象を強調したりするなど，見やすい映像を取得するための重要な技法であるが，本研究では追跡撮影に焦点を絞る．

また構図については，対象を常に画面中央に捉えるように追跡するとし，対象に応じた構図の変化は考えない．机上作業シーン映像で多用される手元や物体のクロースアップでは，対象を画面中央に捉えた構図が一般的であるからである．また，シーン全体を映したショットも，基本的に人物を正面から中央に捉えた構図が採用されている．人物のみを映したショットには「前空き」などの構図が適用されるが，そういったショットの出現率は低い（10 %程度）．

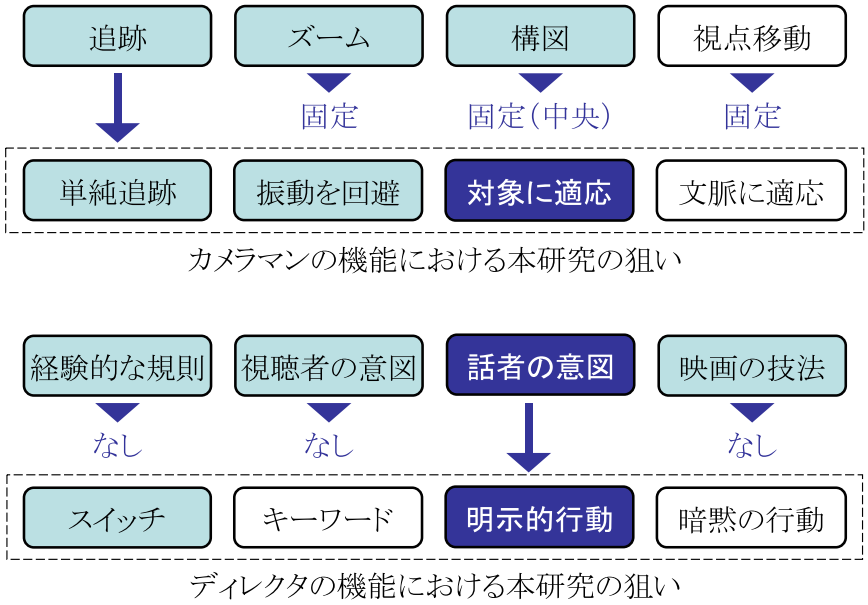


図 2.2: 従来研究に対する本研究の狙い（白抜き文字の項目が本研究の狙い，色のついた項目は従来研究で既に実現されている部分）

ディレクタの機能

本研究ではディレクタの機能として,「話者が明示的に行動で示した注目箇所をクロースアップに切り替えて見せることによって，教示のポイントを上手く強調し，分かりやすく映像を編集すること」に焦点をおく（図 2.2 下段）．以下にその理由を述べる．

シーンを多視点から撮影したショットのうち，いつ・どのショットに切り替えるべきかを判断する編集要因は様々ある．従来研究では，経験的に決定した規則，視聴者の意図，映画の文法などを編集要因とした自動編集手法が提案されてきた．しかし，机上作業シーン映像編集の難しい点として，各ショットの中で映像として使える期間が短いことが挙げられる．これは作業というものが，一般に素早く常に変化していくものであることが原因である．人物の様子や情報が残る黒板・スライドなどに比べて，提示された物体や個々の操作はごく短い期間しかその状態が持続せず，その期間以外のときは見苦しいショットとなっていることが多い．そのため，いつ・どのショットが映像として使えるのかをより正確に知る必要がある．従来研究で提案されてきた編集要因では，いずれもショットが使える瞬間を逃してしまう危険性が高い[2]．

この問題に対して，話者が視聴者に見て欲しいと思っている箇所を映したショットへの切替えに絞れば，話者はそのショットが使える状態であることを正確に把握しており，更には必要なだけその状態を続けることができる．つまり，話者自らが編集者としてショットを選択すれば，上述したような条件の下でも常に見やすいショットに切り替えることが可能である．よって本研究では，自

[2]逆にいうと，従来研究の対象とするシーンでは，切替え候補であるショットのすべてが比較的安定して撮影されているため，編集を失敗しても映像的に不快感を与えるようなことは少ない．

動編集の要因として話者の意図を利用する．

では，どのようにして話者の意図（注目して欲しい箇所）を特定するか考える．テレビ番組を観察すると，指示動作など視聴者の意識を引きつけるために話者がとる行動が多くみられ，これを合図としてマスターショット[3]からクロースアップに切り替わるパターンが頻繁にみられる．そこで本研究では，話者が明示的な行動によって示す注目すべき箇所を特定し，その部分を強調するように編集する．この編集技法により，ショット切替え前後の因果関係が明確で，教示のポイントも上手く強調された分かりやすい映像となる．ただし，話者の自然な言動による示唆だけですべての注目箇所を網羅することは難しいため，意識的に注目箇所を行動で示すよう話者に協力してもらうことを本研究では前提とする．

テレビ番組の教示シーン映像では，この他にもいくつか典型的な編集パターンがみられる．例えば，一つのショットが長く続くと，映像に変化を出すために別カメラからのショットに切り替えられることが多い．しかし，この編集はショット前後の因果関係が明確でなく，一定時間の経過といった単純な方法で自動化すると，不必要な意図を生んでしまう危険性がある．また，娯楽色の強いテレビ番組では教示の途中でも人物の様子を映したショットなどに頻繁に切り替わるが，これも不用意に切り替えると伝達する教示内容の不足が生じてしまうため，本研究では扱わない．

2.4 映像化の観点からみた本研究の位置づけ

前節では，机上作業シーン映像の特徴と従来研究との比較という観点から本研究の狙いを述べた．本節では，その狙いが，映像化の観点からどう位置づけられるかについて議論する．以下，まず映像化の目的を自動化システムが担うべきものと芸術の領域にあるものに分ける．次に前者の目的のために必要となるルールや技法をまとめ，その中で本研究がどこまで取り組むかについて述べる．

2.4.1 映像化の目的

Steven D. Katz の著書「Film directing: shot by shot (1991)」（日本語版「映画監督術 SHOT BY SHOT」津谷祐司 訳）[4][25] の 12〜13 頁に,『映画は，絵画や写真が完全に表現できないもの，つまり，臨在感を伝えることができるのである．絵画や写真を見ているときは，我々は常に図像の表面を意識してしまうので，臨在感はうまく伝わらない．しかし，映画を見ているときの状況は全く異なっている．観客は投影された映像の表面を見るのではなく，あたかも本物の三次元空間を見ているかのように，スクリーン上の映像空間に取り込まれてしまうのである.』とある．これは，目の前にある世界を表現（復元）することにおいて，絵画や写真に比べて映画がより適しているという見方である．これに基づけば，映像化の目的とは，出来事を個々の平面に切り出して繋ぎあわせることによって，元の出来事を連続的な時空間（シーン）としてスクリーン上で復元すること，つ

[3]話者と作業領域の両方を含み，それだけでシーンの全体が把握できるショット．

まり「シーンの復元」ということになる[4].

一方，Daniel Arijon の著書「Grammar of the Film Language (1976)」（日本語版「映画の文法」岩本憲児，出口丈人 訳）[6][7] の 2〜3 頁には,『映画言語が誕生したのは，動きのさまざまな状態にある短い画像をいいかげんに接合したさいに生じる相違や，これら一連の画像を互いに関連させることができるという考えに映画作家たちが気づいたときだった．映画作家たちが発見したことは，二つの異なるシンボルが結合されるとき，それらは新しい意味を担うということ，またそれらは他のコミュニケーション・システムがそうであるように，一プラス一が三になるという感情や概念，事実を伝達する新しい方法を提供するということだった.』とある．これは，映像化によって現実世界を単純に復元する以上の効果を視聴者に与えることが可能であるという見方である．これに基づけば，映像化の目的とは，現実世界を単に再現するだけでは伝わらない感情や概念などの付加的な意味（メッセージ）を伝達すること，つまり「メッセージの伝達」ということになる．

筆者はこのように，映像化の目的として「シーンの復元」と「メッセージの伝達」の二つを考える．これらは独立しているわけではなく，まずは忠実に「シーンの復元」を行い，その上で「メッセージの伝達」を行うことによってストーリーや感情変化を映像で表現するという二段階の構造となる.「シーンの復元」までを目的とする場合，その目指すべき正解は「映像化によって生じる違和感をできるだけなくすこと」といえる．一方,「メッセージの伝達」は芸術の領域にあり，目指すべき正解というものが一つに決まらず，制作者の感性によって様々な様相を呈する．ときには「シーンの復元」のために守るべきルールを破ることで，新たな意味を表現することさえある．

本研究では，目指すべき正解がより明確な「シーンの復元」までを自動化システムで担うべきであるとする．よって，人物の感情やシーンの雰囲気，娯楽性などを伝えるための技法については本論文では議論しない．映画の文法やテレビの慣習を自動化システムで真似ることも可能であるが，その場合，その研究が映像化という観点においてどこまで実現するのかを定めることは難しいだろう．

2.4.2 シーンの復元

以上の議論より，本論文では「シーンの復元」のために必要となる代表的な映像の撮影／編集技術についてまとめる．ただし，映画制作の技術書には「メッセージの伝達」と「シーンの復元」のための技法が混在して述べられており，また実際に明確に分けられるものでもない．よって，ここでは筆者の経験に基づいて整理した一つの考えを述べる．

シーンをできるだけ忠実に復元することを映像化の目的とする場合，視聴者が映像を現実世界として知覚する際に負荷となるものの削減，言い換えれば，映像化によって生まれる現実世界との差異（過不足）をできるだけ少なくするための技術が必要となる．以下に，映像化によって損失するものをまとめる．

[4]「シーンの復元」だけが映像化の目的であると Katz が述べているわけではない．彼の著書では，他の映画制作に関する技術技術書と同様に，映像化によって生まれる新たな効果を上手く扱うための撮影／編集技術が中心に述べられている．ここでは，映像化の目的に二つの段階を設けるための導入として，彼の著書から一文を引用した．

注視点選択権の損失: 人間は立ち位置や目線を移動することで，注目する対象を好きな位置から見ることができる．映像化されることによって，視聴者は立ち位置（視点）や見る対象（注視点）を選択することができなくなる．これは映像化における最も基本的な損失であり，撮影と編集によって視聴者の視点と注視点を上手く決めてやることにより，映像によって違和感なくシーンが復元される．一方で，視点・注視点の不適切な選択は，シーンで起こっている出来事の因果関係の損失に繋がる．

人間視覚能力の損失: 対象を上手く視野に捉える人間の視覚能力は優れており，対象が素早く不規則に動いていても，目線や焦点の調節を適切に行うことで安定した視野を維持する．このような対象を下手に撮影すると，ブレや遅延が生じて視野がめまぐるしく変化してしまい，映像を見ていると気分が悪くなる．また，人間はピント調節や中心視力などの機能により，注目する箇所に高い解像度で観察することによってシーンを理解する手助けとしている．

空間的連続性の損失: ショットが切り替わった際，人物の画面内における位置や動く方向，視線の方向などがその前後のショットで変わると，視聴者はシーンの三次元的構成が把握できず混乱する．このように，映像では様々な位置から撮影したショットを組み合わせるため，三次元空間中から実際に見たときには起こりえないことが起こってしまう危険性がある．また，同じ被写体を捉えたショットの間での切替えでは，その前後のショットサイズが違い過ぎるとシーンの対応関係が掴めず混乱する．逆に，ショットサイズが同じだと，動いていないのに動いたように見えてしまう（ジャンプカット）．

時間的連続性の損失: 編集によって時間的に要約された映像では，時間を隔てたシーンを連続的に視聴者に提示する．この際，見る側がそれに気づかないように，もしくは時間がどの程度経過したのかが分かるようにしないと違和感が生じる．また，複数のイベントが時間的に並列に行われていても，それらを一つの時間軸上で表現しなくてはならない．つまり，映像としては時間的に前後して見せつつ，印象としてはそれらが同時に行われているように感じさせる必要がある．

映画制作の技術書にまとめられた代表的な技法やルールが，以上の損失をどのように復元するかについて表 2.2 にまとめる．ただし，これらは各損失の復元に一対一で対応しているものでない．

2.4.3 本研究の位置づけ

「シーンの復元」という目的に対し，本研究がどこまで実現しようとしているかを以下にまとめる．

注視点選択権の復元: パン・チルト可能なカメラで多視点からシーンを撮影し，それらのショットを切り替えることで視点・注視点を表現する．注視点は，話者が視聴者に注目して欲しい部分を選択する．話者の行動に基づいたショット切替えにより，アクションとリアクションの関係が常に保たれ，ショット切替え前後の因果関係が維持される．
ただし，話者が注目して欲しい部分と視聴者が見たい部分が重複しない箇所は復元できない．また，人間はある箇所を見ていて飽きると別の場所に注視点を移すことがあるが，これを自動シス

表 2.2: シーンの復元のためのルールと技法の一例

パン・チルト: カメラを上下左右に振ることにより，目線の移動（注視点）を復元する．安定な視野で対象を捉えるためには，興味の対象が視野内でできるだけ固定するようにパン・チルト制御を行なう必要がある．従来研究では，対象が視野の中心からある範囲内にある間はカメラを固定する技術が用いられている [26][27][28]．また，パン・チルト制御の速度調節も人間の視覚能力を補償する重要な要因である．
トラッキング・ドリー: カメラ本体をシーン内で動かして撮影（ドリー）する若しくは，被写体を追うように動かして撮影（トラッキング）することによって視点の移動を復元する．対象とするシーンが狭い場合はパン・チルトで代用することができる．
ショット切替え: マスターショットと呼ばれるシーン全体を広く捉えたショットから，インサートショットと呼ばれるシーンの一部をアップしたショットに切り替える操作により，目線の移動や中心視力の機能を代替し，視聴者の注目を重要な箇所に誘導できる．他にも様々なシーンに応じた編集パターンが映画制作の技術書にまとめられているが，ここに述べたパターンは教示シーン映像で典型的に使用されるものである．
アクション・リアクションの維持: 視聴者は常に因果関係を把握しようとしながら映像を見ているため，ショット切替えの前後の事象の因果が正しくない，もしくはそもそも因果がない場合，視聴者を混乱させたり，勝手に因果を作り出したりしてしまう．これを避けるため，映像編集では「アクションとそれに対応するリアクションの連続を維持する」という基本原則がある．
ズーミング: ズーミングにより，中心視力や視点移動などによる解像度の調節を復元することができる．また，一つのショット内で対象が視野にうまく収まるように調整することで，見やすい映像が取得できる．更に，ショット内でショットサイズを大きく変化させることにより，ショット切替えと同じ効果をもたせることができる．
構図: 画面内のどこにメインの対象を置くかによって，視聴者に注目して欲しい個所を把握させることができる．映画制作の技術書にはシーンに応じて様々な構図が挙げられているが，安易に構図を決定すると，対象の心象や周辺との関係などの不要な印象を与えてしまうことがある．一般に，対象がシーン中を動いている場合には，その対象を視野の中心に保つようフォローすることによって興味の対象が強調される．
シーンの一致: 切り替えたショットの前後で，メインとなる被写体はスクリーンの同じ側に位置していなければならず（位置の一致），その動きの方向は同じでなければならず（動きの一致），更に，同一の人物の目線は同じ向きでなければならない（目線の一致）．二人の対話では人物の目線は向かい合っていなければならず，三人以上の対話では興味の中心となっている人物を他の人物全員が見るようにする．
カメラの三角原則: 人物を繋ぐ直線，人物が移動している方向，人物が向いている方向などをイマジナリーライン若しくはアクションラインと呼び，その片側に三角形の形でカメラを配置することをカメラの三角原則と呼ぶ．これを守ることにより，そのカメラ間での切り替えでは上述した動きの一致と目線の一致が維持される．
ショットサイズ: ショットサイズには，大きく分けてロングショット・ミディアムショット・クロースアップの三種類ある．ショット切替えの前後で被写体が同じ場合は，隣合わせたショットサイズで繋ぐのが基本である．これは，サイズの違いすぎるショットを繋ぐことによってシーンの理解を妨げたり，同じサイズのショットを繋いでジャンプカットを生じさせないためである．
エスタブリッシングショットの挿入: 一般に，シーンの始めに置かれたロングショットで，視聴者にシーンの場所が変わったことを知らせ，そのシーンの全般的ムードや登場人物の配置状況などの情報を与える．この技法は，シーンが変わったときの空間的・時間的連続性の復元の役割を果たす．
カットアウェイ: 本来はイマジナリーラインを越えたことによるシーンの不一致を解決するために挿入されるストーリー上必然性のあるショット（シーン内の小物など）を指すが，時間が飛んだ際にもこのようなショットを入れることで時間の経過を表現することができる．エスタブリッシングショットがロングショットなのに対し，カットアウェイで挿入されるショットはクロースアップが多い．
平行編集: 複数のイベントが時間的に並列に行われている場合，それらを交互に切り替えることで，一つの時間軸上でこれを表現することができる．また，出来事や人物の現在と過去を交互に挿入するカットバックという手法もある．

テムで行なうと不要な因果関係を生む可能性があるため本研究では扱わない．なお，カメラ本体を移動しないため同じショット内での視点の移動は表現できないが，ほとんどの教示シーンではパン・チルトでこれを代替できる．

人間視覚能力の復元: 対象追跡のためのカメラワークを提案し，映像酔いするようなショットになることを避ける．これらのカメラワークを実装した様々な解像度のカメラを用意することにより，視点移動や中心視力などによる視野の解像度の調節を補償する．これらにより，安定した視野と中心視力の損失が復元できる．

ただし，ズーム制御によるショット内での解像度変更ができないため，より効果的な中心視力の効果を与えられない．また，ホワイトバランスや露出補正などはカメラの性能に依存する．

空間的連続性の復元: 机上作業シーンに限れば，作業机の正面にすべてのカメラを配置し，また登場人物は作業机より前には出てこない．よって，三角配置とイマジナリーラインのルールは破られることがなく，動きの一致と目線の一致が維持される．また，被写体を画面中央に捉えるため，位置の不一致も起こらない．更に，ロングショットは使用しないためショットサイズの急激な変化は生じず，同じサイズのショット間での切替えは行なわないのでジャンプカットも生じない．

ただし，一般的な教示シーン映像を考えた場合，本研究の枠組みではシーンの一致を維持できない可能性がある．これについては，実シーンを対象とした従来研究でも考慮されていない．

時間的連続性の復元: 本研究では撮影と同時に編集を行なうため，時間の連続性は本質的に失われない．遠隔教示を想定したシステムでは，この条件は常に当てはまる．

ただし，e-Learning などを対象としたシステムでは，ショット撮影後にカットなどの編集を行なうものもあり，その場合は時間的連続性の損失について配慮する必要がある．

2.5 システム

2.5.1 設計概念

まず，本研究で設定する条件を以下に挙げる．これは撮影・編集の技法を基本機能のみに絞り，システムの構成を簡潔にするためであるが，日常の教示シーンの記録や遠隔教示といった用途では十分実用的な条件といえる．

撮影から編集まで一貫して自動化する: 教示シーンを複数のカメラで撮影すると同時に，得られたショットを映像切替器に入力し，そのうち一つを選ぶことで編集する．これは，映像を介した遠隔教示でインタラクティブなやりとりを可能とするためである．よって，編集はショットの切替え（カメラ選択）のみとなり，ショットを短くしたり入れ替えたりする編集は行なえない．ただし，テレビ番組映像でもショットの短縮はあまり使われておらず，ほとんどの部分がショット切替えだけで編集されている．また，撮影から編集まで一貫して行なうことにより，話者は撮影しながらその場で編集結果を確認することができる．この点は，編集が失敗する可能性のある自動化システムでは重要である．

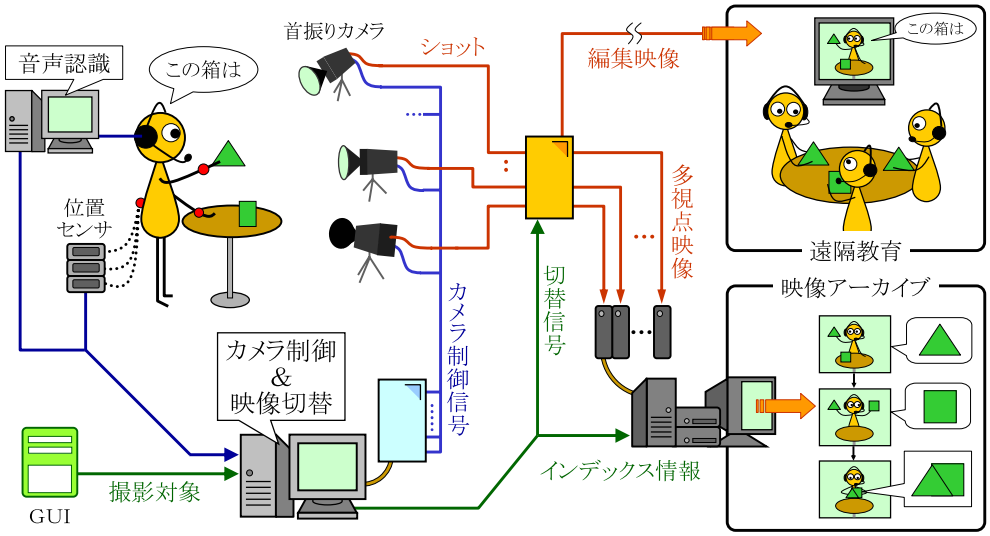


図 2.3: システムの概要

説明・作業する人物は一人とする: 自動撮影では複数人のショットも取得可能であるが，自動編集では一人の話者に関するショットの間での切替えを考える．テレビ番組では娯楽性をもたせるため，話者に加えて数人のアシスタントがいる場合が一般的になっており，複数人で同時に作業する場合も多い．しかし本研究では，個人の知識を映像コンテンツとして記録・送信するという目的から，まずは一人の話者に関するショット切替えに焦点を絞る．

カメラマンとディレクタの連携は考えない: 実際の撮影現場ではカメラマンとディレクタは常に連携しており，ディレクタが欲しいショットをカメラマンに要求し，カメラマンはその曖昧な要望から適切だと思われるショットを判断して撮影する．これに対して本研究では，カメラを必要な数だけ用意しておくことによって，ディレクタの希望するショットはすべて得られているものとする．技術的・社会的背景から普及機の首振りカメラは安価になってきているため，十分に可能な枠組みであると考えられる．

各カメラの撮影対象・カメラワークは 1 種類とする: 本システムでは，人間のカメラマンのように 1 台のカメラを適応的に制御しようとするのではなく，それぞれ違う役割をもって動作する複数台の首振りカメラを使用する．これは，カメラワークの変化や撮影対象の切替えにより，話者がそのショットを使いたいと思った時に使えない状態であることを避けるためである．既に述べたように，カメラを多数用意することで，カメラワークの変化や撮影対象の切替えを補償する．

2.5.2 構成

システムの概要を図 2.3 に示す．カメラ制御・行動検出・映像切替は，表 2.3 に仕様を示した PC 一台ですべて行なっている．カメラに依存する制御遅れ及び音声認識の遅れによる映像切替の遅れ

CPU	Intel Xeon 1.70GHz × 2
メモリ	RDRAM 512MB
OS	Linux kernel 2.4.18
コンパイラ	gcc version 2.95.3

表 2.3: カメラ制御・行動検出・映像切替用 PC の仕様

と比べると，計算処理による遅延は無視できる程度に小さい．なお，使用機材についての詳細とその妥当性については，付録 A を参照されたい．

手先などの素早く動く小さな対象を常に画面内に捉えておくためには，対象の正確な位置をリアルタイムで取得する必要がある．これに対し，本システムでは磁気センサを用いて話者の手や腰などの位置を常に計測し，その位置データを用いて首振りカメラを制御する．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．また，音声認識ソフトを用いて話者の発話内容からキーワードを抽出し，対象の位置やカメラの制御値とともに，映像と同期させて記録する．記録した映像と付加情報は，そのまま映像マニュアルとして利用したり，自動編集して遠隔講義に利用したりすることを想定している [29][30].

2.5.3 キャリブレーション

位置センサから獲得された被写体の位置をカメラで追跡するためには，図 2.4 に示すように，センサ座標系→世界座標系→カメラ座標系の間の変換を行うことが必要となる．これらの変換行列を求めるために，あらかじめキャリブレーションを行う．

まず，空間中に配置した 27 点の参照点の世界座標値，センサ座標値，及びその点を画面中心に捉えたときのカメラのパン・チルト値の組み合わせを計測する．得られたセンサ座標系と世界座標系の座標値の組み合わせから，

$$\begin{bmatrix} W_x \\ W_y \\ W_z \end{bmatrix} = M_{StoW} \begin{bmatrix} S_x \\ S_y \\ S_z \\ 1 \end{bmatrix}$$

で表されるセンサ座標値 (S_x, S_y, S_z) から世界座標値 (W_x, W_y, W_z) への変換行列 M_{StoW} を計算する．次に，カメラのパン・チルト値 (θ_p, θ_t) から，カメラ座標系におけるカメラ中心[5]から参照点への方向ベクトル $(\vec{C}_x, \vec{C}_y, \vec{C}_z)$ を次式によって計算する．ただし，カメラ座標系は垂直方向を C_z 軸とし，カメラは水平に置かれているものとする．

$$\begin{aligned} \vec{C}_x &= \cos\theta_t * \cos\theta_p \\ \vec{C}_y &= -\cos\theta_t * \sin\theta_p \end{aligned}$$

[5]カメラの首振りの駆動中心を指すものとする．

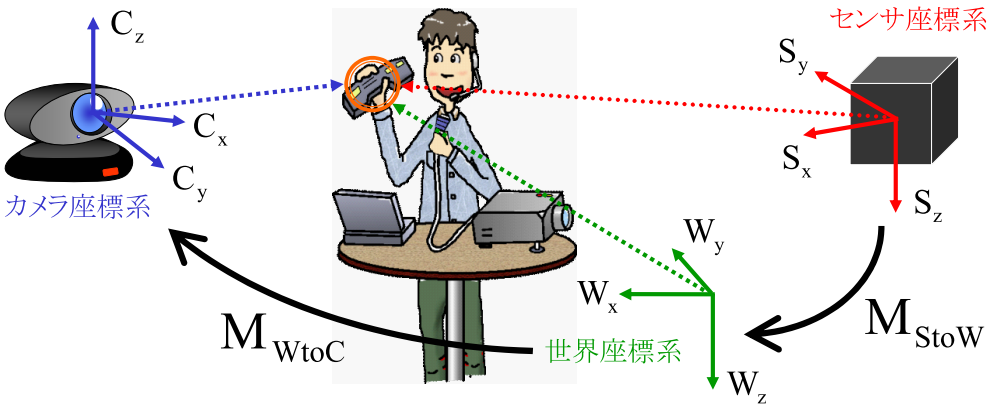


図 2.4: 座標変換

$$\vec{C}_z = \sin\theta_t$$

一方，世界座標系におけるカメラから参照点への方向ベクトル $(\vec{W}_x, \vec{W}_y, \vec{W}_z)$ を計算する．

$$\begin{aligned} \vec{W}_x &= W_x - W_{cx} \\ \vec{W}_y &= W_y - W_{cy} \\ \vec{W}_z &= W_z - W_{cz} \end{aligned}$$

ここで，(W_{cx}, W_{cy}, W_{cz}) はカメラの世界座標系での位置であり，あらかじめ測定しておくものとする．求められた二つの方向ベクトルの組み合わせより，

$$\begin{bmatrix} \vec{C}_x \\ \vec{C}_y \\ \vec{C}_z \end{bmatrix} = M_{WtoC} \begin{bmatrix} \vec{W}_x \\ \vec{W}_y \\ \vec{W}_z \end{bmatrix}$$

で表される世界座標系でのカメラから参照点への方向ベクトル $(\vec{W}_x, \vec{W}_y, \vec{W}_z)$ からカメラ座標系でのカメラから参照点への方向ベクトル $(\vec{C}_x, \vec{C}_y, \vec{C}_z)$ への変換行列を計算する．

以上の手順で求められた変換行列 M_{StoW} と M_{WtoC} により，位置センサから得られた撮影対象の位置をカメラ制御値に変換し，追尾撮影することができる．

2.6 まとめ

本章では，教示シーン映像及び机上作業シーン映像の自動撮影・編集における本研究の考えと狙いを述べた．

まず，映像の定義について，教示シーン映像とは説明者が何かについての教示を行なっているシーンを撮影した映像であり，机上作業シーン映像は机の上で行う作業を説明する人物を撮影した教示シーン映像であるとした．この中で，教示シーン映像の目的は「教示内容を見やすく分かりやすく伝えること」であり，映画やバラエティ映像とは違った技法に主眼が置かれることを述べた．

次に，従来研究と比較して机上作業シーンで顕在化する問題への対処として，1) 見やすい映像を取得するためのパン・チルト制御による追跡撮影，及び，2) 話者の言動によって示唆される注目すべき箇所を強調する編集に焦点を絞ることを述べた．また，映像化という観点から，本研究が「シーンの復元」のために必要となる基本的な課題をクリアしていることを述べた．

最後に本研究のベースとなるシステムの構成について述べた．このシステムでは，磁気センサや音声認識ソフトなど出来合のセンシング技術を組み合わせることで構成している．このシステムの上に 3 章と 4 章で提案する手法を実装することで，初めて自動撮影・編集システムが完成する．言い換えると，出来合の技術を単純に組み合わせるだけでは満足のいく撮影・編集は実現できないということである．

第3章　カメラマンの機能の実現

3.1　はじめに

本章では，カメラマンの機能を自動化システムで実現するための手法を提案する．これまで，机上作業シーンの自動撮影に関する先行研究はあまり行われておらず，何を対象に撮影するカメラが必要で，それらがどのように制御されるべきかという知見が未だ体系づけられていない．そこで本研究では，まずテレビ番組の机上作業シーン映像における典型的なショットとカメラワークの特徴を調べた．その結果を基にして，自動化システムのための撮影対象を定義し，カメラワークを実現するためのカメラ制御手法を提案する．ここで提案した撮影対象の分類とカメラ制御手法は，CGアニメーションで作成した仮想シーンに対する疑似撮影実験と構築したシステムによる実シーンに対する撮影実験の二つの設定において評価した．

ここで提案する制御手法は以下の考えに基づいている．

- 対象に応じたカメラワークを用いて追跡撮影することによってブレや振動のない安定した視野で対象を捉え，見やすい映像を撮影することに焦点をおく．
- 代表的なショットとそれを撮るためのカメラワークの対応関係を予め一覧にまとめておき，映像制作の知識がない人でも簡単に設定できるようにする．
- 人間のカメラマンのように1台のカメラを適応的に制御しようとするのではなく，それぞれ違う役割をもって動作する複数台の首振りカメラを使用する．

3.2　カメラマンの機能の基本要求

テレビの料理番組や工作番組などの机上作業を対象とした番組には，ショットとカメラワークに関するいくつかの典型的なパターンがみられる．本節では，これらのテレビ番組映像を調べた結果を用いて，カメラマンの機能として満たすべき基本要求についてまとめる．

調べたテレビ番組の一覧を表3.1に示す．テレビ番組の机上作業シーン映像で頻繁に使われているショットを表3.2に示す．また，映像全体に対して各ショットが占める時間・回数の割合と，個々のショットの継続時間の統計を表3.3に示す．ショットAとショットCの統計を合わせると，時間・回数の両方で映像全体の90 %以上を占めている．特に，作業時の手元のアップショットであるショットC（以下，手元ショットとも呼ぶ）の重要性が確認できる．

表 3.1: 調査したテレビ番組

テレビ番組	シーン数	ショット数	合計時間
料理番組 A	5	180	49 分
料理番組 B	6	174	41 分
科学実験番組 A	15	356	50 分
科学実験番組 B	8	189	22 分
工作番組 A	21	425	60 分
工作番組 B	5	399	22 分

次に，各ショットで用いられているカメラワークについてまとめる．カメラワークに注目してテレビ番組映像を調べると，カメラマンがどのようなカメラワークを使用しているかは，表 3.2 に挙げたショットのカテゴリではなく，捉えようとする対象の様子（状態）に深く関係していることがわかった．例えば，同じショット C の撮影でも，(a) 作業対象である物体の “外観” に注目する場合には，物体をできるだけ画面中央に捉えることが優先され，(b) 手の “動き” に注目する場合には，その動きの範囲の中心付近を捉えることが優先され，(c) 操作の “周辺関係” に注目する場合には，操作全体とその周辺を含んだ状態でカメラをできるだけ固定することが優先されている．これらのカメラワークの違いは，注目する対象の状態に応じて「対象を画面中央に捉えること」と「視野を固定すること」をカメラマンが適応的に調節していると考えられる．これらの特徴は表 3.2 のショット A〜C のすべてにみられるが，特にショット C において顕著である．

本研究で解決すべき課題は，実際のスタジオではディレクターの指示とカメラマンの技術によって取得されているこれらのショットを，自動化システムによって取得することである．特に変化の激しい手元を適切に撮影することは，講義シーンの撮影では顕在化しない机上作業シーン撮影に特徴的な問題であるといえる．

3.3 撮影対象の分類

3.3.1 撮影対象の定義

カメラマンは，撮影対象のどういう状態に注目するかによって適応的にカメラワーク（追跡方法）を変化させている．自動化システムで実現するカメラワークでも，何を画面に捉えるかだけではなく，その対象のどういう状態をうまく捉えたいのかという意図を反映させたものでなければならない．そのため本研究では，撮影対象を「何（**注目対象物**）のどういう状態（**注目すべき状態**）を見たいのか」，つまり,「撮影対象 = 注目対象物 + 注目すべき状態」として定義する．

注目対象物の分類を表 3.1 に示す.「手先」のカテゴリには，更に右手・左手・両手の区別がある．両手を対象物とする場合は，右手と左手の中点を追跡撮影する．表 3.2 に挙げたショットとの関係を考えると，ショット A とショット B が「話者」に，ショット C が「手先」「物体」「場所」に対応

表 3.2: 机上作業シーン映像を構成するショット

ショット A： 人物と作業領域の両方を含む範囲で撮影されたミディアムショット．話者／アシスタントが単独で含まれる場合と，二人以上が同時に含まれる場合がある．あまりカメラを動かさず，固定した視野で撮影されることが多い．作業全体の流れを掴むためのマスターショットとして利用される．

ショット B： 人物のバストショットかクロースアップで，作業領域は含まないショット．バストショットの場合はあまりカメラを動かさず固定した視野で，クロースアップの場合は顔を追跡しながら，それぞれ撮影されることが多い．

ショット C： 作業領域のみを含んだクロースアップ．作業領域とは，作業中の手元，作業対象の物体，作業に関連する物体や場所をいう．一つのショット中で，それら撮影対象が連続的に変化する場合もある．対象のどのような状態を撮影したいかによって，カメラワーク（追跡方法）が変わる．

ショット D: CG テロップや特殊効果など，カメラで撮影されたものではないショット．主に材料の名前や分量，作業のポイントなどの情報を示す．専門用語では，インサートと呼ばれることもある．

する．本システムでは，注目対象物の指定によって画面中央に捉えるべき対象物とカメラの視点・視野が決まる．

次に**注目すべき状態**の分類を表 3.4 に示す．この分類は，机上作業シーン番組の典型的なショット

表 3.3: 各ショットの出現頻度

	ショットの種類			
	A	B	C	D
出現回数の割合	44.3 %	9.8 %	45.1 %	0.8 %
出現時間の割合	31.6 %	6.0 %	61.1 %	1.3 %
継続時間の平均	6.0 s	5.2 s	11.5 s	14.5 s
継続時間の標準偏差	4.7 s	4.9 s	12.5 s	11.9 s

図 3.1: 注目対象物の分類

の撮影意図が,「対象の外観を見せたい」「対象の動きを見せたい」「対象とその周辺の関係を見せたい」の三つに密接に関係しているという考えに基づいている．注目すべき状態の各々に適したカメラワークを表 3.4 に同時に示した．ここで重要なのは，図 3.2 に示すように，3.2 節で簡単に述べた次の二つの要求がトレードオフの関係にあるという点である（以下，これをカメラ制御のトレードオフと呼ぶ）．

1. 対象の姿がわかりやすいように，対象をできるだけ画面中央に捉える．
2. 視野の激しい変化によって映像が見苦しくならないように，できるだけ視野を固定する．

本研究では，このトレードオフを調整するためのカメラ制御手法を提案する．本手法では，注目すべき状態を指定することによっていくつかのパラメータが決まり，それにより上記トレードオフの調整を含むカメラの動作特性が決まる．

3.3.2 撮影対象の設定

表 3.2 に挙げたショットを撮影する場合の具体的な設定例を以下に示す．

ショット A： 人物と机上の作業領域を含んだショットであるが，人物がアシスタントと話しながら

表 3.4: 注目すべき状態の分類と各々に適したカメラワーク

注目すべき状態	説明	カメラワーク
＜外観＞ ＜appearance＞	提示された物体や動きの伴わないジェスチャ（サイン）などの外観に注目する	対象が停止すると，素早くそれを画面中央に捉えてカメラを固定する．対象が画面中央から外れたら，すぐに追跡を再開する．
＜動き＞ ＜movement＞	操作の様子や動きを伴うジェスチャなど，手や物体の動きに注目する	頻繁に動く対象をカメラで逐一追跡すると見苦しい映像となるため，できるだけカメラを固定して撮影する．対象の動き全体が画面内に入るように，適当な間隔でカメラを微調整する
＜周辺関係＞ ＜circumstance＞	机上での物体の移動や広範囲の動作など，対象とその周辺（背景）との関係に注目する	対象とその周辺の関係を捉えるため，カメラは対象が視野から出そうになるまで固定しておく．対象が周辺と関係をもっていないときは，対象を画面中央に捉えて，ゆっくりと追跡する

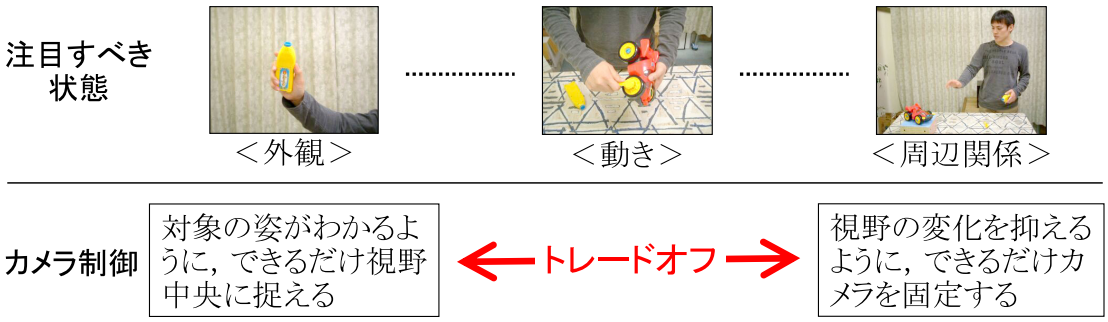


図 3.2: カメラ制御のトレードオフ

作業している，若しくは机上の広範囲にある物体に対して操作している状態が多い．これは注目対象物を「話者」，注目すべき状態を＜周辺関係＞とすることで表現できる．

ショット B： 話者の顔や上半身のみのショットであるが，話者がほとんど動かずに話している状態が多い．これは注目対象物を「話者」，注目すべき状態を＜外観＞とすることで表現できる．

ショット C： 作業時の手元のアップショットであるが，提示された物体の外観や作業時の手の動き，持っている物体と周辺の関係など，表 3.4 に挙げたすべての注目すべき状態が同様に出現する．また，注目対象物も「手先」「物体」「場所」の三つが同様に出現する．ただし,「場所」については固定カメラで撮影できるため,{「手先」,「物体」} ＋ {＜外観＞，＜動き＞，＜周辺関係＞} の組合せで表現できる．

ショット D： CG テロップや特殊効果などのショットであるが，これは CG などで作成されたものであるため，本研究では扱わない．

以上は各ショットの典型的な場合であり，ショット A やショット B でも撮影するシーンによってはすべての組み合わせがあり得る．

実際に撮影を行うときは，ユーザが各カメラの撮影対象の設定を行う．本システムは各カメラの撮影対象を設定するための GUI を備えており，ユーザは各カメラに撮影させたい注目対象物とその注目すべき状態を上記のカテゴリから選択するだけでよい．これについては，3.4.2 節で述べる．

3.4 カメラ制御手法

3.4.1 可変枠制御

表 3.4 に挙げたカメラワークを自動化システムで実現するには，カメラ制御のトレードオフを調整する機能が必要である．また，手先などの素早く不規則に動く対象をアップで撮影するため，カメラ制御の遅れや視野の変化によって見苦しい映像となる問題も解決しなければならない．

これらの問題を解決するため，本研究では**可変枠制御**を提案する．本手法は次の二つの制御モードをもつ．

固定モード: 画面内に仮想的な枠を設定し，画面上の対象がその枠内に留まっている間はカメラを固定する．固定モード中は，定期的（枠調整間隔）に過去数秒間（枠調整区間）の対象の軌跡の重心が，画面中心にくるようにカメラを制御する．画面上の対象が枠外に出たら，追跡モードに切り替える．

追跡モード: 対象をできるだけ画面中央に捉えるように追跡し，対象の反復，若しくは停留を検出すると固定モードに切り替える．反復は，画面上で対象の動き方向が反転した回数を数え，一定回数（反復判定回数）反転することで検出する．停留は，一度停止状態を検出した後，画面上の対象が一定範囲内（停留しきい値）に一定時間（停留判定時間）留まることで検出する．

反復を検出してカメラを固定することにより，手元作業などを撮影した場合に視野が激しく変動し，見苦しい映像となることを回避できる．また，停留を検出してカメラを固定することにより，対象の小さな揺らぎまで追跡することによる視野のぶれを抑えることができる．固定モードと追跡モードの切替えは，パン方向とチルト方向で独立して行なう．

ユーザは，画面上に配置する枠の大きさ（枠サイズ）を画面に対する比率で指定する．制御プログラムの中で枠サイズはパン角度（チルト角度）の範囲に変換され，対象を追跡する際のパン角度（チルト角度）と比較することで「対象が枠内で動いた」若しくは「対象が枠から出た」ことを判断する．画面比で指定された枠サイズを角度範囲に変換する際にカメラの画角の値が必要となるが，これについては，各ズーム値に対する水平（垂直）画角はあらかじめ調べておく．

反復の検出は次のようにして行なう．追跡対象の画面上におけるパン方向（チルト方向）の速度の符号が変化すると，対象が反転したとする．対象が一定回数（反復判定回数）以上連続して反転

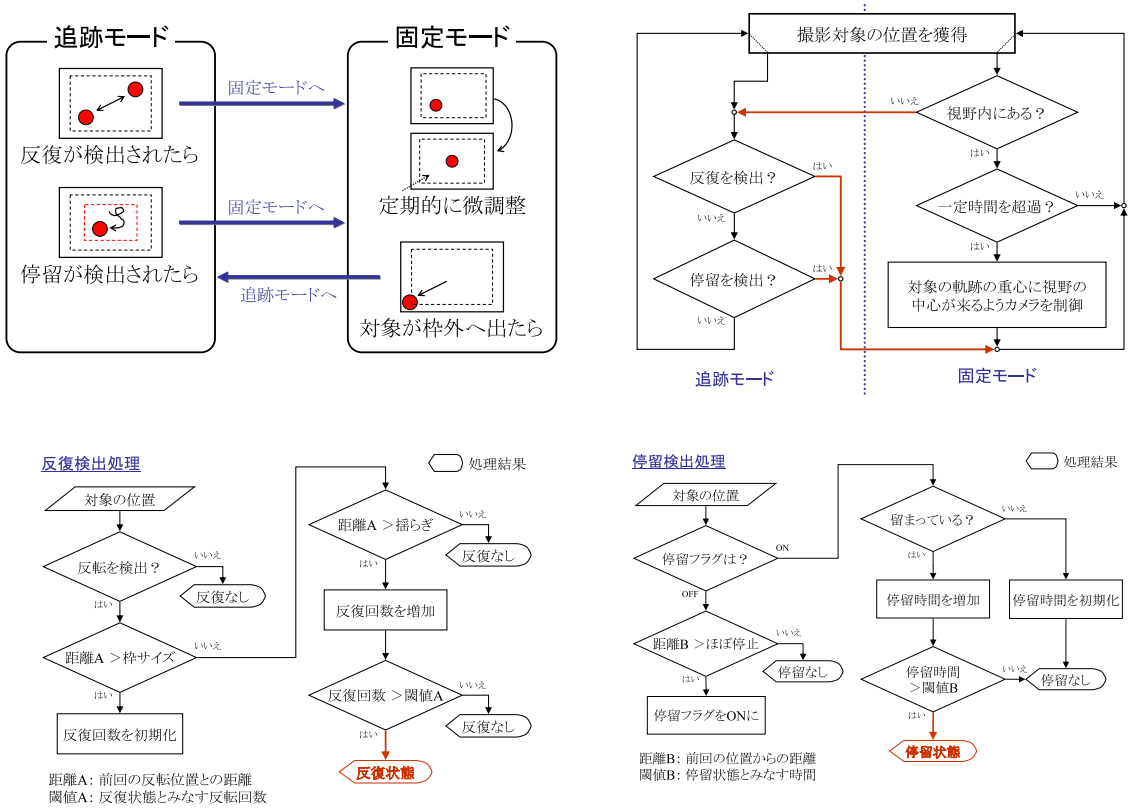


図 3.3: 可変枠制御の流れ（全体の流れを 1 段目に，反復検出と停留検出の処理の流れを 2 段目にそれぞれ示す）

すると，反復しているとみなして固定モードに入る．対象が枠から出た場合，反転回数は初期化される．なお，現在のシステムでは画面の大きさの 5 %以下の動きを停止とみなし，この範囲での反転は無視する．

停留の検出は次のようにして行なう．まず，対象が一度停止した位置を記憶し，そこを停留の中心とする．その後，対象がその位置から一定範囲内（停留しきい値）に一定時間（停留判定時間）留まっていると，停留しているとみなして固定モードに入る．停留とみなされる前に対象が停留しきい値より大きく移動すると，停留の中心を解除し，停留検出の処理を初期化する．

固定モード時のカメラの微調整は次のようにして行なう．固定モードに入って一定時間（枠調整間隔）経過すると，その時点から一定時間（枠調整区間）遡った時点からの対象の軌跡の重心を計算し，その重心の位置に画面の中心がくるようカメラを微調整する．枠調整区間は枠調整間隔に対する割合（0〜1）で指定する．なお，カメラの微調整では追跡の始点と終点が分かっているため，人間のカメラマンの制御を参考にして，動き始めと動き終わりの速度がゆっくりとなるように制御している．

また，可変枠制御では，対象の位置計測過程で生じるノイズを除去するために，剛体の等速運動

表 3.5: カメラワークパラメータ

枠サイズ［画面比］
画面内に仮定する枠の大きさ．これは注目すべき状態まで考慮したときの対象の意味的な大きさを表現している．
反復判定回数［回］
対象を「反復している状態である」とみなす，対象の画面上での反転（方向転換）の回数．
停留しきい値［画面比］
対象を「留まっている状態である」とみなす，画面中央からの一定範囲．対象が一度停止した後に有効になる．
停留判定時間［秒］
対象を「停留している状態である」とみなす時間．対象が一度停止した後，画面上の対象が停留閾値内にいる時間を計測する．
枠調整間隔［秒］
固定モード時にカメラを微調整する時間的な間隔．画面の中心が対象の動いた軌跡の重心位置にくるよう調節する．
枠調整区間［割合］
枠制御間隔に対する割合 (0〜1) で表す．固定モード時の微調整のための重心計算には，過去のこの区間（時間）の軌跡データを用いる．
平滑化度合 ［—］
追跡モード時にどのくらい忠実に対象を画面中央に捉えるかの度合．値が小さいほど平滑化の効果が大きい．

モデルを用いたカルマンフィルタを利用している．更に，カルマンフィルタに与えるノイズ分散比を意図的に調整することで，追跡モード時に対象をどの程度忠実に画面中央に捉えるかの度合（平滑化度合）を設定する．使用しているカルマンフィルタの状態変数 $\mathbf{x}_k$ と状態遷移行列 $\mathbf{F}$ は以下のとおりである．

$$\mathbf{x}_k = \begin{pmatrix} x \\ \dot{x} \end{pmatrix} \quad \mathbf{F} = \begin{pmatrix} 1 & \Delta \\ 0 & 1 \end{pmatrix}$$

ここで，Δ はサンプリング間隔，$\mathbf{x}_k$ は対象の現在の位置，速度を含んだ状態ベクトルである．カルマンフィルタの計算方法とノイズ分散比については付録 A で詳しく述べる．

アルゴリズムの流れを図 3.3 に示す．画面上に仮定する枠の大きさ（枠サイズ）などのカメラワークパラメータを適切に設定することで，カメラ制御のトレードオフを調整することができる．各カメラワークパラメータの意味を表 3.5 に示す．

3.4.2 カメラワークの設定

可変枠制御は次のような特徴をもっている．

- シナリオなどの事前情報を必要とせず，対象の位置情報のみを入力としたアルゴリズムである
- カメラワークパラメータを変化させることにより，一つのアルゴリズムで様々なカメラワークを表現できる

これらの特徴により，対象上の 1 点の位置のみを入力として，表 3.4 に挙げた机上作業シーンに代表される三つのカメラワークを表現することができる．

以下に，それぞれのカメラワークを可変枠制御で実現する方法について述べる．

＜外観＞用カメラワーク (A)： 停留判定時間と枠サイズを小さくすることで，対象の静止に素早く反応し，対象を画面中央に捉えた状態でカメラを固定できる．対象が画面中央から少し離れたら，追跡モードに切り替えて再び画面中央に捉える．

＜動き＞用カメラワーク (M)： 反復判定回数を小さくすることで，手元作業などを撮影した場合に生じるカメラの振動を抑え，対象の動いている様子を安定した視野で捉えることができる．対象の動きの範囲は広がりをもち，その位置は徐々に変化するため，枠サイズをある程度大きく，枠調整間隔を短めに設定する．

＜周辺関係＞用カメラワーク (C)： 対象がその周辺（背景）に対してどのように振る舞っているかを捉えるため，枠サイズを大きく，枠調整間隔を長めに設定する．反復判定回数と停留判定時間を大きくすることで，対象がどの辺りを中心にして移動（または静止）しているのかを推定してからカメラを固定する．

それぞれのカメラワークパラメータ設定を図 3.4 に示す．具体的な値は，実際の撮影環境に依存するため，各実験の節で述べる．実際には，これらのパラメータの値を直接決めることは一般のユーザにとって難しいため，図 3.5 に示すように，各々の撮影対象に適したカメラ制御パラメータセットを予め用意した．一覧表より撮影したい対象を選択することで，各カメラの解像度とカメラ制御パラメータが自動的に設定される．現在のシステムでは,「4 種類の注目対象物× 3 種類のズーム値」の合計 12 種類の撮影対象を用意している．

3.5 仮想シーンにおける評価実験

撮影対象の分類と提案したカメラ制御手法を実装し，それらについての評価実験を行った．評価実験で明らかにするのは次の 2 点である．

1. カメラワークに関する分類の有効性
2. 他の典型的なカメラワークに対する優位性

評価実験の手順を以下に示す．

1. 表 3.4 に挙げた三つの注目すべき状態が典型的に現れているシーンを用意し，評価対象とするカ

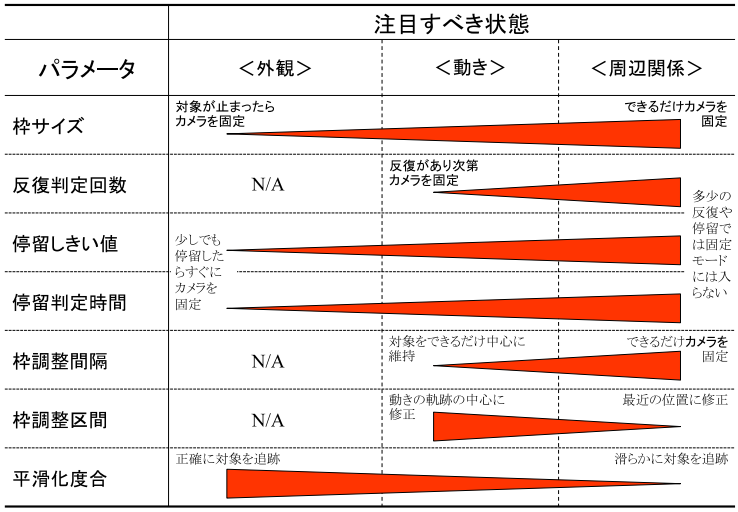


図 3.4: 注目すべき状態のためのカメラワークパラメータの概要（三角形の幅が広くなるに従って，値は大きくなることを表す）

図 3.5: 撮影対象の設定

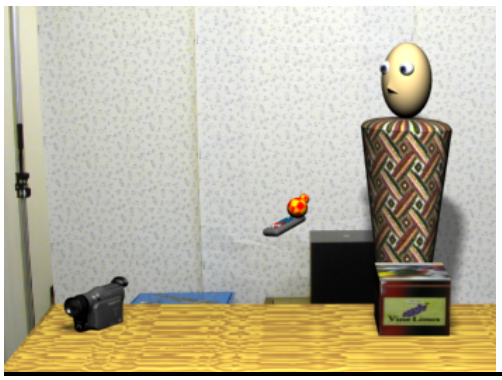
メラワークを用いて撮影する．

2. シーンごとに得られたショット群から任意のペア（ショット対）を作って，被験者に提示する．
3. 被験者は提示されたショット対を見てどちらが良いかを判定し，サーストンの一対比較法によって評価する．サーストンの一対比較法については，付録 B で解説する．

CG アニメーションを用いて仮想シーンを作成することにより，実シーンでは実現が困難な環境で実験することができる．

- 全く同じ位置に全く同じ性能の仮想カメラを何台でも置くことが可能である
- 対象位置の観測に伴うノイズや遅れ，カメラ制御の遅れなどが全くない理想的な条件でのシミュレーションが可能である

表 3.6: 主観評価実験のために用意した CG シーンと評価基準

	**シーン** 1 **内容**：リモコンとビデオカメラを一つずつ持ち上げて見せ，それぞれについて説明するシーン **評価基準**：映像が見苦しくない & 提示物体が画面中央に捉えられている **注目対象物**：「物体」 **注目すべき状態**：＜外観＞
	**シーン** 2 **内容**：大きな箱から小さな箱を取り出してポットを取り出し，ポットを上下に振るシーン **評価基準**：映像が見苦しくない & 作業全体が画面中央に捉えられている **注目対象物**：「手先（両手）」 **注目すべき状態**：＜動き＞
	**シーン** 3 **内容**：話者が作業机を端から端へと歩き回り，リモコンでビデオカメラを操作するシーン **評価基準**：映像が見苦しくない & 話者とその周辺との関係がよくわかる **注目対象物**：「話者」 **注目すべき状態**：＜周辺関係＞

仮想シーンにおける評価実験のために作成したシーンの例とその説明（シーンの内容，評価の基準，注目対象物，注目すべき状態）を表 3.6 にまとめる．仮想シーンを用いた評価実験では 16 人の大学院生に評価してもらった．

3.5.1 カメラワークに関する分類の検討

表 3.4 の注目すべき状態にそれぞれ対応するカメラワークを用いて撮影したショットの間で，有意な違いがみられるかどうかを調べる．＜外観＞用カメラワーク（A），＜動き＞用カメラワーク

表 3.7: 仮想シーンにおける評価実験でのカメラワークパラメータの設定（“—” は機能を無効にしていることを表す）

パラメータ	＜外観＞	＜動き＞	＜周辺関係＞	不感領域制御
枠サイズ［画面比］	0.25	0.7	0.95	0.5
反復判定回数［回］	—	2	6	—
停留しきい値［画面比］	0.2	0.4	0.5	0.5
停留判定時間［秒］	0.5	2.0	5.0	0.5
枠調整間隔［秒］	—	2.0	12.0	—
枠調整区間［割合］	—	1.0	0.25	—
平滑化度合［—］	15	10	5	15

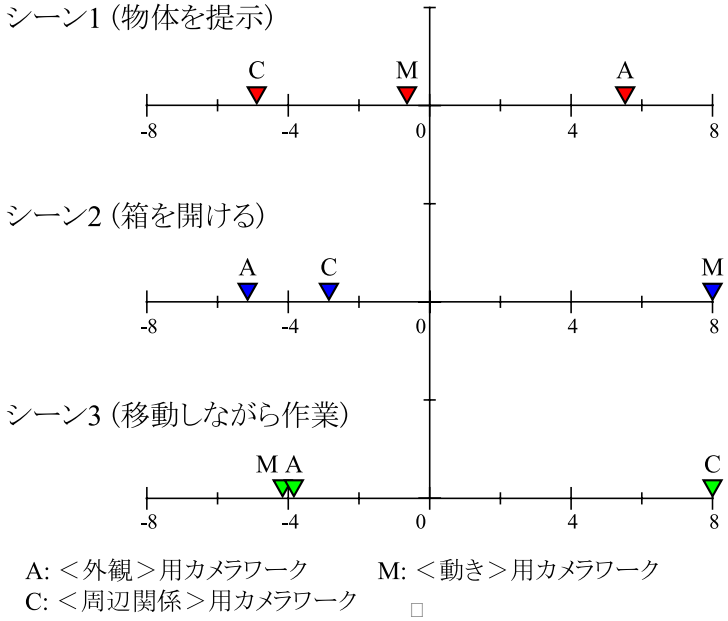


図 3.6: 仮想シーンにおける提案した三つのカメラワークの間での評価結果

（M），＜周辺関係＞用カメラワーク（C）を用いて表 3.6 の Scene1〜3 を擬似的に撮影し，得られたショット群からペアを作って一対比較を行った．評価実験で使用した各カメラワークの具体的な設定を表 3.7 に示す．

図 3.6 に評価の結果を示す．評価は尺度の値が大きくなるほど良いことを表す．Scene1〜3 のすべてにおいて，各シーンの注目すべき状態に対応するカメラワークが最も良い結果となった．以上の結果から，カメラワークに関する本研究の分類が有効であるといえる．

表 3.8: その他の典型的なカメラワーク

遅れなし単純追跡（U）：
対象の位置観測に伴うノイズや遅れ，カメラ制御の遅れがない理想的な条件での単純追跡制御．理想的な単純追跡で撮影すれば，見やすい映像となるのかどうかを確認する．
不感領域制御（T）：
カメラの振動を抑えるだけであれば，ある一定のカメラが反応しない領域（不感領域）をもって追跡すればよい．可変枠制御の機能の一部を利用して，対象が画面中心から 5 割の範囲内にある間は追跡しないように制御する[1]．可変枠制御の停留検出/反復検出の機能が必要かどうか，また，注目すべき状態によってカメラワークを変更する必要があるかどうかを確認する．
平滑化の調整追跡（K）：
可変枠制御のすべての機能を使用しなくても，平滑化度合の強弱だけでカメラ制御のトレードオフを表現できる可能性がある．＜外観＞，＜動き＞，＜周辺関係＞に対応したカメラワークについて，それぞれ平滑化度合を 15，10，5 に設定したカメラワーク．可変枠制御すべての機能が本当に必要かどうかを確認する．

3.5.2 他の典型的なカメラワークに対する優位性

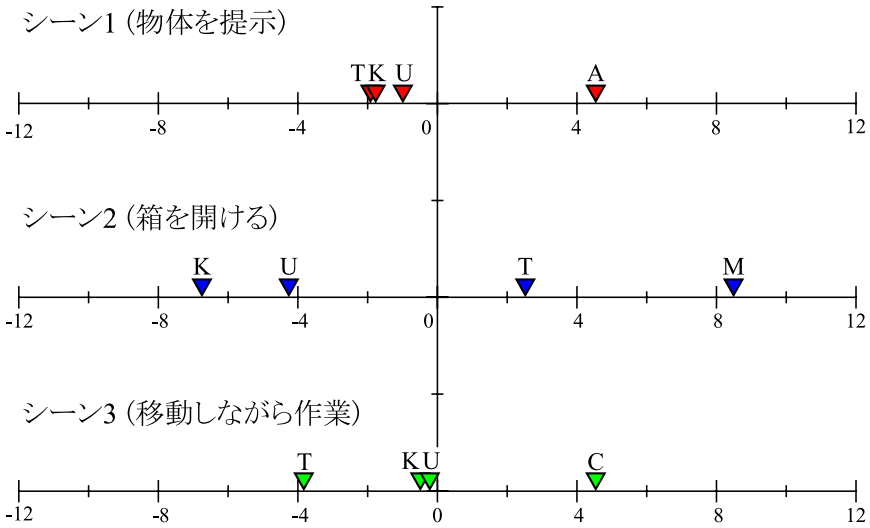
他の典型的なカメラワークと可変枠制御を用いたカメラワークを比較するために，表 3.8 に示す “遅れなし単純追跡（U）”，“不感領域制御（T）”，“平滑化の調整追跡（K）” の三つのカメラワークを用意した．これら三つのカメラワークと可変枠制御によるカメラワークを用いて表 3.6 の Scene1〜3 を擬似的に撮影し，得られたショット群からペアを作って一対比較を行った．

図 3.7 に結果を示す．Scene1〜3 のすべてにおいて，可変枠制御によるカメラワークが最も良い結果となった．遅れなし単純追跡（U）との比較により，ノイズやカメラ制御遅れの影響のない理想的な場合でも，単純追跡では良い映像とはならないことが確認できる．また，不感領域制御（T）や平滑化の調整追跡（K）との比較により，固定サイズの不感領域や平滑化の強弱だけでは，複数の注目すべき状態を捉えるための機能が不足していると考えられる．以上の結果より，ここで提案する可変枠制御が他の典型的な制御方法よりも優れているといえる．

3.6 実シーン撮影による評価実験

仮想シーンで検討したカメラワークを撮影システムに実装し，実際の作業を撮影したショットを用いて評価実験を行った．本実験では，カメラワークに関する分類の有効性について検討する．評価実験の手順は，前節の仮想シーン撮影による評価実験と同様である．ただしノイズの影響を考慮

[1] これは NHK 放送技術研究所の不感領域制御 [31] を論文からわかる範囲で模倣したものである．



T: 不感領域制御　　U: 遅れなし単純追跡　　K: 平滑化の調整追跡
A/M/C: ＜外観/動き/周辺関係＞用カメラワーク

図 3.7: 仮想シーンにおける可変枠制御を用いたカメラワークとその他の典型的なカメラワークの間での評価結果

表 3.9: 実際の撮影システムにおける評価実験でのカメラワークパラメータ設定（“—” は機能を無効にしていることを表す）

パラメータ	＜外観＞	＜動き＞	＜周辺関係＞
枠サイズ［画面比］	0.4	0.7	0.95
反復判定回数［回］	—	2	6
停留しきい値［画面比］	0.2	0.4	0.5
停留判定時間［秒］	0.5	2.0	5.0
枠調整間隔［秒］	—	2.0	12.0
枠調整区間［割合］	—	1.0	0.25
平滑化度合［—］	12	7	3

して，すべてのカメラワークの平滑化度合を大きめに設定し，＜外観＞用カメラワークの枠サイズをやや大きく変更した．カメラワークパラメータの具体的な設定を表 3.9 に示す．

撮影システムを用いた評価実験では，4 台のカメラを近接して配置し，同じ解像度でカメラワークのみ変更して撮影した．＜外観＞用カメラワーク，＜動き＞用カメラワーク，＜周辺関係＞用カメラワーク，単純追跡の四つのカメラワークを用いて表 3.10 の Scene1′〜3′ を撮影し，得られたショット群からペアを作って一対比較を行った（表 3.10 に示すように，Scene1′〜3′ はそれぞれ表 3.6 の Scene1〜3 とほぼ同様の内容である）．

なお，カメラの制御遅れは映像にして 10 フレーム弱[2]，対象位置の観測ノイズは実際の空間にし

[2]使用したカメラは Sony EVI-D100 であり，クロースアップショットの場合（60cm の範囲を画面いっぱいに捉えた

表 3.10: 主観評価実験のために用意した実写シーンと評価基準

	**シーン** $1'$ **内容**：Scene1 を模擬したシーン．ただし，注目対象物を物体ではなく手先（右手）としている． **評価基準**：映像が見苦しくない & 提示物体が画面中央に捉えられている **注目対象物**：「手先（右手）」 **注目すべき状態**：＜外観＞
	**シーン** $2'$ **内容**：Scene2 を模擬したシーン．ただし，ポットの代わりにカメラを用いた． **評価基準**：映像が見苦しくない & 作業全体が画面中央に捉えられている **注目対象物**：「手先（両手）」 **注目すべき状態**：＜動き＞
	**シーン** $3'$ **内容**：Scene3 を模擬したシーン． **評価基準**：映像が見苦しくない & 話者とその周辺との関係がよくわかる **注目対象物**：「話者」 **注目すべき状態**：＜周辺関係＞

て 1cm 程度である．また，Scene1′ の撮影では，注目対象物を物体（リモコン，ビデオカメラ）ではなく，右手にセンサをつけることで擬似的に把持物体を追ったショットを撮影している．この是非に関する議論は付録を参照されたい．

評価実験は，仮想シーンで行った実験と同様の方法で，大学院生 17 人に対して行った．結果を図 3.8 に示す．結果は仮想シーンの場合とほぼ同じであり，仮想シーンで検討したカメラワークが実際のシステムでも有効であることを確認した．

場合），画面比で中心から 2 割ほど遅れた映像となる．一般的な机上作業における手先の速度は最大で約 300cm/s にも達するため，対象が視野から完全に外れることもある．

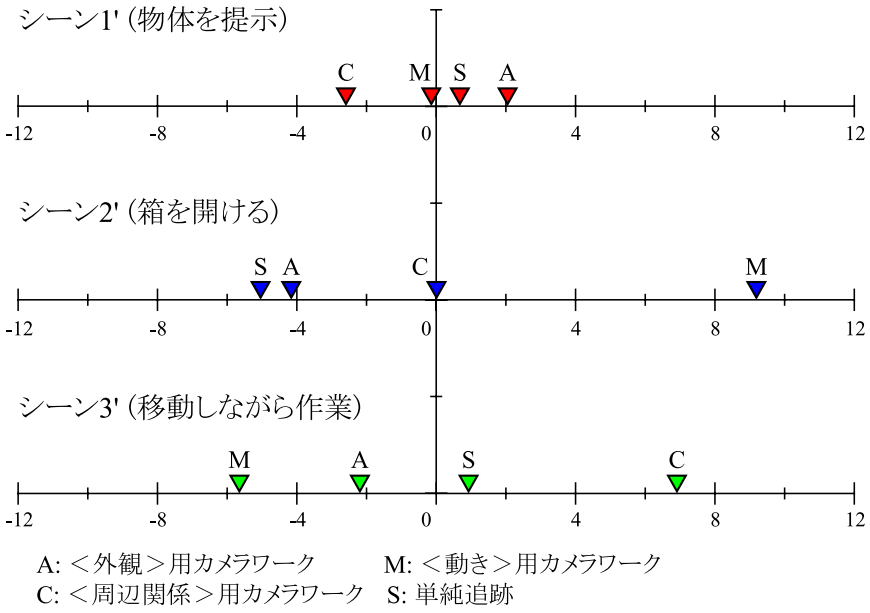


図 3.8: 実際のシステムを用いた撮影における提案した三つのカメラワークと単純追跡の間での評価結果

3.7 プロカメラマンによる映像との比較

プロカメラマンの撮影した映像と本手法を用いて撮影した映像を，5 段階の評定尺度法によって主観的に評価した．評価に使用した映像は，単純追跡で撮影したショット，可変枠制御によるカメラワークで撮影したショット，カメラマンによる料理映像（編集前）のショットの 3 種類で，可変枠制御と単純追跡によるショットは 3.6 節で使用したもの，カメラマンのショットは評価用映像メディアデータベース [32] から表 3.4 の三つの注目すべき状態が典型的に現れているシーンを撮影した部分を切り出したものを用いた．よって，本手法を用いて撮影したショットとカメラマンによるショットは撮影したシーンが異なり，カメラマンによるショットのほうが画質・内容ともに品質が高い．このような条件の下で，本手法を用いて撮影した映像がカメラマンにどこまで迫れるかが焦点となる．

評定要素は，加藤らによるカメラワークを評価するための四つの尺度（連続感・鮮明感・好感・人間的）[27] に，教示シーン映像で重要となる “内容理解のしやすさ” を加えた五つの尺度について，それぞれ二つずつ形容詞句を割り当てた．表 3.11 に五つの尺度と形容詞句を示す．被験者はテレビを日常的に見ている大学院生 14 人とし，可変枠制御・単純追跡によるショット六つとカメラマンによるショット三つの計九つを用いて評価実験を行った．評価値は，尺度ごと（二つの形容詞句ごと）の全カメラワークのスコアの平均 μ と標準偏差 σ がそれぞれ 0 と 1 になるように，カメラワークごとのスコア平均 s を正規化している．

結果を図 3.9 に示す．結果より，“鮮明感”・“好感”・“内容理解のしやすさ” についてはプロカメラ

[3]評価実験では「的確とは具体的にどういう意味か」という質問があったため,「注目すべき対象を的確に捉えている

表 3.11: 形容詞群

尺度	1 点 — 5 点
連続感	ギクシャクした — スムーズな
	途切れた — 連続した
鮮明感	見づらい — 見やすい
	乱れた — 整った
好感	不快である — 快適である
	味気のない — 味気のある
人間的	機械的な — 人間的な
	人工的な — 自然な
内容理解	理解しにくい — 理解しやすい
	的確でない — 的確な[3]

マンの撮影したものと遜色ないが，“連続感”・“人間的” については差がみられた．評価実験後に行ったアンケートより，プロカメラマンの撮影技術には以下のような良い性質があることがわかった．

- わずかなズーム制御が映像に連続感と人間的な印象を与え，また注目すべき部分をうまく強調している
- 対象が画面から出た場合，注目対象の位置を予測しつつ，ゆっくりとカメラを動かすところに連続感がある

今後は，本手法にズーム制御やシーンの 3 次元構成についての知識を用いた制御を取り入れることにより，特に “連続感” についての改善を行っていく必要がある．

3.8 まとめ

机上作業シーン撮影のためのカメラワークを自動化するために，テレビ番組でよく用いられるショットを分類し，それを取得するための基本となる枠組みを提案した．ショットとカメラワークの関係を自動化システムに適用するため,「撮影対象 = 注目対象物 + 注目すべき状態」という考え方を導入し，机上作業シーン撮影に要求される条件を満たすよう，注目対象物と注目すべき状態を分類した．パラメータを調整することで様々なカメラワークを実現することのできる可変枠制御を提案し，仮想シーンと実シーンにおける評価実験，及びプロカメラマンとの比較を通してその有効性を示した．

か」という意味であると補足した．

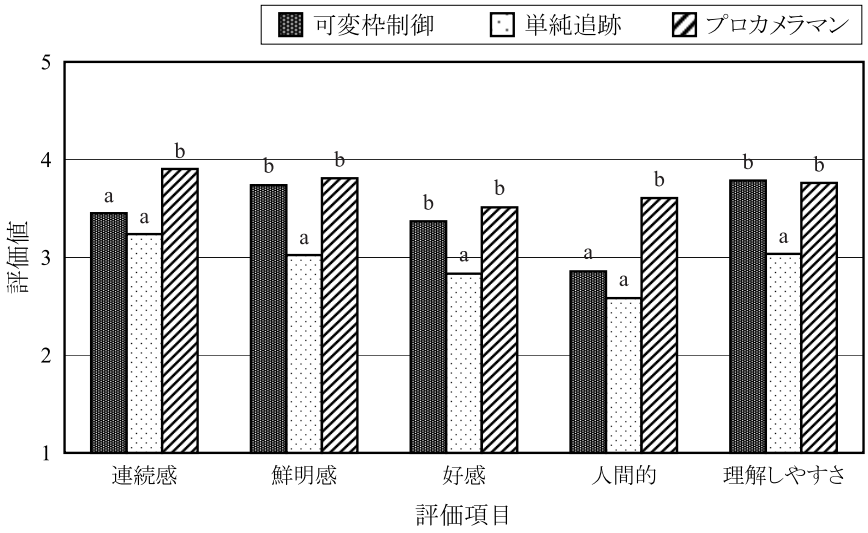


図 3.9: 5 段階の評定尺度法による主観評価の結果

第4章　ディレクタの機能の実現

4.1　はじめに

本章では，ディレクタの機能を自動化システムで実現するための手法を提案する．映像の編集方法には様々あるが，本研究では，一つの連続的なシーンを撮影した複数のショットを切り替える編集を考える．この場合，編集の自動化とは，いつ・どのショットに切り替えるかを決定する要素（**編集トリガ**）を定義し，これを検出若しくは算出することである．

これについて本研究では，編集トリガとして「視聴者の注目を集めるために話者が行う行動（注目喚起行動）」に着目し，その行動が示唆する注目箇所をクロースアップで撮しているショットに切り替えるという編集モデルを提案する．この編集モデルでは，シーン全体を撮したショットと注目箇所をクロースアップで撮したショットの切替えを編集の基本パターンと定義し，これら 2 種類のショットを切り替えるタイミングを話者の注目喚起行動に基づいて決定する．

本手法の特徴は以下の点にある．

- 注目喚起行動の編集トリガとしての有効性と妥当性を網羅的に検証することで，様々な作業映像や教示映像に応用できる知見を提供する．
- 話者の協力を得て自動編集を行うという枠組みについて，話者に大きな負担をかけないことを確認し，映像としても良いものとなることを示す．

4.2　ディレクタの機能の基本要求

本節では，テレビ番組の机上作業シーン映像を調べることで典型的な編集（ショット切替え）のパターンをまとめ，更に各ショットの出現頻度に基づいて編集の最も基本的なパターンを定義する．調べたテレビ番組は，3 章の表 3.1 に挙げたものと同じである．

典型的と考えられる編集パターンを図 4.1 に示す．ショット A とショット C の切替えを中心に，時折ショット B とショット D が挿入される．各ショットの主な役割はそれぞれ次のように推察できる．ショット A で作業の全体的な流れや位置関係を掴ませつつ，ショット C で注目すべき箇所を強調する．ショット B は人物の表情や話している様子を，ショット D は文字にしたほうが分かりやすい情報や作業のポイントなどを伝達することにそれぞれ用いられる．3 章で既に示したように，ショット A とショット C の統計を合わせると，時間・回数の両方で映像全体の 90 %以上を占める．このことから，ショット A とショット C の間での切替えが特に重要であることがわかる．

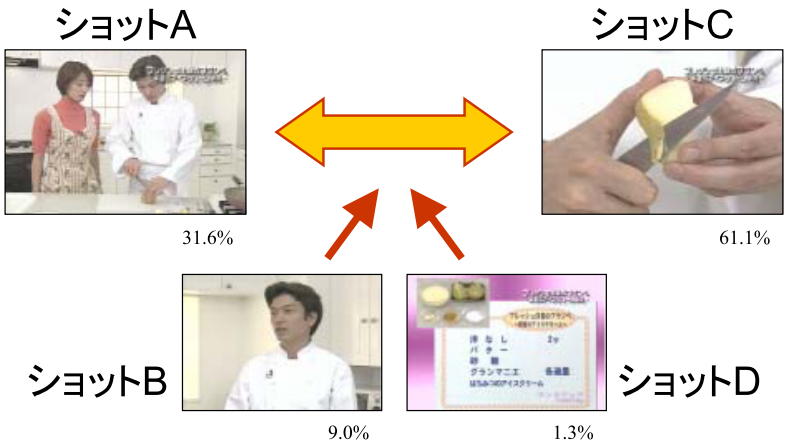


図 4.1: 机上作業シーン映像の典型的な編集パターン

ショット A とショット C の種類を更に細かくみると，まず，ショット A の種類は登場人物の組み合わせの数だけ存在するが，多くの場合，登場人物全員が含まれたショット A で他のショット A を代用することができる[1]．一方，ショット C の種類は，人物の手元，注目物体，注目場所の 3 種類に大きく分けられ，一つの代表的なショット C で他のショット C を代用することはできない．

以上の議論より，一つのショット A と複数のショット C の切替え（ショット A $\Leftrightarrow$ ショット C_i）を机上作業シーン映像における編集の基本パターンと定義する．本研究ではこの基本パターンを扱うための編集手法を提案する．

4.3 注目喚起行動

教示シーンでは，注目すべき箇所の在処が示唆される典型的な状況として，話者が相手（視聴者）の注意を注目箇所に引きつけるための行動が多く現れる．これを**注目喚起行動**と呼ぶことにする．本研究では，話者が注目喚起行動によって注目すべき箇所を特定し，そこをクロースアップで撮したショットに切り換える編集が，視聴者にとっても満足できる映像となると考える．ただし，話者の自然な行動による示唆だけですべての注目箇所を網羅することは難しいため，意識的に注目箇所を行動で示すよう話者に協力してもらうことを本研究では前提とする．

テレビ番組映像を参考にして，編集トリガとして用いる注目喚起行動（動作 4 種類／発話 2 種類）を以下のように選んだ．図 4.2 に例を挙げる．

指示: 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: 注目して欲しい物体を体の前に掲げたり，軽く持ち上げたりする動作.

[1]例えば話者とアシスタントの二人が登場する場合は，話者＋作業領域，アシスタント＋作業領域，話者とアシスタント＋作業領域の主に 3 種類となり，カメラの台数によっては更に別角度からのショットも加わる．しかし多くの場合，正面からの「話者とアシスタント＋作業領域」ショットでこれらを代用することができる．

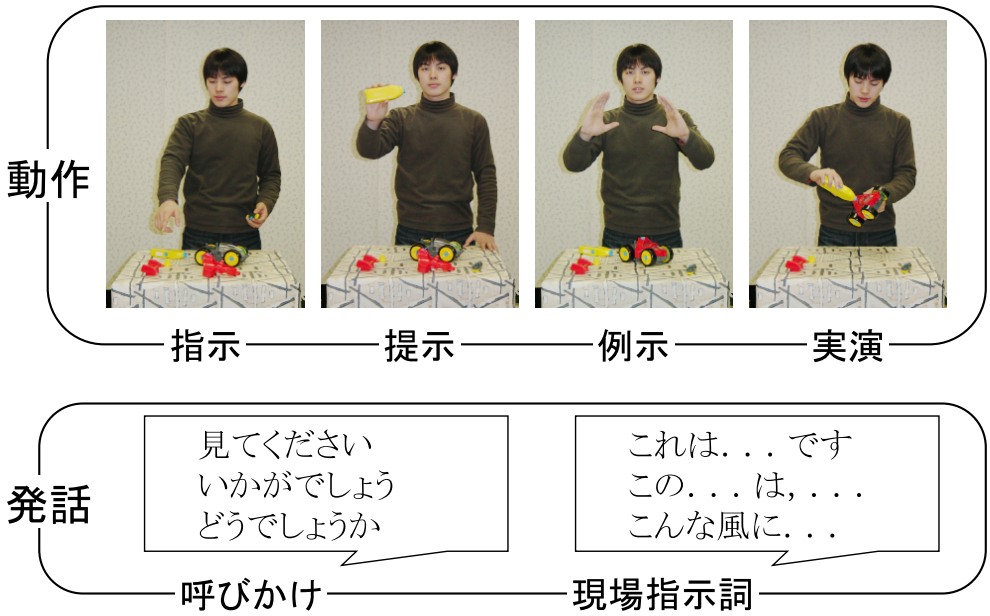


図 4.2: 机上作業にみられる典型的な注目喚起行動

例示: その場にはない物体の形や大きさをジェスチャで表現したり，操作などを身振りだけで表現する動作.

実演: 作業の一部や一連の作業が特に重要であることを強調し，実際に行ってみせる動作.

呼びかけ: 「見てください」,「いかがでしょう」などと呼びかけて，現在の状態に注目して欲しいことを示す発話.

現場指示詞: 「これが〜です」,「このように〜」などと強調して，操作している対象や操作自体に注目して欲しいことを示す発話.

これらの行動は，机上作業シーンに頻繁に現れるもの，若しくは作業映像で重要となる物体や操作に対して注目を集めるものを選んだ．ただし，六つの行動は完全に独立して行われるわけではなく，多くの場合，動作と発話は同時に行われる．特に，“実演” についてはそれ特有の動きはなく，発話を伴って初めて注目喚起行動であると判断できる．そこで本研究では，編集トリガとして用いるという観点からは「話者が強調している箇所をクロースアップで見せる」という点でこれらの行動の役割は共通していると考え，個々の行動を区別せず，まとめて 1 種類の編集トリガとして扱う．これら以外にも注目喚起行動といえる行動はあるが，まずこれら代表的な行動に絞って編集トリガとしての妥当性と有効性を検討する．

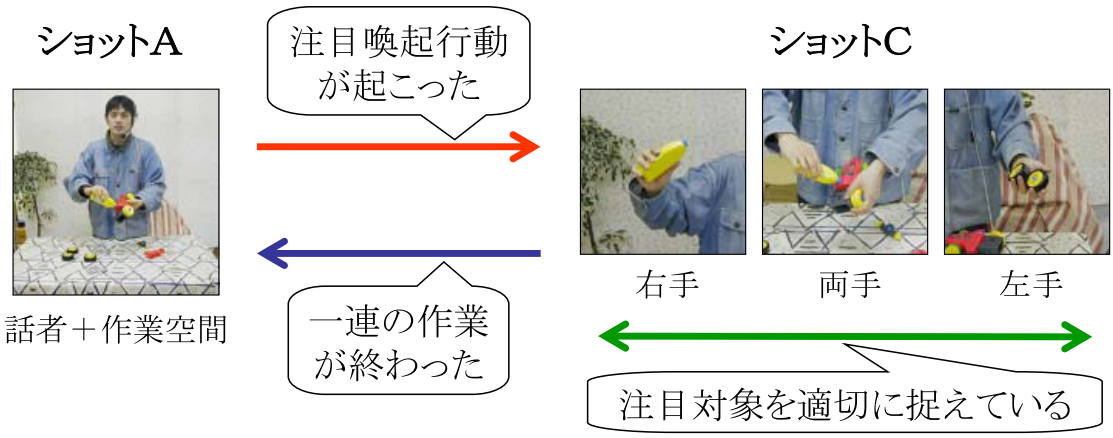


図 4.3: 編集モデル（ショット C が 3 種類の場合）

4.4 注目喚起行動に基づいた編集

4.4.1 編集モデル

4.2 で述べた「ショット A ⇔ ショット C_i」という基本パターンを扱うための編集モデルを提案する．ここでは，一つのショット A と必要数のショット C（C_1〜C_n）は複数台のカメラで同時に撮影されているとし，撮影すると同時にこれらを切り換えることで編集を行う場合を考える．

ここで考えなければいけないことは，ショット A から “いつ”・“どの” ショット C に切り換えるか（A→C_i），そして，ショット C_i から “いつ” ショット A に切り換えるか（C_i→A）の 2 点となる．前章までの議論より，これらをそれぞれ次のように決定する．

A→C_i: 注目喚起行動が起こった時にショット C_i へ切り換える．その際，注目すべき対象を最も適切に捉えているショットを選択する．

C_i→A: 注目喚起行動から続く一連の作業が終わった時にショット A へ切り換える．

ここでは，C_i→A のタイミングを注目喚起行動自体が終わった時ではなく，一連の作業が終わった時としている．これは注目喚起行動が，注目すべき箇所が “いつ”・“どこ” に現れたかを示唆するものであり，それ自体の終了が注目すべき状態の終了を示唆するわけではないからである．

4.4.2 自動化手法

注目喚起行動の起こりとそれに続く一連の作業の終了を自動検出することで，提案した編集モデルを自動化することができる．本節では注目喚起行動の自動検出手法を提案し，注目喚起行動を編集トリガとして用いることの有効性を更に検証するために，注目喚起行動を用いた自動編集の可能性を示す．ここで，検出手法は以下の条件を満たすものが望ましい．

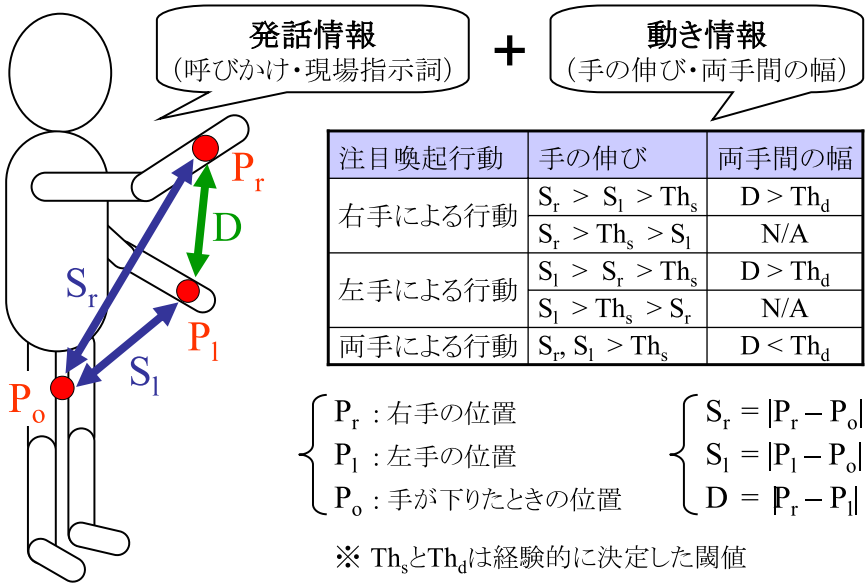


図 4.4: 注目喚起行動の自動検出

- 制約が少なく，話者への負担が少ないほうが良い．つまり，自然な行動を検出することができ，検出のために特別な行動をとる必要がないことが望ましい．
- 検出に失敗した時に話者が再試行できるほうが良い．つまり，話者が直感的に理解できる単純な検出手法であることが望ましい．

これに対して本研究では，簡単な動き情報と発話情報を組み合わせることで注目喚起行動を検出する手法を提案する．前節の A→C_i と C_i→A における検出処理をそれぞれ以下に示す．なお，この手法では注目すべき物体や場所は手の付近にあると仮定し，ショット C_i は右手・左手・両手の中間を撮した 3 種類とする．

A→C_i: 注目喚起行動の検出では，動き情報に「手を前方に伸ばす動き」を，発話情報に注目喚起行動の “呼びかけ” 若しくは “現場指示詞” の単語をキーワードとして用いる．これらの情報を用いて，手を伸ばしている時にキーワードが発声されたら注目喚起行動が起こったとする．どのショット C に切り替えるかは，注目喚起行動がどの手によって行われたかで決定する．概要を図 4.4 に示す．

C_i→A: 一連の作業の終了は，カメラの視野から手が出たこと，もしくは話者が手を下に降ろしたこと（手が伸びていないこと）で判断する．作業の終了時以外にも手がカメラの視野から出ることがあるが，それが作業の終了であるか途中であるかを識別することは今後の課題とする．

動き情報の検出には位置センサから得られるデータを用い，発話情報の検出には市販の音声認識ソフトウェアを用いる．図 4.4 の P_r・P_l・P_o は，それぞれ次のように計算する．

P_r ／ P_l**:** 右手／左手の手首に装着した位置センサのデータをそのまま用いる．

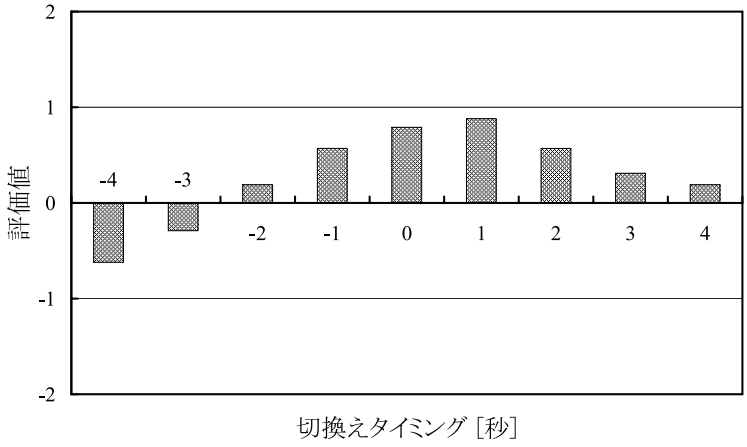


図 4.5: 切替えタイミングの評価（注目喚起行動が起こった時が 0 秒）

P_o: 垂直位置（高さ）は右手を降ろした時の位置を撮影開始時に取得し，水平位置は腰に装着した位置センサのデータをそのまま用いる．

カメラ制御モジュールではカメラの視野のどこに対象が位置しているかを常に計算しているため，対象が視野から外れたか否かはそこから得ることができる．

4.5 注目喚起行動とショット切替えの共起性

注目喚起行動が編集トリガとして妥当であるというためには，これらが机上作業シーンにおいて自然に現れる行動であること，またそれを撮影した映像のショット切替えにそれらの行動が深く関連していることが望ましい．これを確認するため，テレビ番組の机上作業シーン映像を調べ，注目喚起行動とショット C との共起性を調べた．

4.5.1 共起とみなす範囲の検討

共起率を計算する前に，まず，注目喚起行動とショット C の切替えが何秒以内の範囲で同時に起こっていれば共起とみなすかを検討する．ここでは，視聴者が「良い」と感じる切替えタイミングが注目喚起行動の生起時刻に対してどのように分布しているかを調べる．そのため，注目喚起行動が一回だけ起こる作業シーンの映像を被験者に見てもらい，どのタイミングで切り替わるのが良いか答えてもらった．実験の手順を以下に示す．

1. 注目喚起行動 1 回を含むシーンを撮影した 20 秒程度の映像を 9 本用意する．
2. 各映像について，注目喚起行動が起こる前後 4 秒ずつの間を 1 秒間隔でショット A からショット C へ切り替えた映像を用意する．
3. 22 人の被験者に編集パターンが重複していない 9 本の映像を見てもらい，切替えのタイミング

表 4.1: 共起率の計算に使用したテレビ番組映像

テレビ番組	シーン数	ショット数	合計時間
料理番組 A	4	97	28 分
料理番組 B	4	119	28 分
科学実験番組	9	146	24 分
工作番組	10	196	16 分

表 4.2: 表 4.1 に挙げた映像の各ショットの出現頻度

	ショットの種類			
	A	B	C	D
出現回数の割合	41.2 %	4.8 %	53.2 %	0.7 %
出現時間の割合	25.9 %	3.5 %	69.4 %	0.2 %
継続時間の平均	6.7 s	7.8 s	14.2 s	29.3 s
継続時間の標準偏差	4.9 s	6.3 s	14.4 s	9.7 s

を 5 段階（−2:悪い…0:普通…2:良い）で 1 人 2 回ずつ評価してもらった．

各切替えタイミングの評価値の平均を図 4.5 に示す．注目喚起行動の生起時刻に対して，−1 秒〜2 秒の範囲で高い評価が得られている．これより，注目喚起行動とショット C への切替えが −1 秒〜2 秒の範囲で起こっていれば共起とみなすことにする．

また，共起範囲の他に以下の二つの示唆がこの結果より得られる．

- 注目喚起行動の自動検出がこの範囲内で完了すれば，オンライン編集として許容されることを示す．
- 被験者には注目喚起行動に基づいて編集したことは伝えていないのにも関わらず，評価値がこのように注目喚起行動の生起時刻付近を頂点とした山の形を描いていることから，視聴者にとっても注目喚起行動はショット C に切り替える妥当なトリガであることが推測できる．

4.5.2 共起率の計算

調べたテレビ番組映像を表 4.1 に挙げる．まず，注目喚起行動が現れる頻度を調べた．調べた映像（計 6084 秒／ショット数 559 ／ショット C 数 297）では，注目喚起行動が合計 809 回現れており，動作／発話はそれぞれ 285 回／ 524 回であった．このうち動作と発話が同時（1 秒以内）に起こった数は 254 回であり，動作の約 89 %が発話と同時に起こっていた．つまり，同時に起こった動作と発話を 1 回の行動とみなしても，約 11 秒間に 1 回の割合で注目喚起行動が起こっていることになり，一般的な机上作業シーンの中にも十分な数の注目喚起行動が自然に現れているといえる．

次に，共起率を計算した結果を表 4.3 に示す．共起率は，各ショットの開始・終了時刻と注目喚起

表 4.3: ショット C と注目喚起行動の共起率（[ ] 内の数値は，ショット C への切替えと注目喚起行動がそれぞれランダムに起こったと仮定したときの期待値）

ショット C	
ショットが共起した割合	72.1 %
ショット切替えが共起した割合	51.5 % [27.3 %]
注目喚起行動	
ショットと共起した割合	85.7 %
ショット切替えと共起した割合	32.3 % [14.6 %]

行動が起こった時刻を 1 秒単位で調べ，共起している数を数えることで算出した．ここで，ショット C に切り替わっている間に注目喚起行動が起こった場合を “ショットとの共起” とし，更にショット C の開始と注目喚起行動が同時に起こった場合を “ショット切替えとの共起” とする．なお，同時に起こったとは，注目喚起行動に対して −1 秒〜2 秒の間にショット C への切替えがあった場合とする．その決定方法については付録を参照されたい．

結果より，全ショット C の中でその切替えが注目喚起行動と共起した割合は約 51.5 %であるのに対し，ショット C への切替えがランダムに起こったと仮定した場合の期待値が約 27.3 % [2]であることから，ショット C への切替えに注目喚起行動が深く関連しているといえる．

一方，ショット切替えと共起した注目喚起行動の割合は約 32.3 %であり，注目喚起行動がランダムに起こったと仮定した場合の期待値（約 14.6 %[3]）と比べると 2 倍以上共起しているものの，全行動に対する割合としては少ない．しかし，注目喚起行動の約 85.7 %はショット C に切り替わっている間に起こっており，注目喚起行動を編集トリガとした場合に誤編集する確率は 15 %未満となる．

以上の結果より，一般の机上作業シーンにおいて十分な数の注目喚起行動が自然に現れており，またそれを撮影した映像のショット切替えにもそれらの行動が関連していることが確認できる．

4.6 編集結果の主観評価

4.6.1 実験手順

本節では，テレビ番組を模擬したシーンを本システムで撮影し，種々の方法を用いて編集した映像に対する視聴者の反応を確かめることで，話者の注目喚起行動に基づいた編集によって視聴者も満足する結果が得られることを確認する．

まず，模擬するシーンは表 4.1 に挙げた 27 シーンからランダムに 4 シーンを選び，時間の長いものについては更にその中から 1 分半程度の一連の作業シーンを切り出した．用意した編集映像（約

[2]約 11.0 秒間に 1 回の割合で注目喚起行動が起こっており，共起とみなすのは 3 秒間であることから，$3 \div 11.0 \times 100 \approx 27.3(\%)$ となる．

[3]6084 秒の映像中でショット C への切替えは 297 回起こっており，約 20.5 秒間に 1 回の割合でショット C への切替えが起こっていることになる．共起とみなすのは 3 秒間であるから，$3 \div 20.5 \times 100 \approx 14.6(\%)$ となる．

表 4.4: 評価項目

A. 見たい部分が見えましたか？
B. 切替えのタイミングは良かったですか？
C. 飽きませんでしたか？
D. めまぐるしくなかったですか？
E. 作業内容は良く分かりましたか？
F. 雰囲気（様子）は伝わりましたか？

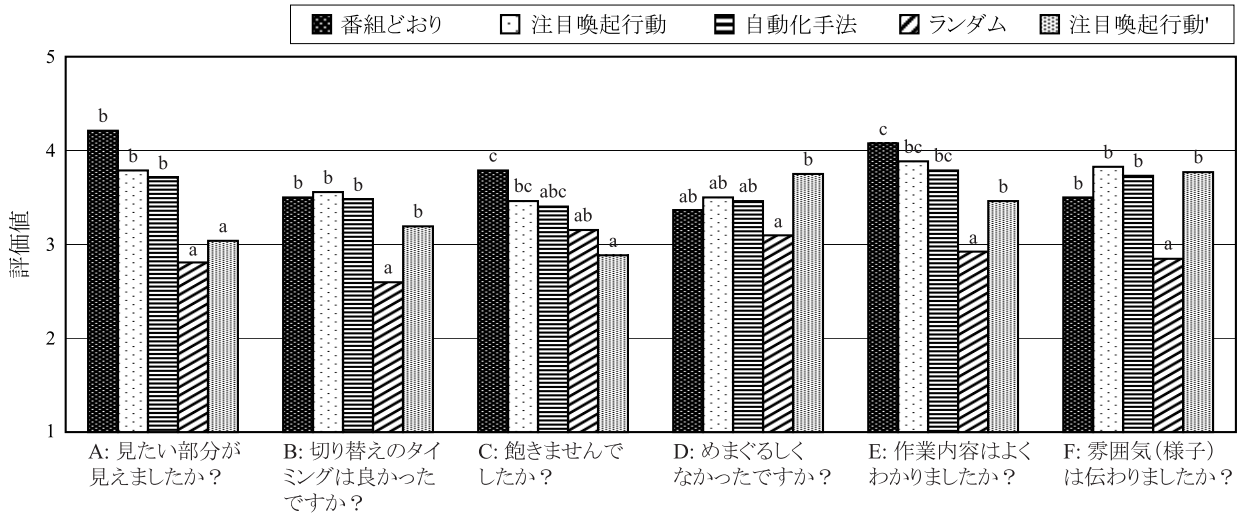


図 4.6: 編集映像の評価実験の結果（1:悪い…3:普通…5:良い）

1 分半 × 4 本）を 13 人の大学院生に見てもらい，表 4.5 に挙げた六つの項目に対して 5 件法（1:悪い…3:普通…5:良い）で答えてもらった．評価は個人毎に Web ページで行ってもらい，映像の順番による影響をできるだけ減らすため，評価する 4 本の編集映像の順番を被験者毎にランダムに変更した．

比較した 5 種類の編集方法を以下に挙げる．

番組どおりの編集: 元のテレビ番組映像に基づいて手作業で編集する．映像制作の専門家によるものであり，唯一の正解ではないものの理想的な編集例であるといえる．

提案手法: 注目喚起行動に基づいて手作業で編集する．提案した編集モデルに従って，理想的に編集した場合の評価を調べる．この編集方法が番組どおりの編集にどこまで近づけるかを調べることが本実験の第一の目的である．

自動化手法: 自動検出の結果を用いて自動編集する．検出ミス（本実験での検出率は約 75 %）や検出の遅延（1〜2 秒）により，提案編集とは若干違った編集結果となっている．提案した編集モデルの有効性を示す一つの指標として，その自動化への可能性を示す．

ランダム編集: ある時点のショットの種類とその継続時間を表 4.2 の統計に基づいてランダムに決定し，これを映像の全時間が埋まるまで繰り返すことで自動編集する．シーンから得られる情報を用いずにショット切替えの統計のみに基づいた自動編集手法として提案手法と比較する．

提案手法′: 注目喚起行動を 2 分の 1 の確率でランダムに選び，それに基づいて手作業で編集する．自動検出が失敗した時の影響を調べる．自動化する場合に，注目喚起行動の検出率がどの程度必要であるかを調べる目安とする．

なお，理想的な編集として元番組とは違った編集パターンも考えられるが，本実験では，本手法が手本とした現在の映像文化（テレビ番組）との比較に焦点をおくことにする．

4.6.2 実験結果

評価実験の結果を図 4.6 のグラフに示す．グラフの縦軸は全被験者の評価値の平均である．また，グラフ上のアルファベットは多重比較法による検定結果を表し，各項目について同じアルファベットを含まない結果の間には有意水準 5%で有意差があることを示す．

各項目の評価について，提案手法と番組どおりの編集の比較・考察を中心に以下にまとめる．

A. 見たい部分が見えましたか？

有意差はないものの，提案手法と番組どおりの編集の間に差がみられる．これは，注目喚起行動だけでは特定できない注目箇所や，話者が強調していなくとも視聴者が見たいと思う部分があることを示している．

B. 切替えのタイミングは良かったですか？

番組どおりの編集と同等の評価が得られた．注目喚起行動と合わせてショットを切り替えることが，視聴者にとっても適切な編集のタイミングとなっていることがわかる．

C. 飽きませんでしたか？

提案手法とランダム編集との差が明確には見られない結果となった．提案手法では，ショットの継続時間を考慮していないことや，ショット B（人物のみのショット）や別視点からのショット A やショット C などを扱えないことが理由であると考えられる．

D. めまぐるしくなかったですか？

全体的に有意差がみられない結果となった．提案手法′の評価が良い傾向があることから，本実験では単に「切替えが少ない」という意味で採点された可能性もある．

E. 作業内容は良く分かりましたか？

番組どおりの編集よりやや評価が悪い傾向があるが，項目 A と比べると接近している．項目 A と本項目の違いは「作業内容に関して」見たい部分が見れたかどうかということであり，注目喚起行動が作業内容に関する注目の要求を効率的に示唆していることがわかる．

F. 雰囲気（様子）は伝わりましたか？

提案手法の評価が番組どおりの編集よりもやや良い傾向となった．クロースアップが多いと全体

的な雰囲気・様子を把握しづらくなることが被験者から指摘されており，番組どおりの編集では提案手法に比べてクロースアップへの切替えが多いことからこのような結果になったと考えられる．

評価をまとめると，提案手法と自動化手法の両方において，全体的には番組どおりの編集に劣っている傾向があるものの，13 人の被験者による実験では統計的に有意差がないという結果になった．これにより，話者の注目喚起行動に基づいた編集（話者の視点からの編集）が視聴者にとっても良い映像となることが確認できる．

ただし，提案手法′の結果をみると，自動編集で満足できる結果を得るためには，75%程度の検出率は必要であることがわかる．また，アンケート項目「C. 飽きませんでしたか？」及び「D. めまぐるしくなかったですか？」では明確な有意差が現れていない．これらの項目をより正確に評価するには，長時間の映像を用いた評価実験が必要であると考えられる．これについては，評価映像の長さと被験者への負荷の関係の考察も含め，今後の課題としたい．

4.7 ユーザインタフェースの評価

本手法の特徴は話者自らが映像編集も行う（話者＝編集者）という点であり，編集結果の有効性だけでなく，そのユーザインタフェースの側面からの検討も不可欠である．本節では，この「話者＝編集者」という枠組みにおける編集手法のユーザインタフェースはどのようなものが良いか，そして，それがプレゼンテーションの形式によってどのように変化するかについて調べた結果を述べる．

関連研究として，講義撮影システムによる講師・受講生への影響についての報告などがある [33]-[34]．これらに対して本研究は，話者＝被撮影者としての影響ではなく，話者＝編集者としての影響を明らかにしようという点に特徴がある．

4.7.1 目的

講義やプレゼンテーション，机上作業などを対象とした自動撮影システムのオンライン編集には，少なくとも次の三つの形態が考えられる．

(1) 編集ルールに基づいてシステムが自動的にショットを選択

(2) システムの補助により視聴者側が見たいショットを選択

(3) システムの補助により話者側が見せたいショットを選択

話者が撮影や編集を気にすることなく講義や解説ができるという点から，多くの自動撮影システムでは (1) か (2) の形態を用いた編集手法が使われている．しかし，説明内容のポイントを最もよく把握しているのは話者であることから，話者による編集をシステムが上手くサポートできれば，(3)

の形態を用いた編集手法も非常に有効である．

本研究では，このような視点から，話者の注目喚起行動に基づいた編集手法を提案し，その編集結果の有効性を明らかにしてきた．しかし，(1) や (2) に対して (3) の形態を用いた編集手法では，編集結果の有効性だけでなく，どのような編集トリガを用いて話者が編集の意図をシステムに伝えるのが良いかというユーザインタフェースの側面からの検討が不可欠である．

そこで，典型的な自動編集手法をシステムに実装して被験者に実際に使ってもらい，各編集手法についてのアンケート評価を行った．本実験では，3 種類のプレゼンテーション形式の下で，4 種類の編集手法を比較する．このような設定により，次の二つの観点から編集手法のユーザインタフェースを検討する．

- 話者が編集の意図をシステムに伝える編集トリガとして，どのようなものが良いか
- それがプレゼンテーション形式によってどのように変化するか

他の比較条件として，作業の内容や時間の違い，立った状態 ／座った状態の組み合わせなども考えられるが，被験者の負荷を考えるとすべての組み合わせを一度に行うことは難しい．そこでまずは，プレゼンテーションの内容を 1 種類（車の模型の組立て作業：4 分程度：立ち作業）として，プレゼンテーション形式と編集手法の評価の関係を調べることに焦点を絞る．

4.7.2 比較する編集手法

編集トリガとして，注目喚起行動に加え，編集用の特別な行動を用いることがまず考えられる．そこで編集手法の比較では，話者がシステムに編集の意図を伝えるのに，1) 普段から自然に現れる行動を使うのが良いのか，編集用の特別な行動を使うのが良いのか，そして，2) 特別な行動としては発話を用いるのが良いのか動作を用いるのが良いのか，という点に注目する．本実験で比較する編集手法の概要を以下にまとめる．

(A) 注目喚起行動法： 普段から自然に現れる行動による編集手法．物を掲げる動作や指示詞の発話など，話者が視聴者の注目を集めるためにとる行動を用いる．実装方法については既に述べたとおりである．

(B) キーワード法： 編集用の特別な発話による編集手法．右手・左手・両手・全体というカメラを直接指定するキーワードを用いる．本システムではカメラと撮影対象が一対一で対応しているため，それぞれのカメラが撮影しているショットの内容（右手・左手・両手・全体）をキーワードとしてカメラを発話で指定する．ここで「全体」とはマスターショット（ショット A）を指す．

(C) 足スイッチ法： 編集用の特別な動作による編集手法．作業で両手が自由に使えるよう，足下に置いたスイッチの ON/OFF とその際の手の位置を用いる．典型的な方法としてスイッチを使うことがまず考えられるが，机上作業シーンでは両手を使ったままショットを切り替えたい場面が頻繁に現れるため，手ではなく足でスイッチを押すことにする．ただし，四つのスイッチを足で区別して押すことは難しいため，ショットの選択は注目喚起行動法と同じ方法を用いて決定する．

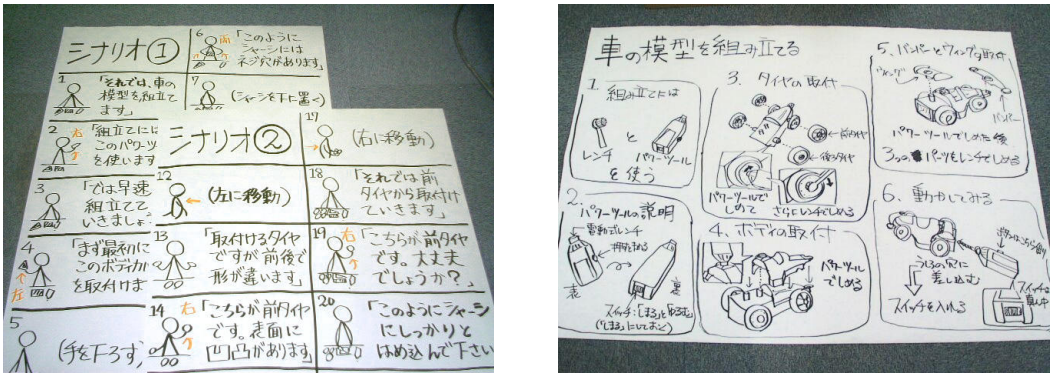


図 4.7: 詳細シナリオ（左）と概略シナリオ（右）

スイッチには，市販されている直径 8cm 程度の足で押すタイプのスイッチ（以後，足スイッチと呼ぶ）を使用した．詳細シナリオ形式と概略シナリオ形式の作業では話者の移動があるため，作業を主に行う場所の足下 2 箇所にスイッチを置いた．

(D) 手動切替え法： 人間の編集者による手動編集．参考のために上述の自動編集手法とともに比較する．編集者は筆者の一人が担当し，スタジオの様子をモニタで見ながら映像切替器のボタンを押すことで，切替え・切戻しタイミングとショット選択を決定する．なお，この筆者は机上作業シーン撮影の研究に 4 年以上携わっており，手動によるオンライン編集には十分慣れている．

4.7.3 比較するプレゼンテーション形式

机上作業シーンを撮影する主な用途として，1) シナリオに従って映像マニュアルを作る場合，2) シナリオを参考にして遠隔地に向けて一方的に説明する場合，3) シナリオを用いずに遠隔地と双方向でやり取りする場合を想定する．これらプレゼンテーション形式の違いによって，話者にかかる作業説明の負荷が変わる．話者にとって本来のタスクである説明の負荷の変化に対して，各編集手法の評価がどう変わるかに注目する．本実験で比較するプレゼンテーション形式を以下にまとめる．

詳細シナリオ形式： 作業内容・台詞・編集箇所（8 箇所）を漫画のように詳細に図示したシナリオ（図 4.7 の左）に従って机上作業の説明を行う．アーカイブを目的とした映像マニュアル制作などを想定する．

概略シナリオ形式： プラモデルの組立説明書のように作業手順の概略だけを図示したシナリオ（図 4.7 の右）を参考にして机上作業の説明を行う．一方的に映像を送る形式の遠隔講義などを想定する．

インタラクション形式： シナリオを用いず，遠隔地で送信映像を視聴している聞き手との質疑応答の形で机上作業の説明を行う．双方向のやり取りを含む遠隔講義などを想定する．

表 4.5: 話者役の被験者へのアンケート項目

a) 思いどおりに切り替わりましたか？
b) 作業は妨げられませんでしたか？
c) 長時間（30 分以上）使い続けられますか？
d) 使い慣れれば便利だと思いますか？

表 4.6: 聞き手役の被験者へのアンケート項目

e) 話者の振る舞い（言動）は自然でしたか？
f) 切替えのタイミングは良かったですか？

4.7.4 実験手順

本実験では，本システムを使ったことのない 6 人の大学院生を話者役の被験者として集め，車の模型を組み立てる作業（4 分程度）を説明してもらい，表 4.5 のアンケートに 5 件法（1:いいえ…3:どちらともいえない…5:はい）で回答してもらった．プレゼンテーション形式は，詳細シナリオ形式 → 概略シナリオ形式 → インタラクション形式の順番で，それぞれ 2 週間〜1ヶ月程度の間隔をあけて実験を行った．

詳細シナリオ形式と概略シナリオ形式では，話者役の被験者に，前方に掲示したシナリオを見ながらカメラに向かって作業を説明してもらった．インタラクション形式では，別の 6 人の大学院生を聞き手役の被験者として集め，あらかじめ用意した質問表（各編集手法につき 7 問ずつ）を読み上げてもらった．それに対して話者役の被験者は，その回答の中で映像を一回以上切り替えることとした．また参考のため，聞き手役の被験者に表 4.6 に示す編集結果についてのアンケートに回答してもらった．なお，どの手法が現在使われているかは聞き手役の被験者には知らせていない．

以上の実験手順をまとめたものを表 4.7 に，実験の様子を図 4.8 に示す．

4.7.5 実験結果

まず，編集の成功率を表 4.8 に示す．キーワード法の成功率がやや低いが，全体的に 80〜90%程度の成功率となっており，ユーザインタフェースの是非を議論できる程度には各被験者とも編集手法を使いこなせているといえる．注目喚起行動法とキーワード法における失敗の主な原因は，音声認識のミスであった．足スイッチ法における失敗は，足スイッチの押し遅れと，手の位置関係を用いたショット選択に失敗したことに因る．同じショット選択の方法を用いている注目喚起行動法ではミスが比較的少なかったが，これは足スイッチ法が足と手を同時に使うためにうまくショットを選択できなかったことが原因と考えられる．手動切替え手法でも 100%とはなっていないことから，手元など素早く不規則に動く対象のクロースアップショットを含んだオンライン編集を失敗なく行うことが，編集者に相当の技術と集中力を要することがわかる．

表 4.5 のアンケート評価の結果を図 4.9 のグラフにまとめる．被験者数が少ないことから，統計

表 4.7: 実験手順

1) 手動切替え法を通してオンライン編集を体験してもらう
2) 3 種類の自動編集手法について以下を繰り返す（編集手法の順番は被験者毎に変更）
   1. ［詳細シナリオ形式と概略シナリオ形式のみ：］編集手法について説明し，練習を兼ねて一度リハーサルを行う
   2. 本番を 2 回行う（インタラクション形式では 1 回）
   3. 話者役の被験者に表 4.5 のアンケートに回答してもらう
   4. ［インタラクション形式のみ：］聞き手役の被験者に表 4.6 のアンケートに回答してもらう

図 4.8: 6 人の被験者による机上作業の例

的検定による明確な有意差はみられなかったが[4]，各手法の傾向は十分に現れている．この結果についての考察は次節で行う．

表 4.8: 編集の成功率

詳細シナリオ形式				
	(A)	(B)	(C)	(D)
切替え	94.8 %	78.1 %	86.4 %	100.0 %
切替え&切戻し	92.7 %	75.0 %	85.4 %	93.8 %
概略シナリオ形式				
	(A)	(B)	(C)	(D)
切替え	96.6 %	88.2 %	84.5 %	95.9 %
切替え&切戻し	94.3 %	83.9 %	81.6 %	89.8 %
インタラクション形式				
	(A)	(B)	(C)	(D)
切替え	83.7 %	82.4 %	81.6 %	96.0 %
切替え&切戻し	81.6 %	78.4 %	73.5 %	90.2 %

(A) 注目喚起行動法　(B) キーワード法
(C) 足スイッチ法　(D) 手動切替え法

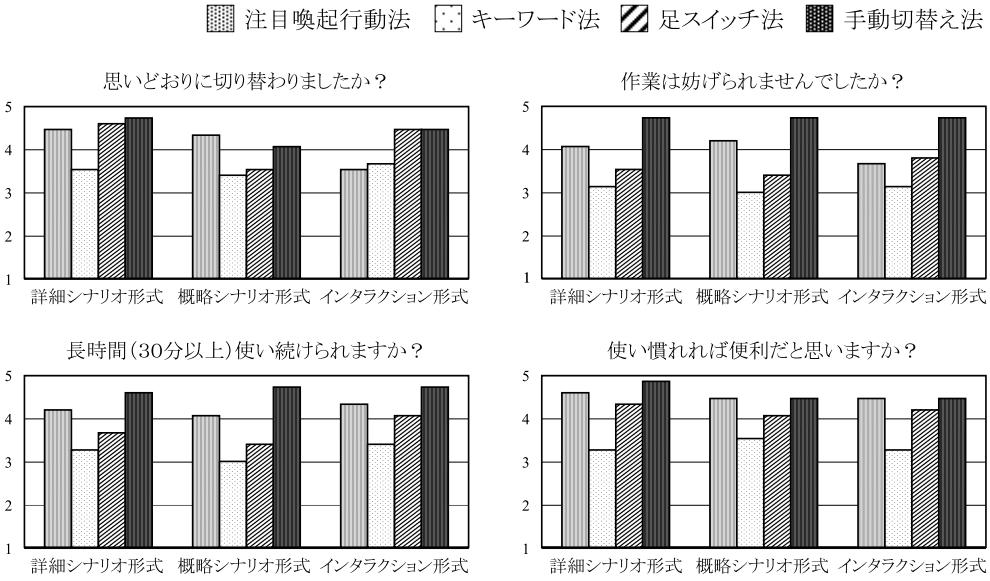


図 4.9: 表 4.5 のアンケート評価の結果（グラフの縦軸の値が大きいほど評価が良いことを表す）

4.7.6 考察

ユーザインタフェースの比較

図 4.9 に示したアンケート評価の結果，自動編集手法の中では，注目喚起行動法の結果が全体的に良い傾向となり，足スイッチ法が次点の評価を得た．対して，キーワード法の結果は全体的に低

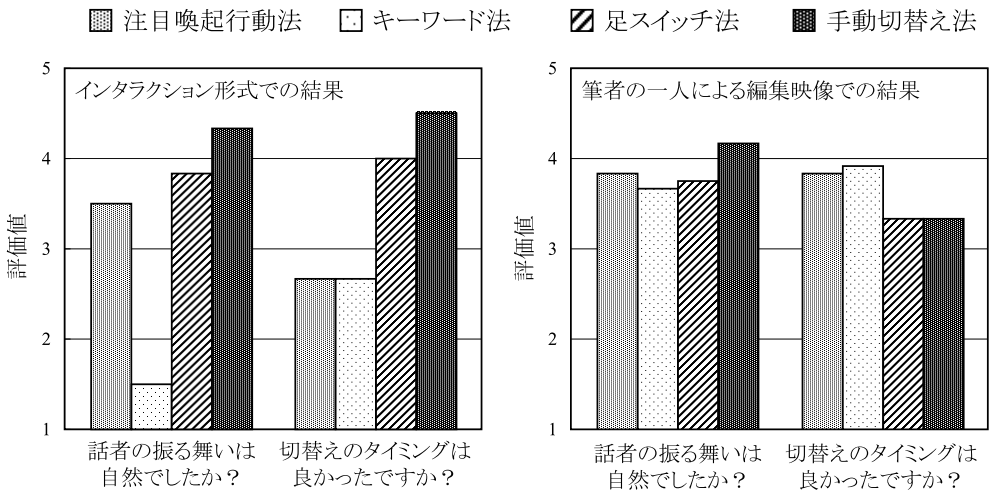


図 4.10: 表 4.6 のアンケート評価の結果

く，有意に評価が低いものが多かった．一方，プレゼンテーション形式間の比較では,「思いどおりに切り替わりましたか？」という項目以外では，傾向の違いはほとんどみられなかった．

以上の結果より，話者が編集の意図をシステムに伝えるユーザインタフェースとしても，注目喚起行動が有効であることがわかる．これは，普段から自然に現れる行動を編集トリガとして利用することで，作業が妨げられにくく，また長時間使えそうだと判断されたと考えられる．しかし，シナリオがない場合は，自然な行動は反って使いづらく，編集用に特別な行動をとるほうが使いやすいという意見もあった．このことは，シナリオを用いないインタラクション形式における「思いどおりに切り替わりましたか？」という項目の結果で，注目喚起行動法が最も悪い結果となっていることからも伺える．よって，様々なプレゼンテーション形式に対応するためには，注目喚起行動法と足スイッチ法を組み合わせ，場面によって上手く使い分けられる手法が望ましいと考えられる．

編集結果の評価

インタラクション形式の実験の際に行った聞き手役の被験者による編集結果のアンケート評価の結果を図 4.10 の左のグラフに示す．結果より，注目喚起行動法が切替えタイミングの項目で，キーワード法がすべての項目で，平均 3 を下回った．これについては，注目喚起行動法やキーワード法を十分に使いこなすまでには話者役の被験者が慣れておらず，不自然さが目立ったことが原因と考えられる．

そこで，ユーザインタフェース実験の被験者とは別の 12 人の大学生を集め，全編集手法に十分に慣れている筆者の一人が作った編集映像を比較してもらった．結果を図 4.10 の右のグラフに示す．

[4]有意水準 5%で有意差があると判断されたものは,「詳細シナリオ形式のアンケート項目 (d) で注目喚起行動法がキーワード法よりも評価が高い」「概略シナリオ形式の項目 (b) で注目喚起行動法がキーワード法より評価が高い」「インタラクション形式の項目 (a) で足スイッチ法が注目喚起行動法より，項目 (d) で注目喚起行動法がキーワード法より，それぞれ評価が高い」という結果となった．

インタラクション形式での実験結果に比べ，編集手法間での評価の差が小さくなったことがわかる．分散分析の結果でも有意な差（有意水準 5%）は認められなかった．

以上の結果より，いずれの自動編集手法の編集結果も，手動切替え法と同等の評価が得られることを確認した．

4.8 まとめ

本論文では，シーン全体を撮したショット A と注目箇所をクロースアップで撮したショット C の切替えを机上作業シーン映像の編集の基本パターンと定義し，いつ・どのショット C に切り替え，いつショット A に切り替えるかを注目喚起行動に基づいて決定する編集モデルを提案した．また，注目喚起行動の編集トリガとしての妥当性と有効性について，いくつかの実験を通して網羅的に検証した．実験で明らかにしたことを以下にまとめる．

- テレビ番組映像における注目喚起行動とショット C の共起性を調べることで，十分な数の注目喚起行動が机上作業シーンに自然に現れ，またそれを撮影した映像のショット切替えの約 50%にそれらの行動が関係していることを確認した．
- いくつかの代表的な方法で編集した映像を視聴者に見比べてもらうことで，話者の注目喚起行動に基づいた編集が視聴者にとっても良い編集となることを確認した．注目喚起行動とショット切替えの共起率は 50%程度であったが，注目喚起行動だけを用いた編集でも視聴者に十分良い印象を与えていることがわかる．
- 複数のカメラを用いた自動撮影システムにおいて，話者が何かを説明しながら同時に編集も行うという枠組みで用いる編集手法について，そのユーザインタフェースの側面から検討した．実験では，3 種類のプレゼンテーション形式の下で 4 種類の編集手法を実際に使ってもらい，アンケート評価を行った．その結果，注目喚起行動法の評価が全体的に良い傾向にあること，シナリオのない状況では足スイッチ法の評価が良い傾向にあることから，場面に応じてこれらの手法を使い分けることのできるインタフェースが望ましいことがわかった．

今後の課題として，4.5 節で注目喚起行動と共起しなかった残りの約 48.5 %のショット C を扱うため，注目喚起行動以外の編集トリガの検討が挙げられる．また，より多くのショットを扱う枠組みについても検討する必要がある．これについては，まず人物を映したショットであるショット B を編集モデルに含め，面白味のある映像作りを目指していく．また，映像だけでは伝えづらい情報を補完するという役割を果たすショット D についても，より高度な映像コンテンツの制作に向けて将来的に考慮していく必要がある．

更に，ユーザインタフェースについての実験は,「話者＝編集者」という枠組みにおけるユーザインタフェース評価の試みの第一歩であり，残された課題は多い．被験者を増やして統計的に有意な結果を得ることをはじめ，プレゼンテーションの内容や時間の違いによる評価結果の変化も調べる必要がある．また，編集結果の評価実験でみられたように，編集手法の慣れによって評価結果が変

化してくる可能性がある．今後も評価実験を継続し，この点に注目していきたい．一方，長時間の使用における評価を行うためには，現在のシステムは位置センサの装着や音声認識の遅れなどの問題があるため，これらの改善も行っていく予定である．

第5章　映像インデキシングとその利用

5.1　はじめに

本章では，机上作業シーン映像の自動取得と同時に，シーンに関するメタデータを自動取得し，索引（インデックス）として映像と共に記録する（映像インデキシング）手法について説明する．また，そのようにして得られたインデックス付き映像の利用例をいくつか挙げる．

5.2　物体情報のインデキシング

5.2.1　机上作業シーン映像のメタデータ

作業などの映像マニュアルを考えた場合，部品や操作，またそれに対する説明が最も重要な情報となる．このような映像マニュアルを利用する際に，重要な部品や操作が説明されたり，使われたりしている部分に簡単にアクセスすることができれば便利だろう．このような高度な映像コンテンツを実現するためには，作業に登場する物体について，以下のようなメタデータを自動的に取得し，インデックスとして映像に付加しておく必要がある．

- 物体の位置とテクスチャ
- 物体を使った作業の内容と区間
- 物体への注目度

物体を検索キーとして使用するためには，個々の物体の各時点での位置，それが使って行われた作業の区間，物体が注目されていた区間などの情報が映像をインデックスとして記録しておく必要がある．ここで物体のテクスチャが得られていれば，例えば，映像閲覧の際に物体像のアイコンをクリックすることで，物体が説明されている部分や操作に使われている部分にアクセスすることができる．また作業の内容がインデックスとして記録されていれば，作業内容をキーとして映像を検索することができる．

5.2.2　条件設定

本研究では以下のような机上作業シーンを対象とする．

1. 図 5.1 にあるように，話者が物体を前に掲げたり指さしたりする．

2. 物体を参照しながら，その名前や使い方などについて説明する．
3. 紹介した物体（部品や道具）を使って，組み立てや分解などの作業を行う．

このような場面では，物体は回転や変形などによって常にその外観を変化させる．また，シーン内には説明を行う人物以外にも他の人物が近傍にいることが多い．そのため，以下のような環境条件の下で物体追跡に取り組む必要がある．

- 物体の大きさ・色・形状などに関する予備知識はない．種々の物体が形を変えながら出現するため，すべての物体に関する知識をあらかじめ与えておくことは難しい．
- 作業中，背後に複数の人物が登場したり，物体の位置や姿勢が変化するため，背景は常に変化する．

ただし，机上作業シーンであることから次の前提条件が仮定できる．

- 重要な物体（注釈付けすべき物体）のほとんどは人の手によって動かされたり操作されたりする．
- 作業は一カ所に留まって行われることが多いため，注目すべき物体が存在する可能性の高い空間はある程度特定できる[1]．

これら二つの前提条件を加えても，先に述べた二つの環境条件下における物体追跡は簡単ではない．これに対して本研究では，可視光カメラ・赤外線カメラ・ステレオカメラの 3 種類の画像センサを相互補完的に用いることで，このように厳しい条件の下でもロバスト物体追跡を実現する手法を提案する．

5.2.3 システム構成

図 5.1 に概要を示す．自動撮影・編集システムと物体追跡システムをネットワークで繋ぎ，自動撮影・編集システムで取得した映像に対して物体に関する情報を自動インデキシングする．3 種類の画像センサ（可視光カメラ・赤外線カメラ・ステレオカメラ）は作業台の正面に隣接して設置する．

5.3 物体情報の取得

本研究では，物体像を物体のモデルと直接照合するのではなく，ある特定の空間中で手と共に移動する領域を検出し，そこから手の領域を分離することで人が手に持っている物体（把持物体）を検出する．そのために，まず 3 種類の画像センサから以下の要素領域を抽出する．

肌色領域： 肌色の領域．可視光カメラから得られる画像より抽出する．

動領域： 動いている領域．可視光カメラから得られる画像より抽出する．

肌温領域： 人間の肌の表面温度以上の温度を持つ領域．赤外線カメラから得られる画像より抽出する．

[1]例えば作業台上で行われる作業では，重要な物体のほとんどは作業台の上にあると仮定できる．

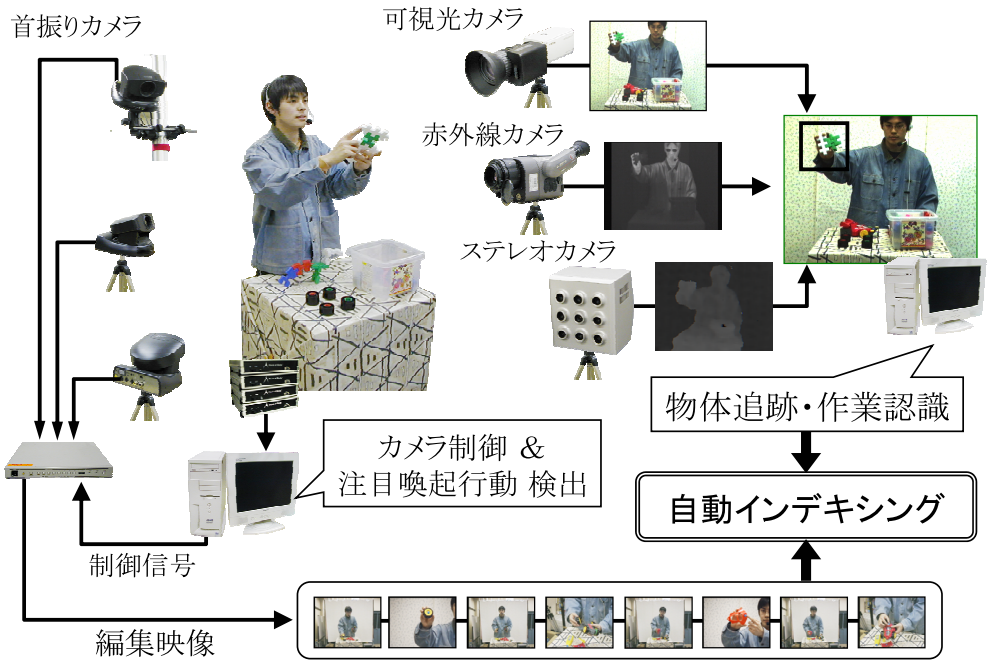


図 5.1: 自動撮影・編集システムと物体追跡システム

特定距離領域： 作業が行われると想定される空間にある物体の領域．ステレオカメラから得られる距離画像より抽出する．

要素領域の論理積を以下のようにとることで，手領域と把持物体領域を検出することができる．

$$\begin{aligned} \text{手領域} &= \text{特定距離領域} \wedge \text{動領域} \\ &\quad \wedge \text{肌温領域} \wedge \text{肌色領域} \end{aligned} \tag{5.1}$$

$$\begin{aligned} \text{把持物体領域} &= \text{特定距離領域} \wedge \text{動領域} \\ &\quad \wedge \neg \text{手領域} \end{aligned} \tag{5.2}$$

ただし，一部の “∧” は厳密な論理積を表さない．これについては以下の「手と把持物体の検出」の部分で説明する．処理の概要を図 5.2 に示す．この手法では三つの画像センサの視点位置が一致している必要があるが，可視光カメラと赤外線カメラおよびステレオカメラのすべての視点を一致させることは難しい．そこで本研究では，センサから取得された画像を補正することで近似的に視点を一致させる．本論文の実験で使用した補正方法については 5.5.1 節を参照されたい．

以下，各要素領域の抽出方法をまず説明し，その後に手と把持物体の検出方法について述べる．

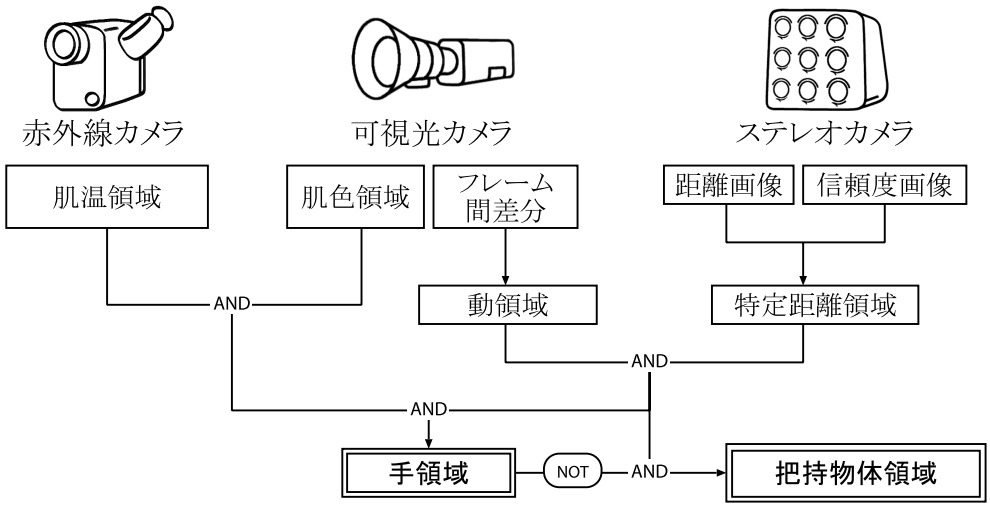


図 5.2: 各要素領域の抽出と把持物体検出の流れ

5.3.1 肌色領域の抽出

肌色は RGB 色空間において rg 平面[2]にまとまりのある分布を示すことが知られている．そこで rg 色度平面上に肌色モデルを作り，肌色領域を抽出するのに用いる [35]．ただし，近接物体からの反射光や照明の当たり具合いにより，実際に観測される色はかなり変化する．そこで本研究では，厳密なモデルを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ようにする．

我々の実験環境において腕と手から抽出した多数の肌領域のサンプルを基に肌色の分布を求めた．その平均値と共分散行列 Σ の例を以下に示す．

$$mean(\bar{r}, \bar{g}) = (0.437773, 0.334845)$$
$$\Sigma = \begin{pmatrix} 0.003915 & -0.000230 \\ -0.000230 & 0.000935 \end{pmatrix}$$

これを基に，肌色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．

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

図 5.3 の左側のグラフは，可視光カメラ画像から手動で抽出した肌領域と背景領域のマハラノビス距離 $D^2(r, g)$ の各値における画素分布を示している．他の要素領域との論理積をとることから，候補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．ただし，我々の実験環境で用いている肌色に近い色合いの壁紙の影響を避けるため，そのピークが含まれないよう閾値 $Th_{\text{skin-c}}$ を決定した[3]．

[2]正規化色空間．$r \equiv \frac{R}{R+G+B}$, $g \equiv \frac{G}{R+G+B}$ で定義される．
[3]ベージュなど明るい暖色系の壁紙は，我々の実験環境に限らず一般的によくみられる．

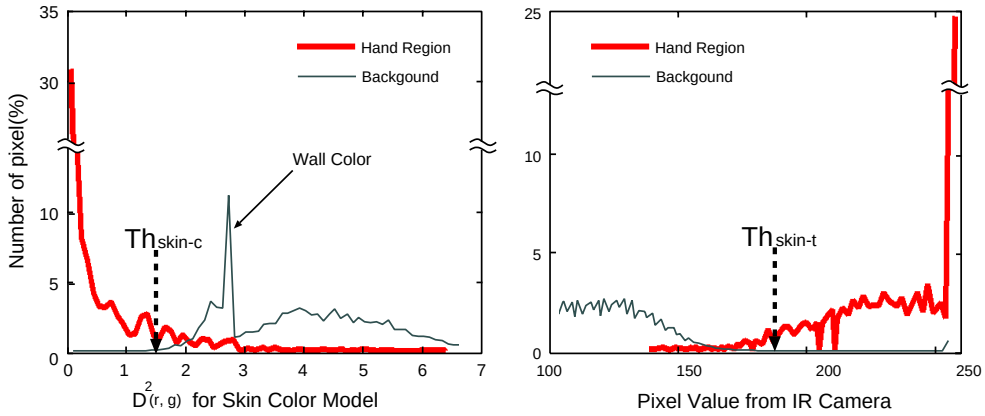


図 5.3: 典型的な机上作業シーンにおける，肌領域と背景領域のマハラノビス距離による画素の分布（左）と肌領域と背景領域の温度による画素の分布（右）

5.3.2 動領域の抽出

動領域は可視光カメラから得られた画像のフレーム間差分を用いて抽出する．差分は 4 フレーム間隔で行い，閾値は検出される可能性のある物体のうち最小のものを基に計算した．

5.3.3 肌温領域の抽出

本研究で使用する赤外線カメラは検出波長が 8〜14 μm であり，人体が出す赤外線の 7〜10 μm に対して高い感度を持つ．そこで赤外線カメラ画像を閾値処理することによって，人体の肌表面温度に相当する画素のみを抽出し，得られた部分を手や顔の候補領域とする．

図 5.3 の右側のグラフは，赤外線カメラ画像から手動で抽出した肌領域と背景領域の画素分布を示す．他の要素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ため，肌領域のほとんどが抽出されるよう閾値 $Th_{\text{skin-t}}$ を決定した．

5.3.4 特定距離領域の抽出

ステレオカメラから得られる奥行き画像を用いて，作業が行われていると想定される空間（以下，特定距離空間と呼ぶ）を立方体の空間として切り出す．5.5 章の実験と 5.6 章の応用例では，特定距離空間の幅・奥行きをそれぞれ作業台の幅と奥行き[4]に，特定距離空間の高さを作業台の天板上から人物の頭までの高さとした．ステレオカメラは作業空間に正対して配置されているため，特定距離領域（特定距離空間内にある物体の領域）は単純な閾値処理によって抽出できる．

[4]作業台の手前の端に意味なく手を置くことが多いため，厳密には，作業机の奥行き幅よりもステレオカメラからみてやや浅めに特定距離空間の奥行きを設定している．

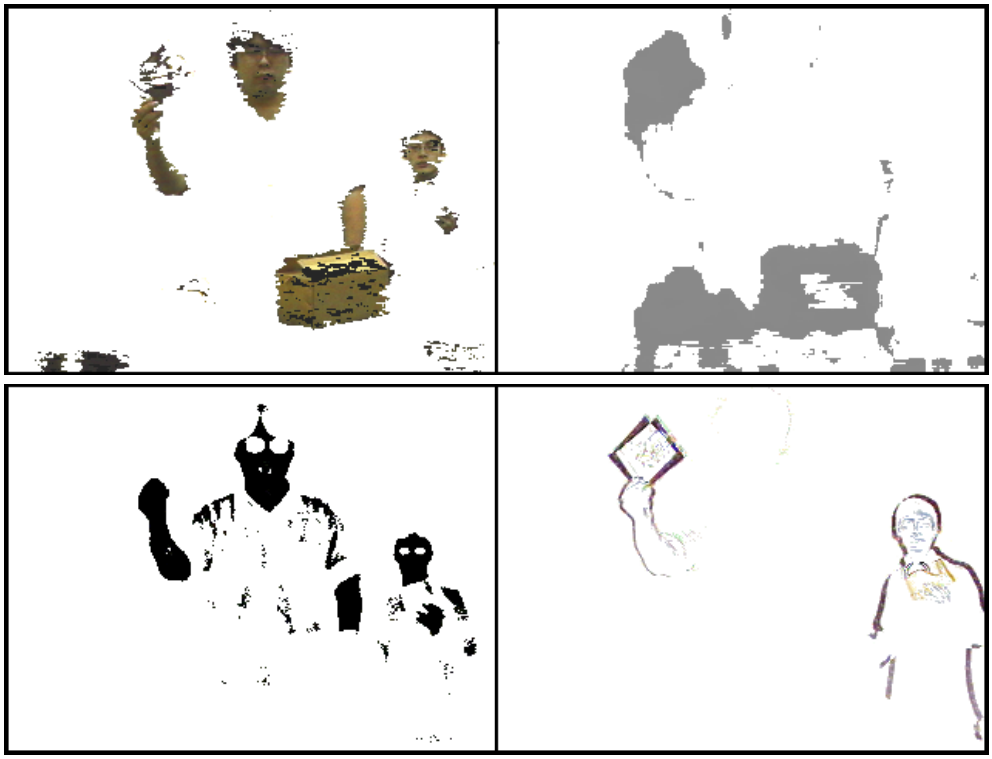


図 5.4: 各画像センサから得られる領域（左上：肌色領域，左下：肌温領域，右上：特定距離領域，右下：動領域）

5.3.5 手と把持物体の検出

手領域の検出については，まず前述の式 (5.1) に示した論理積をとることによって候補領域を抽出する．次に候補領域に対して収縮膨張処理を行い，細かいノイズを削減する．最後に残った候補領域の面積をラベリングによって計算し，面積が閾値 Th_{hand} よりも大きい領域を二つまで手領域として検出する．ここで閾値 Th_{hand} は，我々の実験環境における手の大きさを基に計算した．

把持物体領域についても，まず式 (5.2) に示した論理積をとることによって候補領域を抽出する．ただし，式 (5.2) の最初の “∧” は単なる論理積処理ではなく，まず特定距離領域をラベリングによって各領域に分離し，各領域の中で動領域と判定された画素の割合（画素数の比）を計算する．この割合が閾値よりも大きく，且つ手領域ではない小領域を把持物体の候補領域とする．抽出された領域候補は収縮膨張処理によりノイズを削減し，残った候補領域の面積が閾値 Th_{obj} よりも大きなものを二つまで把持物体領域として検出する．ここで閾値 Th_{obj} は，我々の実験環境において検出される可能性のある最小の物体を基に計算した．

手や把持物体の位置は検出された領域の重心を計算することによって求める．この重心計算を毎フレーム繰り返すことで手と把持物体を追跡する．求めた位置はカルマンフィルタを用いて平滑化し，滑らかな追跡撮影を行う．

表 5.1: 作業検出の概要

作業内容	数の変化	位置関係
提示	1	-
分離	1 → 2	離れていく
組付	2 → 1	一緒になる

各画像センサから抽出された領域の例を図 5.4 に示す．この図からわかるように，肌色の領域を検出するだけでは，肌色の箱や背後の人物など手以外の領域も検出されてしまう．温度情報と奥行き情報を組み合わせることで，誤検出された領域を候補から外し，手と把持物体のみを正確に検出できていることがわかる．

5.3.6 作業の検出

物体追跡システムから得られる「把持物体の数の変化」と「把持物体の位置関係」の組み合わせより，机上作業シーンで重要な「提示・分離・組付」の三つの作業を検出・識別する．作業検出の概要を表 5.1 に示す．

ただし，これだけのルールでは単なる物体の移動や接触を作業として誤検出してしまうため，話者が注目喚起行動を行なってから一連の作業を終えるまでの間に表 5.1 の状態が起こった場合にのみ作業検出を行なう．注目喚起行動とそれに続く一連の作業の終了は自動撮影・編集システムで検出され，その検出時刻はネットワークを通して把持物体追跡システムに送信される．

5.4 映像インデキシング処理

物体追跡システムと机上作業シーン映像の自動撮影・編集システムから得られる情報を組み合わせることにより，映像インデキシングを行なう．各システムから得られる情報を以下に示す．

物体追跡システム： 物体の位置，物体のテクスチャ，物体が把持されている期間，把持物体間の距離
自動撮影・編集システム： 注目喚起行動が行われている期間，発話内容，多視点映像

映像インデキシングは，ある物体に関連する注目喚起行動の開始・終了，把持の開始・終了の時刻，作業内容を記録することで行う．取得映像におけるこれらの検出時刻を物体の ID・位置・テクスチャとセットにしてデータベースに記録していく．位置やテクスチャの照合によって，このデータベースから映像を検索することなどができる．図 5.5 に映像インデキシングの流れを示す．

実際に映像インデキシングを行った例を図 5.6 に示す．内容は，一人の人物がもう一人の人物と対話しつつ，机の上にある料理のサンプルを順に説明していくものである．図中の写真は物体追跡の結果を示しており，その左下の写真はインデキシングする対象として選択されている映像である．「把持物体の検出→注目喚起行動の検出→インデキシング」という流れに対応して，物体上に表示

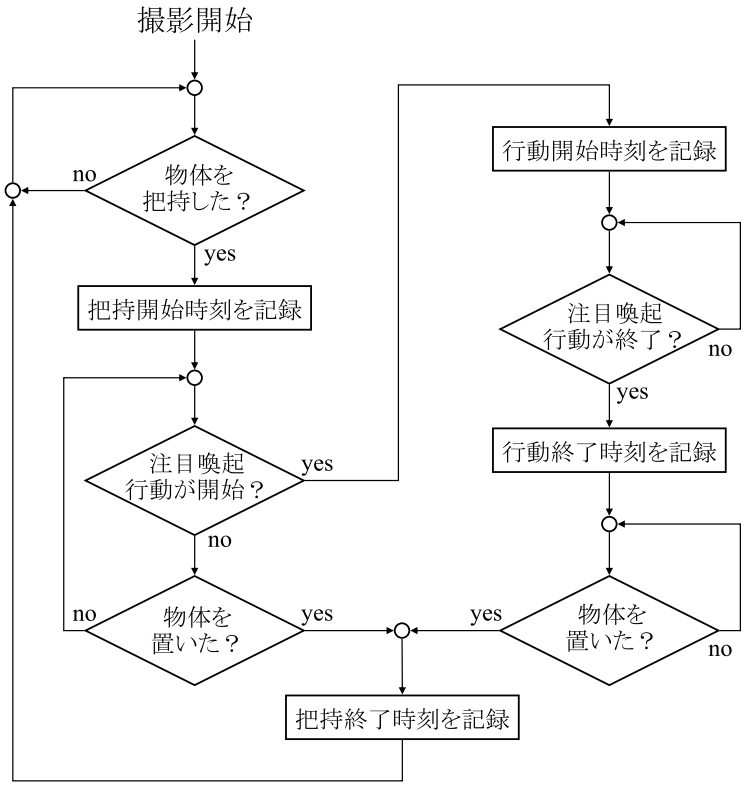


図 5.5: インデキシングの流れ

されている枠が「破線→太い実線→実線」と変化している．

5.5 物体追跡の評価実験

5.5.1 レンズ歪みと視点位置の補正

各画像センサにレンズ歪みがあることと，三つのセンサの光軸を完全に一致させるのは難しいことから，各センサから得られた画像を補正する必要がある．本研究では，可視光カメラから得られた画像を基準として，赤外線カメラとステレオカメラから得られた画像を変換する．補正には，レンズ歪みを補正するために 2 次の幾何補正を，視点位置の不一致を補正するために 2 次元射影変換を用いる．ここで視点の位置合わせに 2 次元射影変換を用いるのは，各画像センサから見て作業空間の奥行き範囲が比較的狭いためである．

実際の計算には，2 次幾何補正と 2 次元射影変換の両方の機能を兼ねる 3×5 の行列を用いる．補正行列を求めるためにキャリブレーションボードを用いるが，これには作業空間の中心（作業机の中心）に置いたときに各画像センサの視野全体が埋まる大きさのものを用意し，その上に 25 点の特徴点を配置した．これらの特徴点を各センサから得られる画像からそれぞれ抽出し，以下の補正行

(a) 写真は把持物体の追跡経過を表示した例．左下に上書きされている写真は，自動撮影・編集システムで取得されている映像．

(b) 把持物体が検出されると，その開始時刻がインデックスとして付加される．検出された物体（下中央）の上に枠が表示されている．

(c) 注目喚起行動が起こって一連の作業が終了すると，その開始・終了時刻が作業対象となっていた物体の ID と共に記録される．

(d) 物体が机に置かれたところで把持物体の位置・テクスチャ・作業内容が記録される．

(e) 前に検出された物体の情報は保持したまま，新たな把持物体が検出され，同様の処理が始まる．

図 5.6: 映像インデキシングの例

列 $M_{3\times5}$ を赤外線カメラとステレオカメラについて求める．

$$\begin{pmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_n \\ 1 & \dots & 1 \end{pmatrix} = M_{3\times5} \begin{pmatrix} u_1 & \dots & u_n \\ {u_1}^2 & \dots & {u_n}^2 \\ v_1 & \dots & v_n \\ {v_1}^2 & \dots & {v_n}^2 \\ 1 & \dots & 1 \end{pmatrix} \tag{5.3}$$

ここで，$(x_1, y_1) \sim (x_n, y_n)$ は可視光カメラから得られる画像上の特徴点座標，$(u_1, v_1) \sim (u_n, v_n)$ は補正前の赤外線カメラまたはステレオカメラから得られる画像上の特徴点座標である．計算には最小二乗法を用いた．

5.5.2 把持物体の追跡精度

まず物体追跡手法の精度を調べるために，(a) 机の上に一つだけ物体が存在し，背景中に人物が存在しないシーンと，(b) 机の上に複数の物体が存在し，背景中で人物が動いているシーンを用意した．シーンの例を図 5.7 に示す．

実験では，3 人の被験者に同時に一つもしくは二つの物体を掴んで自由に動かしてもらった．各

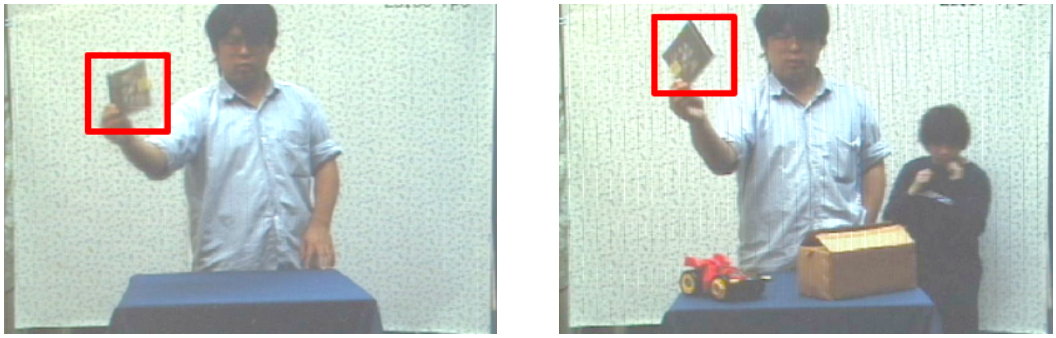


図 5.7: 追跡例（左：人物のみ，右：人物に加え，机上に静止物体があり背景に別の人物が動いている）

表 5.2: 追跡結果（単位はフレーム数）

	把持物体のみ	複雑な環境
全フレーム数	1350	1350
正解数	1316 (97.5 %)	1259 (93.3 %)
検出漏れ数	30 (2.2 %)	11 (0.8 %)
誤検出数	4 (0.3 %)	80 (5.9 %)

シーンについて 15 秒の映像（450 フレーム）を評価に使用し，フレームごとに正解・検出漏れ・誤検出の 3 通りに分けて数えた．ここで「検出漏れ」とは把持物体が検出されなかった場合,「誤検出」とは把持物体以外の領域が検出された場合を表す．

結果を表 5.2 に示す．結果より，背景で人物が動いており，作業空間中に似たような物体が存在する複雑な環境でも，精度良く把持物体を追跡できていることがわかる．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，これは十分に良い結果であるといえる．

5.5.3 作業内容の検出精度

??章で述べた作業内容の検出について，その検出率を調べた．今回の実験では，机の上にある 4 つの物体の中から適当に 1 つもしくは 2 つの物体を選び，それらを用いて「提示」「分離」「組み合わせ」の作業を各 80 回ずつ行った．各作業について，正解数・検出もれ数・誤検出数を数え，作業総数に対する割合を計算した．ここで,「検出もれ数」とは作業が検出されなかった場合,「誤検出数」とは検出された作業の種類が実際に行われたものと違う場合を表す．

結果を表 5.3 に示す．どの作業についても 7〜8 割の正解率となった．把持物体の検出もれに起因する作業の検出もれは 1〜2 割ほどあるものの，作業の誤検出はほとんどなかった．実際に検出された組み付け作業の例を図 5.8 に示す．左の図から順に，組み付け前の物体が 2 つある状態，組み付けている途中の状態，組み付けが終了し物体が 1 つとなった状態を表す．

表 5.3: 作業内容検出結果

	提示	分離	組付
作業総数	80	80	80
正解数	70 (87.5%)	61 (76.3%)	68 (85.0%)
検出もれ数	10 (12.5%)	19 (23.8%)	11 (13.8%)
誤検出数	0 (0.0%)	0 (0.0%)	1 (1.3%)

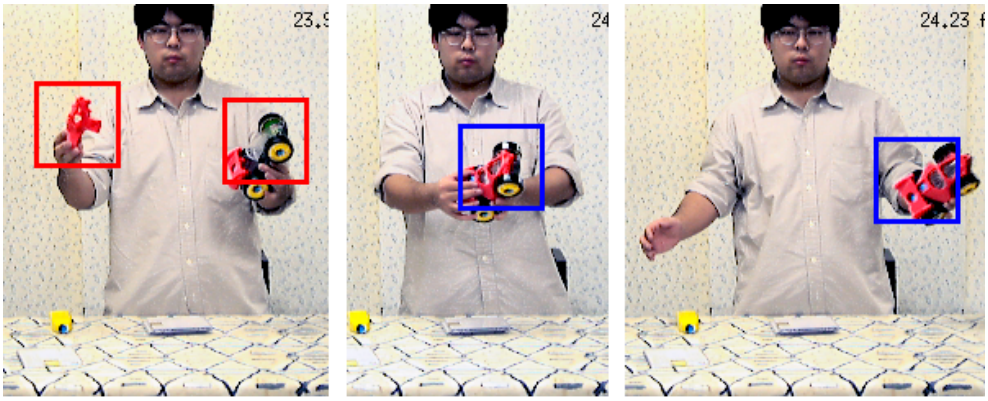


図 5.8: 組み付け作業の例

5.6 インデックス付き映像の利用

インデックス付き映像の利用例として，次の三つを紹介する．

- 物体アイコンを用いた映像ビューア
- 対話的映像メディアのためのコンテンツ
- 複合コミュニティ空間における注釈の共有提示

5.6.1 物体アイコンを用いた映像ビューア

机上作業の映像マニュアルにおいて，ある物体に関する作業方法を検索したいとき，その物体のテクスチャが表示されたアイコンをクリックすることで該当する映像クリップが再生されれば便利である．この目的を実現するビューアを実際に作成した．インデキシング処理の流れは図 5.5 と同様である．図 5.9 左にその様子を示す．図 5.9 右に示すように，右のウィンドウに作業に登場した物体像のアイコンが並んでおり，これらのどれか一つを選んでクリックすると，それに関連する映像セグメントが簡単に閲覧できる．

これを応用すれば，例えば，人の作業をロボットやコンピュータが見守り，分からないところがあれば手に持っている物体をカメラに向けることで，その解決方法を映像で見せてくれるといった

図 5.9: 映像インデキシングの様子と物体アイコンによる注釈映像の検索

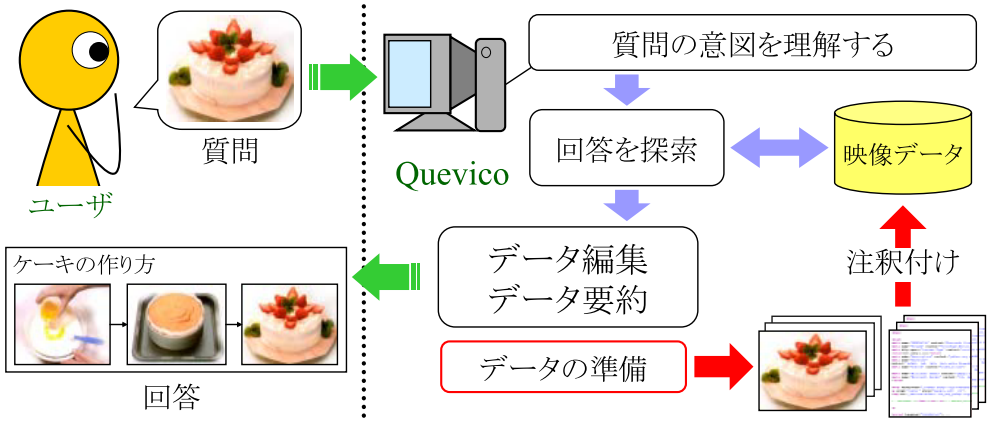


図 5.10: QUEVICO：対話型映像メディアの概要

ことも考えられる．

5.6.2 対話型映像メディアのためのコンテンツ

伊津野らは，組立て作業や料理などを行う際に，作業者の質問に対して適切な答えを与えてくれる “対話型メディア” を扱うモデルとして，“QUEVICO —マルチモーダルデータのための QA モデル—” を提案している [36]．QUEVICO は，図 5.10 に概要を示すように，(1) ユーザから発話や文章で与えられた質問を解析し，(2) 予め用意されたデータベースから答えとなるデータを取得，(3) それらを要約・編集してユーザに回答を提示する．この際，(2) の部分で用いるデータベースの映像コンテンツは，シナリオと映像内容に基づいて手作業で QUEVICO 形式のインデックスを付加している．これに対して本研究では，映像の自動インデキシングの結果を QUEVICO のデータとして利用することで，映像の撮影〜編集〜インデキシング〜提示までを一貫して行う試作システムを構築

```
<quevico>
<scene>
  <task id="t1" name="assemble" output="#o1" method="#v7"
        instrument="#o2">
    <object id="o1" name="toy car"/>
    <object id="o2" name="power tool" description="#a3"
            state="#v8 #v13"/>
      <task id="t2" name="attach" patient="#o3" instrument="#o2"
            input="#o4" method="#v2">
        <object id="o3" name="chassis" state="#v9 #v10"/>
        <object id="o4" name="body cover" state="#v10"/>
      </task>
    ...
</scene>
<stream name="video(action)" src=" ">
    <vsegment id="v1" begin=" " end=" "/>
        <point id="p1" time=" " x="" y=" " object=" "/>
        <point id="p2" time=" " x="" y=" " object=" "/>
        ...
    </vsegment>
    <vsegment id="v2" begin=" " end=" "/>
    ...
</stream>
<stream name="video(patient)" src=" ">
...
</stream>
<stream name="audio(speech)" src=" ">
  <asegment id="a1" begin=" " end=" ">
    I will now start to explain how to assemble a toy car.
  </asegment>
  <asegment id="a2" begin=" " end=" ">
    For a start, we use this power tool to assemble a toy car.
  ...
</stream>
...
</quevico>
```

図 5.11: シナリオから得られた QUIEVICO 形式のインデックス

した.

まず，料理や組立て作業のためのシナリオを用いて,「タスクの名前」「タスクの意味」「物体の名前」「物体の役割」「材料」「完成品」などの意味的なデータを用意する．これは既存の構文解析技術などで自動化することができる．次に，シナリオと実際の作業進行との整合をとることにより，シナリオのある部分が映像や発話データのどの部分に当たるのかを知る．この処理を本システムで担う．教示シーンの状況（作業進行状態や物体位置）を認識することにより，シナリオからは得られないデータを補完できる．

以下，車の模型を組み立てる作業について，具体的な QUEVICO への適用方法について説明す

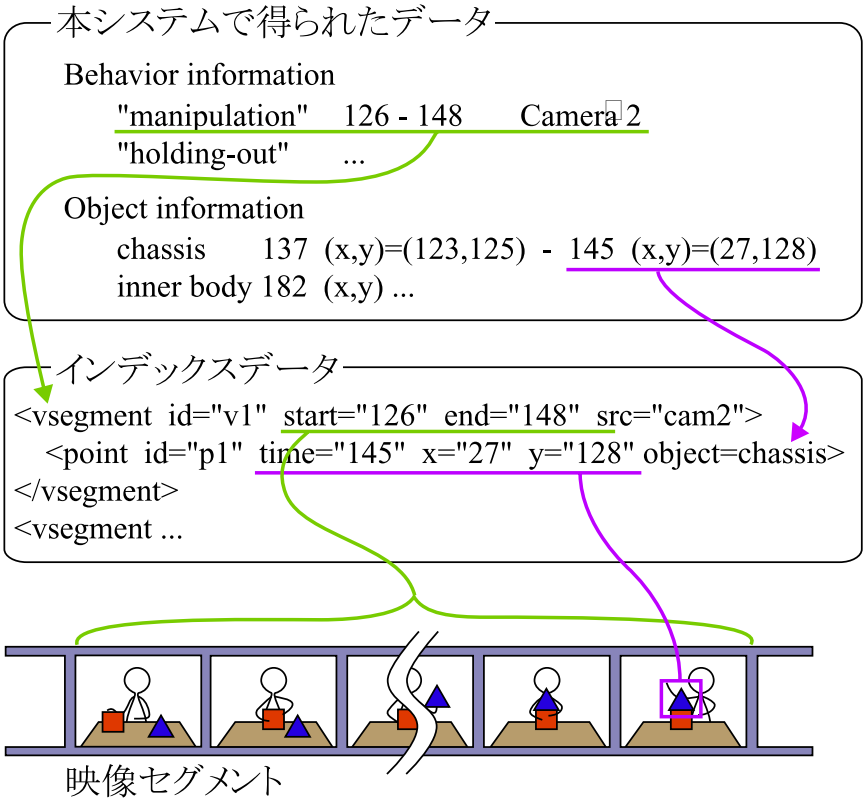


図 5.12: 本システムによるインデックスの補完

る．まず，シナリオから抽出できる情報から生成したインデックスを図 5.11 に示す．QUEVICO にはタスクとオブジェクトに関するインデックスがあり（task タグと object タグ），それぞれ ID と名前に加えて映像セグメント（vsegment タグ）と音声セグメント（asegment タグ）の開始・終了時刻を記述する必要がある．また，オブジェクトについては，映像上における位置のインデックス（position タグ）もある．図 5.11 に示すように，シナリオから抽出できるデータだけでは，vsegment タグ，asegment タグ，position タグが記述できない．本システムを用いて空白となっていたインデックスを補完する様子を図 5.12 に示す．asegment タグは，シナリオと発話データの DP マッチングによって補完する．これらは手作業では最も苦労する部分であるため，自動化することにより労力が大幅に削減できる．

実際に上述の手順で得られたインデックス付き映像を QUEVICO のデータとして用い，質問に対して回答を表示させた結果の一例を図 5.13 に示す．前述したインデキシングにより，例えば次のような質問に答えることができる．

組み立てには何を使いますか？ パワーツールの写っているショットの 1 フレームか，物体追跡処理の結果として得られるパワーツールのテクスチャ画像を表示する．

パワーツールとは何ですか？ パワーツールの使い方を説明している映像セグメントか，その部分の

図 5.13: QUEVICO のコンテンツとしての利用

発話を表示する．

ボディを取りつけるにはどのようにすればいいですか？ ボディをシャーシに取りつける部分の映像セグメントを表示する．

タイヤを取りつけるときに注意することはなんですか？ タイヤの取りつけ前に注意事項を説明している部分の発話を表示する．

前タイヤとはどのようなタイヤですか？ 前タイヤについて説明している部分の文章を表示する．

5.6.3 複合コミュニティ空間における注釈の共有提示

“複合コミュニティ空間” とは，現実世界と仮想世界が融合した複合現実の感覚を複数の人間が共有することのできる人工空間である．図 5.14 に示すように，ここでは複数の人間同士が現実世界で行っている視覚的な情報交換に加え，現実世界にはない新しい視覚情報を同時に共有する．本節では，この複合コミュニティ空間において “注目の共有” を支援する機能として，インデックス付き映像を利用する．例えば，ある人が注目しているものを他の人に強調して提示することによって，話題の中心を明確にし，意志の疎通を円滑にすることができる．また，注目が必要とされている場面を注釈情報として撮影し，その映像を物体と関係付けて記録しておくことによって，時間や場所が異なる場合にも注目対象に関する情報が利用できる．複合現実感を利用した注釈付けについて，これまで数多くの研究が行われてきた [37]–[38]．これらに対し本研究は，人間の自然な行動から注目すべき状態（注釈情報を提示・記録すべき状態）や注目すべき物体を検出しているところに特徴がある．

本研究では，複数人が複合現実空間の中で対話している場面において，注目が要求されている状態（人物が注目を誘導している状態）を検出し，その様子を注釈情報として記録したり，他のユーザに提示することを考える．

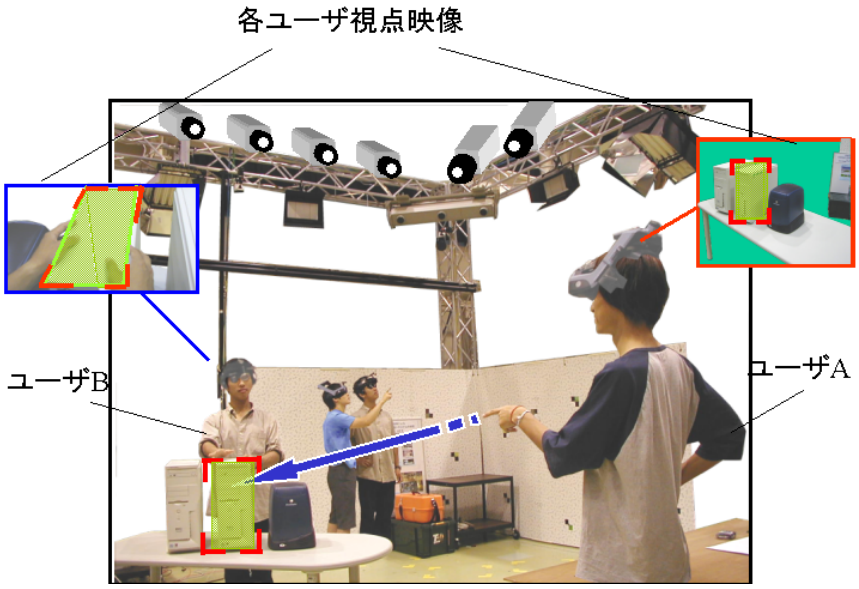


図 5.14: 複合コミュニティ空間の例

このようなことを実現するために必要となる機能には次の四つがある（図 5.15）.

(a) 注目対象を複数カメラで追跡・撮影する機能： 注目対象となり得る部分，例えば人物の主要部位や主要物体・場所等を自動的に追跡し撮影しておくことが必要である．ここでは 3 章で述べたカメラ制御手法が利用できる.

(b) 注目が必要となる状態を検出する機能： 人物の発話や動作を認識して，その人物やその周辺に注目することが要求されている状態を検出する必要がある．ここでは 4 章で述べた注目喚起行動を対象とする.

(c) 映像を物体と関係付けて記録する機能： 機能 (a) によって撮影された物体に関する説明やその際の人物の様子を，注釈映像としてその物体に関連付けて記録する．そのためには，各時点で注目されている物体を常に検出し追跡しておく必要がある．このようにして物体に関連付けられた注釈情報は，時間や場所を隔てたコミュニケーションに対して大きな補助となる.

(d) 注目対象の強調や注釈情報の提示機能： 同じ場所で共に行動している他の人物に対しては，撮影されている映像をそのまま提示するだけでもコミュニケーションの補助となる．また時間や場所を隔てた場合にも，機能 (c) で記録しておいた物体に関する過去の説明映像を提示することができる.

具体的な例として，ある人物が物体を掲げながらその利用方法を説明する場合を考えてみよう．まず複数のカメラがその人物がいる空間・手先・その他の重要な部分を常に追跡・撮影している［機能 (a)］．ある時刻にその人物が物体を掲げながらその物体に関する説明を始めると，システムはそ

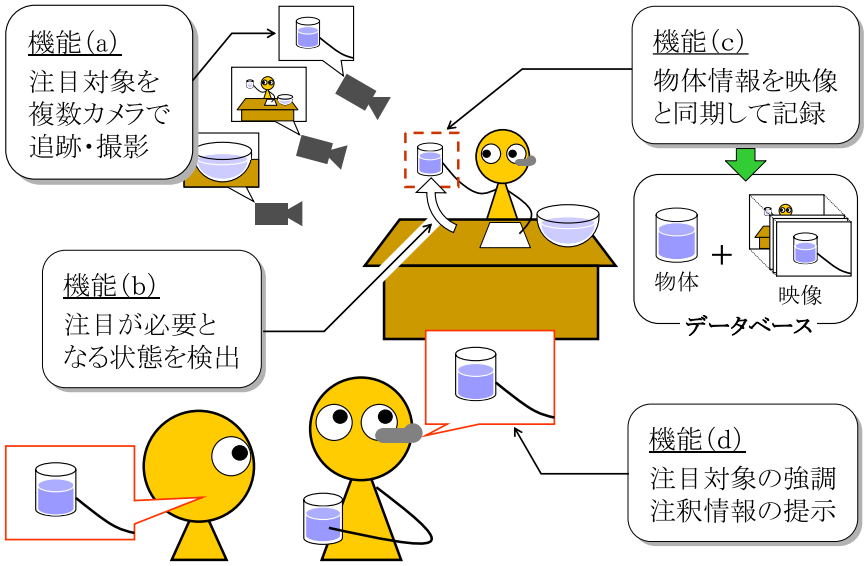


図 5.15: 注目の共有を実現するための四つの機能（右）

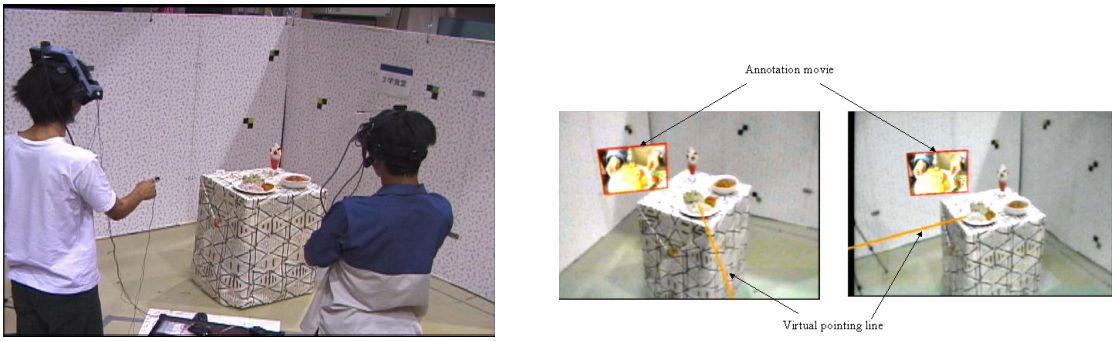


図 5.16: 複合コミュニティ空間における注釈映像提示（右の二つの写真は各人の HMD に表示されている映像）

の動作を注目が必要な状態として検出する［機能 (b)］．これをトリガとして，各カメラで撮影されている映像からその説明を最も良く捉えている映像が選択され，掲げられた物体と関連付けて記録される［機能 (c)］．これらの映像は同時刻に複合コミュニティ空間にいる他の人物には実時間で提示され，時間や場所を隔てた空間にいる人物には物体の位置やテクスチャなどを検索キーとしてデータベースから注釈映像を検索し提示する［機能 (d)］．

実際に複合コミュニティ空間で注釈情報を共有した例を図 5.16 に示す．この例は，図 5.6 の場所に後から別の二人の人物が訪れ，机の上の料理について注釈映像を見つつ話し合っている場面である．

注釈提示システムでは，人物の指差し動作とその方向を頭部の HMD と手先につけた磁気センサにより検出し，指し示された先にある物体をデータベースからテクスチャマッチングにより検索する．データベースに似通った物体が存在すれば，それに注釈付けされた映像クリップがその物体付

近に表示され，その場にいる複数のユーザでその情報を共有できる．図 5.16 からわかるように，各ユーザの視点に合わせて映像クリップと指示方向を示す線が表示される．このシステムの詳細については文献 [39] を参照されたい．

5.7 まとめ

机上作業シーン映像の自動撮影・編集システムと物体追跡システムを組み合わせることにより，物体に関する情報（位置・テクスチャ・把持状態・作業内容）を映像と同期して記録する映像インデキシングを実現した．可視光カメラ・赤外線カメラ・ステレオカメラから肌色領域・動領域・肌温領域・特定距離領域を抽出して論理積をとることで，物体の大きさ・色・形状などの予備知識がなく背景が常に変化するという条件の下でも精度良く物体を検出できる手法を提案した．

また，物体に関するインデックスが付加された高度な映像コンテンツの利用例として，物体のアイコンをキーとした映像検索，対話型映像メディアのコンテンツ，複合現実空間におけるコンテンツを紹介し，その有用性の一端を示した．

第6章　結論

本研究では，机上作業シーン映像の自動撮影・編集手法を提案し，試作システムでの評価実験を通してその有効性を明らかにした．本論文では，2章で本研究の狙いと独自性を示した後，大きく三つの課題に分けて，その考え方・提案手法・成果について述べた．以下，それぞれの結論をまとめる．

カメラマンの機能の実現

机上作業シーン映像では，撮影する対象が何であるかによってではなく，対象のどのような状態に注目するかによってカメラワークが決まるという考え方を提案し，その注目すべき状態の典型として＜外観＞・＜動き＞・＜周辺関係＞の三つを定義した．また，それらの状態を適切に撮影するためのカメラワークは，対象をできるだけ視野中央に捉えるよう追跡することと，視野ができるだけ動かないよう固定することのトレードオフを調整することによって得られるという考え方を導入した．これをカメラの自動制御で実現するために，対象の動き（反復・停留）に応じて追跡モードと固定モードを切り替える「可変枠制御」を提案した．

また，本研究の三つのカメラワーク分類に意味があること，及び可変枠制御が他の代表的なカメラ制御手法よりも優れていることを，CG世界と現実世界のそれぞれで撮影したショットを主観評価してもらうことで確認した．更に，本システムで撮影したショットとプロカメラマンが撮影したショットと比較した結果，“鮮明感”・“好感”・“内容理解のしやすさ”についてはプロカメラマンの撮影したものと遜色ないが，“連続感”・“人間的”については改善の余地があることがわかった．

ディレクタの機能の実現

テレビ番組の分析より，机上作業シーン映像は，ショットA（マスターショット）とショットC（手元・物体・場所のクロースアップ）の繰り返しでその90%以上が構成されていることを明らかにした．一方，机上作業シーンにおけるアクションの代表的なものとして，話者がある箇所に注目を集めようとする行動に着目し，注目喚起行動と名付けた．注目喚起行動は，他のアクション（話者の行動）に比べて話者の意図が明確であり，テレビ番組のショット切替えとの間にも約50%の共起率がみられた．本研究では，この注目喚起行動をショットAとショットCの切替えのトリガとする編集モデルを考案し，話者の発話と手の位置関係から注目喚起行動を自動検出する手法を提案した．

また，注目喚起行動のみに基づいて編集した映像がテレビ番組に基づいた編集結果と同程度の評価が得られることを示し，注目喚起行動が机上作業シーン映像の編集トリガとして高い有効性を持

つことを確認した．更に，注目喚起行動によるショット切替えと話者が意図的にショット切替えを行う他の手法との使用感を比較した結果，シナリオのある・なし，聞き手とのインタラクションのある・なしに関わらず，高い評価が得られることを確認した．

映像インデキシングとその利用

机上作業シーン映像の自動撮影・編集システムと物体追跡システムを組み合わせることにより，物体に関する情報（位置・テクスチャ・把持状態・作業内容）を映像と同期して記録する映像インデキシングを実現した．可視光カメラ・赤外線カメラ・ステレオカメラから肌色領域・動領域・肌温領域・特定距離領域を抽出して論理積をとることで，物体の大きさ・色・形状などの予備知識がなく背景が常に変化するという条件の下でも精度良く物体を検出できる手法を提案した．

また，物体に関するインデックスが付加された高度な映像コンテンツの利用例として，物体のアイコンをキーとした映像検索，対話型映像メディアのコンテンツ，複合現実空間におけるコンテンツを紹介し，その有用性の一端を示した．

本研究を通して，誰でも手軽に自分の持っている知識・技術を映像コンテンツとして記録するための自動撮影・編集システムを実現した．本システムは，これまでに数多くの撮影実験とデモンストレーションを通してその有用性を確認しており，実利用を考えることのでできる段階にあるといえる．今後は，研究と実利用の場を密接に結び付け，高度情報化社会におけるコンテンツの充実およびコンテンツの新たな利用についての研究へと発展させていきたい．

また，知識・技術の映像化において，教示シーンから何を切り取り（撮影），それらをどう再構成するか（編集）について，根本となる考え方と技術を提供した．本研究で提案したカメラワークの考え方と注目喚起行動という編集トリガは，基本的な要素であると同時に，これらだけでも専門家による映像に比肩する映像を生成することができる．これらの要素の体系的な評価は，教示映像の自動制作における有意な学術的知見となろう．このような知見を一つずつ丁寧に吟味して積み重ねていくことにより，自動映像制作の基礎となる学術的体系が形成されるだろう．

残る課題として，教示映像の第一の目的である「教示内容を見やすく分かりやすく伝えること」については，カメラマンの機能ではズーム制御や速度制御などが必要であり，ディレクタの機能では注目が喚起されない作業の開始を扱える枠組みが必要である．また，カメラ台数を削減して無駄の少ない撮影を実現するには，カメラマンの機能とディレクタの機能の連携も考える必要がある．

謝辞

6 年間に渡り懇切丁寧なご指導を賜りました筑波大学システム情報工学研究科 大田友一教授，及び実質指導教官として親身になって研究を支えてくださった京都大学学術情報メディアセンター 中村裕一教授に感謝の意を表します.

本論文を査読してくださり，また貴重なご意見を頂きました筑波大学システム情報工学研究科 鬼沢武久教授，国立情報学研究所 佐藤真一教授，筑波大学システム情報工学研究科 葛岡英明助教授に心からお礼申し上げます.

また，共同研究者として共に研究してきた里雄二君，伊藤雅嗣君，伊津野英克君，同じ研究グループとして数多くの実験に協力してくれた尾形涼君，津吹陽介君，西崎隆志君，評価映像の出演者として協力して頂いた服部繭さん，鈴木梓さん，常盤ひかりさん，桑野美智子さんに深く感謝致します．そして，数々のご助言を頂き，また多くの評価実験に快く協力してくださった本学画像情報研究室の皆様に感謝致します.

なお本論文では，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ます.

参考文献

[1] IT 戦略本部. e-japan 重点計画. *http://www.kantei.go.jp/jp/singi/it2/index.html*.

[2] M. Davis. Garage cinema and the future of media technology. *Comm. ACM*, Vol. 40, No. 2, pp. 42–48, 1997.

[3] Wikipedia. *http://ja.wikipedia.org/*.

[4] Steven D. Katz. *Film directing shot by shot*. Michael Wiese Productions, 1991.

[5] 吉田直哉. 映像とは何だろうか：テレビ制作者の挑戦. 岩波書店, 2003.

[6] Daniel Arijon. *Grammar of the Film Language*. Focal Press, Boston, 1976.

[7] ダニエル・アリホン. 映画の文法. 紀伊國屋書店, 1980. 岩本憲児, 出口丈人 訳.

[8] A. Bobick and C. Pinhanez. Controlling view-based algorithms using approximate world models and action information. *Proc. CVPR*, pp. 955–961, 1997.

[9] 椋木雅之, 西口敏司, 池田克夫, 美濃導彦. テンプレート映像に基づく一定移動パターンの自動撮影手法. 画像電子学会誌, pp. 806–814, 2002.

[10] M. Gleicher and J. Masanz. Towards virtual videography. *Proc. ACM Multimedia*, pp. 375–378, 2000.

[11] 三塚和幸, 山村恵一, 山中徳唯. インテリジェンスロボットカメラの開発. テレビジョン学会技術報告, Vol. 17, No. 51, pp. 33–37, 1993.

[12] 加藤大一郎, 山田光穂ほか. スタジオ番組における放送カメラマンのカメラワークと視線の動きの分析. テレビジョン学会誌, Vol. 49, No. 8, pp. 1023–1031, 1995.

[13] 加藤大一郎, 山田光穂ほか. 被写体を追尾撮影時の放送カメラマンのカメラワーク分析. テレビジョン学会誌, Vol. 50, No. 12, pp. 1941–1948, 1996.

[14] S. Mukhopadhyay and B. Smith. Passive capture and structuring of lectures. *Proc.ACM Multimedia*, pp. 477–487, 1999.

[15] 宮崎英明, 亀田能成, 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. 信学論 D-II, Vol. J82, No. 10, pp. 1598–1605, 1999.

[16] Q. Liu, Y. Rui, A. Gupta, and JJ. Cadiz. Automating camera management for lecture room environments. *Proc. ACM CHI*, pp. 442–449, 2001.

[17] 錦織修一郎, 菅沼明, 谷口倫一郎. 黒板講義を対象とした講義自動撮影システムの構築. 信学技報, pp. 79–86, 2000.

[18] 大西正輝, 村上昌史, 福永邦雄. 状況理解と映像評価に基づく講義の知的自動撮影. 信学論 D-II, Vol. J85, No. 4, pp. 594–603, 2002.

[19] 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*, pp. 1237–1240, 2000.

[20] 井上智雄, 岡田謙一, 松下温. テレビ番組のカメラワークの知識に基づいた tv 会議システム. 情処論, Vol. 37, No. 11, pp. 2095–2104, 1996.

[21] 尾形涼, 中村裕一, 大田友一. 制約充足と最適化による映像編集モデル. 信学論 D-II, Vol. J87, No. 12, pp. 2221–2230, 2004.

[22] 西崎隆志, 尾形涼, 中村裕一, 大田友一. 会話シーンを対象とした映像コンテンツの取得と編集. 画像の認識・理解シンポジウム, Vol. I, pp. 451–456, 2004.

[23] L. He, M. F. Cohen, and D. H. Salesin. The virtual cinematographer: A paradigm for automatic real-time camera control and directing. *Proc. SIGGRAPH 96*, Vol. Computer Graphics Proceedings, pp. 217–224, 1996.

[24] S. M. Drucker and D. Zeltzer. Camdroid:a system for implementing intelligent camera control. *Proc. Symposium on Interactive 3D Graphics*, pp. 139–144, 1995.

[25] スティーブン・D・キャッツ. 映画監督術 SHOT BY SHOT. フィルムアート社, 1996. 津谷祐司 訳.

[26] D. Kato, M. Yamada, et al. Analysis of the camerawork of broadcasting cameramen. *SMPTE Jounal*, pp. 108–116, 1997.

[27] 加藤大一郎, 石川秋男, 津田貴生, 福島宏, 山田光穂. カメラワーク分析と映像の主観評価実験. 映情学誌, Vol. 53, No. 9, pp. 1315–1324, 1999.

[28] 熊野雅仁, 岩本健, 有木康雄, 塚田清志. ボールと選手に着目したディジタルカメラワークの実現法—ディジタルシューティングによるサッカー解説映像生成システムに向けて—. 画像の認識・理解シンポジウム, Vol. II, pp. 341–346, 2004.

[29] M. Ozeki, M. Itoh, Y. Nakamura, and Y. Ohta. Tracking hands and objects for an intelligent video production system. *Proc. of Int'l Conf. on Pattern Recognition*, pp. 1011–1014, 2002.

[30] 尾形涼, 尾関基行, 中村裕一, 大田友一. 遠隔サイト間で注目を共有するための映像撮影・選択・伝送システム. *Forum on Information Technology*, pp. 43–50, 2002.

[31] D. Katou, T. Katsuura, and H. Koyama. Automatic control of a robot camera for broadcasting based on cameramen's techniques and subjective evaluation and analysis of reproduced images. *Journal of Physiological Anthropology and Applied Human Science*, Vol. 19, No. 2, pp. 61–71, 2000.

[32] 馬場口登ほか. 映像処理評価用映像データベースについて. 信学技報, Vol. PRMU2002-30, , 2002.

[33] 村上正行, 田口真奈, 溝上慎一. 日米間遠隔一斉講義における講師・受講生の評価変容の分析. 日本教育工学会論文誌, Vol. 25, No. 3, pp. 199–206, 2001.

[34] 望月俊男, 中原淳, 山内祐平, 西森年寿, 松河秀哉, 一色裕里, 松浦匡, 朝川哲司, 八重樫文, 加藤浩. 教室の授業と連携した e-learning とその評価分析－東京大学 iii online における社会人学生とフルタイムの学生の評価に着目して－. 教育システム情報学会誌, Vol. 20, No. 2, pp. 132–142, 2003.

[35] 近藤博仁, 孟洋, 佐藤真一, 坂内正夫. テロップ認識と顔照合を統合したニュース映像中人物の自動索引付けシステム. 電子情報通信学会 総合大会, Vol. D-12-190, , 1999.

[36] H. Izuno, Y. Nakamura, and Y. Ohta. Quevico qa model for video-based interactive media. *Proc. of Third International Workshop on Content-Based Multimedia Indexing*, pp. 413–420, 2003.

[37] R. T. Azuma. A survey of augmented reality. *Teleoperators and Virtual Environments*, Vol. 6, No. 4, pp. 355–385, 1997.

[38] 天目隆平, 神原誠之, 横矢直和. 赤外線ビーコンと歩数計測を用いたウェアラブル型注釈提示システム. 信学技報, Vol. IE2002-54, , 2002.

[39] 里雄二, 北原格, 中村裕一, 大田友一. 複合コミュニティ空間における注目の共有〜指示動作による注目の強調提示システム〜. VRSJ 第 6 回大会論文集, pp. 235–238, 2001.

[40] 有本卓. カルマン・フィルター. 産業図書, 1988.

[41] 西山清. パソコンで解くカルマンフィルタ, 第 5.2 章, pp. 59–64. 丸善, 1993.

[42] L. L. Thurstone. A law of comparative judgement. *Psychological Review*, Vol. 34, pp. 273–286, 1927.

[43] 大串健吾, 中山剛, 福田忠彦. 画質と音質の評価技術, 第 2.5 章. 昭晃堂, 1991.

発表論文リスト

査読付雑誌論文

1. 尾関基行，中村裕一，大田友一，
   “注目喚起行動に基づいた机上作業映像の編集，” 信学論 D-II.（採録決定）

2. M.Ozeki, Y.Nakamura, and Y.Ohta,
   “Automated Camerawork for Capturing Desktop Presentations,” IEE Proceedings on Vision, Image & Signal Processing. （条件付き採録）

3. 尾関基行，中村裕一，大田友一，
   “話者の注目喚起行動による机上作業映像の自動編集 —ユーザインタフェースの側面からの評価—，” 情報科学技術レターズ， pp.269-272，Sept. 2004.（FIT 論文賞）

4. 尾関基行，伊藤雅嗣，里 雄二，中村裕一，大田友一，
   “複合コミュニティ空間における注目の共有 〜注目誘導行動による物体への注釈付け〜，” 日本バーチャルリアリティ学会論文誌，Vol.8，No.4，pp.369-377，Dec. 2003.

5. 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，” 信学論 D-II，Vol.J86，No.11，pp.1606-1617，Nov. 2003.

査読付国際会議

1. M.Ozeki, Y.Nakamura, and Y.Ohta,
   “Video Editing based on Behaviors-for-Attention –Approach to professional editing by a simple scheme–,” Proc. of IEEE Int’l Conf. on Multimedia and Expo (ICME), TP9-4(cdrom), Taipei, Taiwan, 30 Jun. 2004.

2. M.Ozeki, Y.Nakamura, and Y.Ohta,

"Automated Camerawork for Capturing Desktop Presentations – Camerawork design and evaluation in virtual and real scenes," Proc. of 1st European Conf. on Visual Media Production (CVMP), pp.211-220, London, UK, 16 Mar. 2004.（優秀論文として，IEE Proceedings on Vision, Image & Signal Processing に推薦）

3. M.Ozeki, M.Itoh, H.Izuno, Y.Nakamura, and Y.Ohta,
"Object Tracking and Task Recognition for Producing Interactive Video Content – Semi-automatic Indexing for QUEVICO," Proc. of Knowledge-Based Intelligent Information & Engineering Systems (KES), pp.1044-1053, Oxford, UK, 5 Sept. 2003.

4. M.Itoh, M.Ozeki, Y.Nakamura, and Y.Ohta,
"Simple and Robust Tracking of Hands and Objects for Video-based Multimedia Production," Proc. of IEEE Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI), pp.252-257, Tokyo, Japan, 1 Aug. 2003.

5. M.Ozeki, Y.Nakamura, and Y.Ohta,
"Human Behavior Recognition for an Intelligent Video Production System," Proc. of 3th Pacific-Rim Conf. on Multimedia (PCM), pp.1153-1160, Hsinchu, Taiwan, 18 Dec. 2002.

6. M.Ozeki, M.Itoh, Y.Nakamura, and Y.Ohta,
"Tracking Hands and Objects for an Intelligent Video Production System," Proc. of 15th(16th?) Int'l Conf. on Pattern Recognition (ICPR), pp.1011-1014, Quebec, Canada, 14 Aug. 2002.

7. M.Ozeki, Y.Nakamura, and Y.Ohta,
"Camerawork For Intelligent Video Production – Capturing Desktop Manipulations," Proc. of IEEE Int'l Conf. on Multimedia and Expo (ICME), pp.41-44, Tokyo, Japan, 23 Aug. 2001.

国内学会発表

1. 尾関基行，中村裕一，大田友一，
"話者の意図に基づいた自動編集のためのユーザインタフェース，" 信学技報 (PRMU)，広島，21 Oct. 2004.

2. 尾関基行，中村裕一，大田友一，
"注目喚起行動を用いた机上作業映像のための自動編集手法，" 画像の認識・理解シンポジウム (MIRU)， pp.457-462，函館，27 Jul. 2004.

3. 尾関基行，中村裕一，大田友一，
“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—，” 信学技報 (PRMU)， PRMU2003-45，pp.7-12，高知，17 Jul. 2003.

4. 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)，PRMU2003-45，pp.13-18，高知，17 Jul. 2003.

5. 伊藤雅嗣，尾関基行，中村裕一，大田友一，
“映像インデキシングのための手と把持物体のロバストな認識と追跡，” 画像センシングシンポジウム (SSII)， pp.277-282，神奈川，12 Jun. 2003.

6. 尾形 涼，尾関基行，中村裕一，大田友一，
“遠隔サイト間で注目を共有するための映像撮影・選択・伝送システム，” 情報科学技術フォーラム (FIT)， pp.43-50，東京，25 Sept. 2002.

7. 伊藤雅嗣，尾関基行，中村裕一，大田友一，
“映像メディア取得のための手と把持物体の追跡と認識 —多種類の画像センサによるロバストな実時間追跡—，” 信学技報 (PRMU)， PRMU2002-26，pp.43-50，広島，27 Jun. 2002.

8. 尾関基行，伊藤雅嗣，中村裕一，大田友一，
“複合コミュニティ空間における注目の共有 ～人物動作理解による物体への注釈付け～，” 日本バーチャルリアリティ学会 第 6 回全国大会，pp.239-242，長崎，20 Sept. 2001.

9. 伊藤雅嗣，尾関基行，中村裕一，大田友一，
“プレゼンテーションにおける手と把持物体の認識と追跡，” 電子情報通信学会 ソサエティ大会，Vol.D-12-39，東京，18-21? Sept. 2001.

10. 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，” 電子情報通信学会 総合大会（シンポジウム発表），Vol.SD-5-4，pp.357-358，滋賀，29 Mar. 2001.

11. 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 —動作認識による映像のタグ付け—，” 第 6 回知能情

報メディアシンポジウム (IIM)， pp.69-74，東京，6 Dec. 2000.

12. 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—，” 信学技報 (PRMU)， PRMU2000-104，pp.31-38，茨城，16 Nov. 2000.

著書など

1. Y.Nakamura, M.Ozeki, and Y.Ohta,
“An Intelligent System for Capturing Presentation on Desktop Manipulations — Supporting for Video Contents Production,” In ”Virtual Environments for Teaching and Learning” (World Scientific Publishing) Chapter 10, 1 Nov. 2002.

付 録A　使用機器

A.1　使用機器の仕様

本システムで使用している機器・ソフトウェアを表 A.1 に，カメラの仕様を表 A.2 にそれぞれ示す．

後で議論するように，机上作業シーンに登場する速く動く対象（手振り動作時の手先）に比べて，このカメラの最大速度は十分に速い．しかし，初動時の速度が遅いことやシステム構造に起因する遅延が生じるため，それらをできる限り補償するため，カメラは常に最大速度で制御している．ズーム値は，注目対象物それぞれに光学ズームの範囲で 3 種類予め用意する．シャッタースピード・露出・ホワイトバランスは，それぞれオートに設定している．

位置センサとして用いる The Flock of Birds（以下，FOB）は，三次元位置（X・Y・Z）と各軸まわりの回転角を測定できる 6 自由度の磁気センサである．FOB 本体では 100Hz で測定を行なっているが，本システムではそこから 30Hz でデータを受け取って処理している．本システムのカメラ制御と行動検出手法では，30Hz で位置が獲得できれば十分である．また，移動精度は 1.8mm (RMS) であり，3.7m 離れた位置から撮影する場合（本スタジオの環境），最大ズーム（水平画角 6.6°）でも一画素が約 1.85mm の範囲に対応するため，十分な精度であるといえる．ただし，追跡精度についてはカメラ制御遅れの影響が大きく，位置センサの精度は問題とならない．これについては次節で説明する．

A.2　システム構成上の制約

A.2.1　手首に装着したセンサによる物体・場所の追跡

3 章 3.3 節で注目対象物（追跡対象）として話者・手先・物体・場所の四つを挙げたが，すべての物体と場所に位置センサをつけることは現実的に難しい．一方，机上作業シーンで物体が自律的に移動することは稀であり，物体の移動は基本的に人の手によって行われる．また，注目すべき場所や置いてある物体もほとんどが手の届く範囲にあるため，指差しなどで位置を示すことができる．そこで本研究では，注目すべき物体や場所の位置を手首に装着したセンサから求め，追跡撮影することを考える．机上作業では多くの物体が登場し組み合わされたり調理されたりして形を変化させるが，把持／指差しされた物体を追跡することで，事前情報を使うことなく効率良く注目すべき物体の位置を取得することができる．

本節では，手首につけたセンサで把持物体を追跡することに問題がないかを確認する．ここでは，

機器	品名	メーカー
カメラ	DVI-D100	ソニー
位置センサ	The Flock of Birds	Ascension Technology Corporation
映像切替器	ACS-310	朋栄
映像分割器	MV-94	朋栄
シリアルポート	Cyclades-Z	Cyclades
音声認識	Via Voice	IBM

表 A.1: 使用機器の一覧

映像信号	NTSC カラー JEITA 標準方式，1/4 インチ カラー CCD
レンズ	光学 10 倍，デジタル 40 倍，f=3.1〜31mm，F1.8〜F2.9
水平画角	6.6°〜65°
シャッタースピード	1/4〜1/10,000 秒（VISCA コントロール時）
パン機能	水平 ±100°，最大速度 300°/秒
チルト機能	垂直 ±25°，最大速度 300°/秒

表 A.2: DVI-D100 の仕様

物体を直接追跡した場合と手首を追跡した場合で，取得した映像に違いが出るかどうかを調べた．図 A.1 に例を示す．静止画からも分かるように，物体を直接追跡したショットと手首のセンサを元に追跡したショットの間での違いは主観的にはほとんどわからない（動画でも確認した）．ただし，指差しで示された物体や場所については手首側にずれた映像となってしまうため，指を差した方向にオフセットを加えた位置を狙う必要がある．本システムで使用している位置センサは方向も計測できるため，オフセットを加えたショットを撮影することは可能である．しかし，例えば右手で注目喚起行動を行なった際に，右手先のショットと指差した物体のショットのどちらに切り替えるかは，画像処理などを組み合わせて判断する必要がある．現時点では，指差しの際に手先を物体や場所に触れるまで伸ばすことで，把持した場合と同様のショットを取得している．

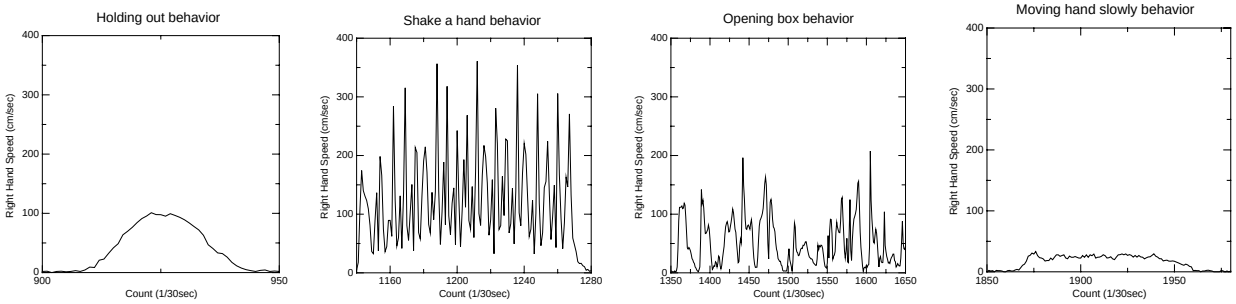
A.2.2 机上作業時の手先の速度とカメラ制御の遅れ

2 章で述べたように，手先などの素早く不規則に動く対象を撮影することは机上作業シーンの特徴の一つである．本節では，机上作業シーンの代表的な動作における手先の速度を調べ，本システムで用いる首振りカメラでどこまで追従できるか確認する．

ここでは,「提示する動作」「手を振る動作」「箱を開ける動作」「ゆっくりの移動」の四つの動作の速度を磁気センサの位置情報より計算した．結果を図 A.2 のグラフに示す．それぞれの最大速度と平均速度の概算値は表 A.3 のようになった．人の歩く速度は速くて 130cm/sec 程度であり，速度だけみても，講義シーンにおける人物追跡撮影に比べて机上作業シーンにおける手先の撮影が難し

右手先を追跡　物体を追跡　両手の中点を追跡　物体を追跡

図 A.1: 手首と物体にセンサを付けて追跡した映像



提示動作　手振り動作　箱空け動作　手の移動

図 A.2: 机上作業時の手先の速度のグラフ

いことがわかる．

「提示する動作」と「ゆっくりの移動」を追跡撮影している様子を図 A.3 と図 A.4 に示す．提示する動作の場合，クロースアップショットでは一時的に対象が画面から外れてしまう．また，ゆっくりの移動の場合，しばらく動作を続けても一定以上カメラが追いついてこなかった．画面中での磁気センサの位置を計測してカメラの追従がどの程度遅れているかを概算すると，それぞれ最も遅れている時点で約 7°/秒（提示する動作）と約 2°/秒（ゆっくりの移動）の遅れとなった．最も速い対象（360cm/秒）を追跡する場合でも約 50°/秒の速度があれば追跡はできるはずであり[1]，EVI-D100 は最大 300°/秒の速度が出るためカメラの性能で追いつけないものではない．よって，これはカメラの構造的な制御遅れであると考えられる．

また，カメラの制御開始の遅れを調べたところ，対象が動き始めてからカメラが追従を開始するまでに 10 フレーム弱（0.2〜0.3 秒）の遅れがあることがわかった．本システムで使用している位置センサの遅れはほとんど無視できるため，これもカメラの構造的な遅れであると考えられる．

以上のように，廉価なカメラで素早い対象をクロースアップで追跡する場合には，カメラ制御遅れの問題が無視できない．特に，対象が上下左右に反復する場合，カメラ制御遅れの影響が相まって非常に見苦しい映像となる．これは，人物など比較的ゆっくりと動く対象の撮影では顕在化しない問題であり，また，CG シーンを対象とした撮影では起こり得ない問題である．これに対して本研究で提案した可変枠制御では，反復やブレを検出してカメラを固定する機能を備えているため，カ

[1]約 3.7m 離れた位置から撮影する場合（本スタジオの環境）．

表 A.3: 机上作業時の手の速度

動作	最大速度	平均速度
提示する動作	100 cm/秒	40 cm/秒
手を振る動作	360 cm/秒	120 cm/秒
箱を開ける動作	210 cm/秒	50 cm/秒
ゆっくりの移動	30 cm/秒	15 cm/秒

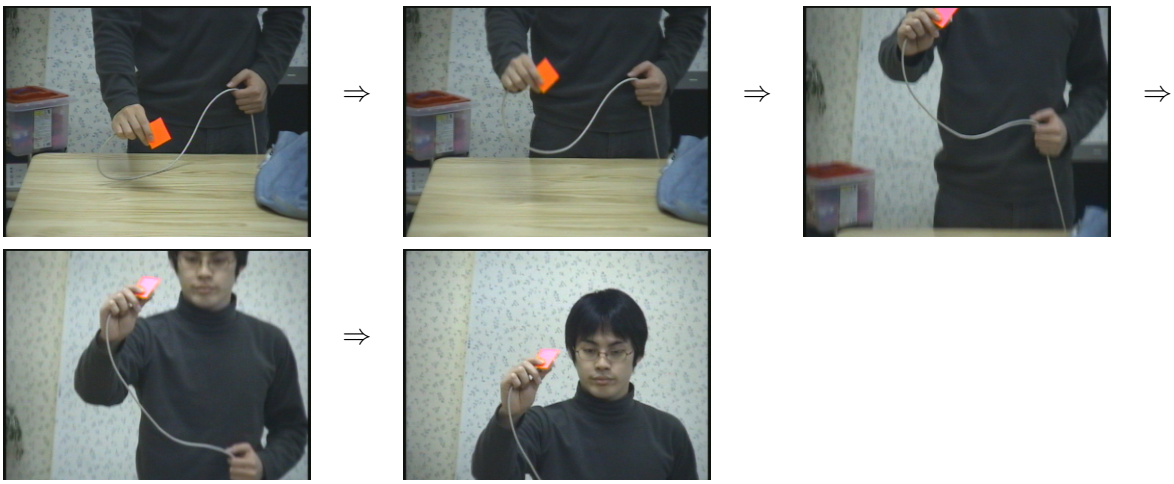


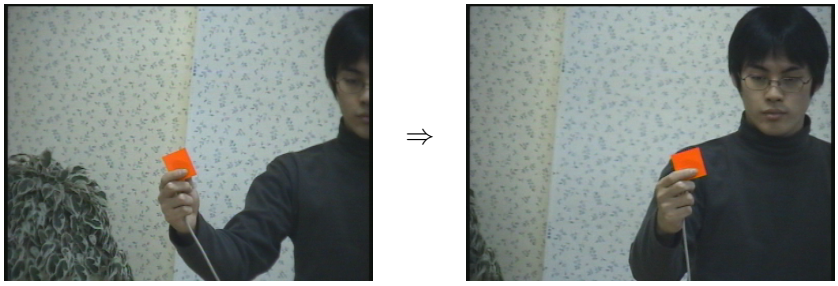
図 A.3: 提示動作でのカメラの遅れ

メラ制御遅れの影響を回避することができる．また，注目喚起行動に基づいてショットを選ぶため，遅れを含んだ手の移動中のショットに切り替えてしまう危険性も低い．このような工夫により，素早く不規則に動く対象を上手く映像に捉えることができる．

A.2.3 音声認識の遅れと行動検出

本システムでは，音声認識ソフトとして IBM ViaVoice8 Pro を用いており，発声から認識までに 2 秒弱ほど遅れる．しかし，映像の文法の観点からは，アクション（本研究では注目喚起行動）から少し遅れてショットが切り替わるのは，典型的な切替えタイミングの一つであり問題とはならない．4 章 4.5 節の「共起とみなす範囲の検討」では，注目喚起行動の生起時刻に対して前後何秒で切り替えた映像が良いかを視聴者に選んでもらう実験を行なったが，アクションから 1 秒後が最も良い結果となっている．このように，音声認識の多少の遅れは映像としては逆に良い方向に働いているといえる．

しかし，話者がショットを選ぶユーザインタフェースとしては，音声認識の遅れの間，注目喚起行動が正しく検出されたかどうか不安になるというアンケート結果が得られている．よって，この観点からは，音声認識はできるだけ遅れがないことが望ましい．この音声認識の遅れの問題は，プ



移動前 移動中

図 A.4: ゆっくりした移動でのカメラの遅れ

ログラムソースが公開されている音声認識エンジン（京都大学の Julius など）を本システム用にカスタマイズするなどの方法が考えられる．

以上のように，映像としては多少切替えが遅れた方が良く，ユーザインタフェースとしてはできるだけ速く検出結果を知りたいという相反した要求がある．これに対しては，例えば，音声認識（行動検出）の結果はライトの点灯などの手段で話者に示し，ショット切替えは 1 秒ほど遅らせて行なうことが考えられる．

付 録B　カルマンフィルタと一対比較法について

B.1　カルマンフィルタ

3 章 3.4 節で述べているように，可変枠制御では，対象の位置計測過程で生じるノイズを除去するためにカルマンフィルタを利用している．更に，カルマンフィルタに与えるノイズ分散比を意図的に調整することで，追跡モード時に対象をどの程度忠実に画面中央に捉えるかの度合（平滑化度合）を設定する．以下，本手法で使用しているカルマンフィルタの計算方法及びノイズ分散比について述べる [40][41].

本手法では，対象物の動きを剛体の運動モデル（等速度運動）で近似し，速度の変化をシステムノイズとして入力する．以下にダイナミクスモデルの式を示す.

$$\mathbf{x}_{k+1} = F\mathbf{x}_k + Cu$$

$$textbf{x}_k = \begin{pmatrix} x \\ \dot{x} \end{pmatrix} \quad \mathbf{F} = \begin{pmatrix} 1 & \Delta \\ 0 & 1 \end{pmatrix} \quad C = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

ここで，u はシステムノイズ，Δ は計測のサンプリング間隔，$\mathbf{x}_k$ は対象の現在の位置と速度を含んだ状態ベクトルである．

カルマンフィルタは以下の式で構成される．

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + K_k(y_k - H\hat{\mathbf{x}}_{k|k-1}) \tag{B.1}$$

$$\hat{\mathbf{x}}_{k+1|k} = F\hat{\mathbf{x}}_{k|k} \tag{B.2}$$

$$K_k = P_{k|k-1}H^T(I + HP_{k|k-1}H^T)^{-1} \tag{B.3}$$

$$P_{k|k} = P_{k|k-1} - K_kHP_{k|k-1} \tag{B.4}$$

$$P_{k+1|k} = FP_{k|k}F^T + \frac{\sigma_u^2}{\sigma_w^2}\Lambda \tag{B.5}$$

$$H = (\,1, 0, 0\,) \quad \Lambda = \mathrm{diag}\{0, 0, 1\}$$

$\hat{\mathbf{x}}_k$ は対象の推定状態であり，σ_u^2 はシステムノイズ，σ_w^2 は観測ノイズ，H は観測行列である．観測値 y_k が計測される毎に，式 (B.2)〜(B.5) の計算によって中間推定値 $\hat{\mathbf{x}}_{k|k-1}$ とカルマンゲイン K_k を更新し，式 (B.1) によって推定値 $\hat{\mathbf{x}}_{k|k}$ を計算する．$P_{k|k-1}$ は現在の観測値が得られた時点での誤差の共分散値であり，これが最小となるように推定値が計算される．

カルマンフィルタを上式で表すとき，その性質は σ_u^2 と σ_w^2 の比で決まる．この σ_u^2/σ_w^2 をノイズ分散比といい，追跡のスムーズさを表すカメラ制御パラメータとして利用する．ノイズ共分散比が小さいほどスムーズに追跡し，大きいほど対象物を忠実に追跡する．

B.2 サーストンの一対比較法

3 章の 3.5 節の評価実験で使用したサーストンの一対比較法について説明する [42][43].

一対比較法では，複数個の測定対象から総当たりで作られた刺激対を被験者に提示し，被験者はある判断基準（明るさ・大きさ・美しさなど）によってどちらかの対象を選択する．その結果をサーストンの比較判断の法則で処理することにより，測定対象の間隔尺度を構成する，つまりどのくらい差があるかを決めることができる．

ある刺激 S_j と S_k が被験者に提示されたとき，その反応連続体上の分布は，それぞれ平均 $\bar{R}_j$・標準偏差 σ_j の正規分布と平均 $\bar{R}_k$・標準偏差 σ_k の正規分布に従う．ここで二つの刺激 S_j が S_k の大小判断を行なった場合，二つの分布が少しでも重なる限り，どちらかが 100 %大きいとはいえない．

そこで，R_j と R_k の差 $(R_j - R_k)$ の分布を考え，その平均値の差 $\bar{R}_{jk} = \bar{R}_j - \bar{R}_k$ を二刺激間の距離とする．分布 $(R_j - R_k)$ の標準偏差 σ_{jk} は以下のように求められる．

$$\sigma_{jk} = \sqrt{\sigma_j^2 - 2r_{jk}\sigma_j\sigma_k + \sigma_k^2}$$

$$\text{ただし, } r_{jk} = \frac{\sum(\bar{R}_j - R_{ji})(\bar{R}_k - R_{ki})}{n\sigma_j\sigma_k}$$

ここで，n は被験者数，i は被験者のカウンタ，r_{jk} は R_j と R_k の相関係数をそれぞれ表す．

分布 $(R_j - R_k)$ の面積で $R_j > R_k$ になる確率と $R_j < R_k$ になる確率を分岐する点を 0 点とし，そこから分布の中心までの距離を σ_{jk} を単位として z_{jk} と表現すると，

$$z_{jk} = frac\bar{R}_j - \bar{R}_k\sigma_{jk}$$

となる．従って，σ_{jk} と上記の式から，反応連続体上の距離 $\bar{R}_{jk}$ は，

$$\bar{R}_{jk} = z_{jk}\sqrt{\sigma_j^2 - 2r_{jk}\sigma_j\sigma_k + \sigma_k^2}$$

と表現できる．z_{jk} は実験結果から計算される $R_j > R_k$ の割合と正規分布表から求まるため，残りの σ_j，σ_k，r_{jk} の値を決めれば，尺度値 $\bar{R}_{jk}$ が求まる．

この σ_j，σ_k，r_{jk} の求め方としてサーストンは五つのケースに分類しているが，ケース V という方法がその簡便さから最もよく用いられる．ケース V では，刺激対の反応連続体上の分布 R_j と R_k が無相関（$r_{jk} = 0$）かつ等分散（$\sigma_j = \sigma_k$）を仮定する．この場合，尺度値 $\bar{R}_{jk}$ を求める計算は，

$$\bar{R}_{jk} = \sqrt{2}z_{jk}\sigma$$

となり，更に $\sqrt{2}\sigma$ を単位とすれば，

$$\bar{R}_{jk} = z_{jk}$$

と簡略化される．

よって，ケース V でサーストンの一対比較法を用いる場合，各刺激対の大小判定回数を比較回数で割った値をもとに正規分布表から z_{jk} を求め，刺激毎にその平均値を計算することで尺度値 $\bar{R}_{jk}$ が得られる．

付 録C　本システムで取得した映像

本システムで取得した映像の例を挙げる．

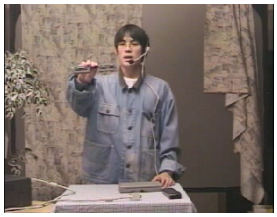
1. ［作業］ノートパソコンへの IO アダプタの取付け
2. ［作業］車の模型の組立て 前半・後半（ユーザインタフェースの評価実験で被験者にプレゼンテーションしてもらった内容）
3. ［模擬料理］キャベツとチーズのオーブン焼き（編集結果の評価実験で使用）
4. ［模擬料理］豚肉と竹の子の炒めもの料理（編集結果の評価実験で使用）
5. ［科学実験］アルミホイルと炭で電池を作る 前半・後半（編集結果の評価実験で使用）
6. ［工作］紙で作ったブーメラン（編集結果の評価実験で使用）
7. ［工作］アルミホイルの筒で作ったモノレール（編集結果の評価実験で使用）
8. ［工作］封筒で作った空飛ぶこいのぼり


1. それではノートＰＣにプロジェクタを取り付けます.


2. このノートＰＣには,このように,


3. ディスプレイを取り付けるコネクタがついていません.

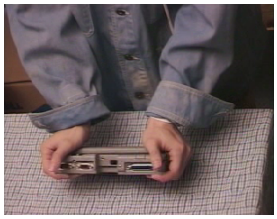

4. そこで,


5. このＩＯアダプタを取り付けます.

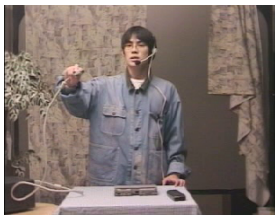

6. ここにディスプレイを取り付けるコネクタがあります.


7. このように,


8. ＩＯアダプタの上からノートＰＣを重ねて,


9. カチッと音がするまでしっかりと取り付けてください.


10. 次に,


11. このディスプレイケーブルを取り付けます.


12.


13. 左から二番目のコネクタに取り付けて,


14. こうやって,


15. しっかりとネジを締めてください.


16. これで完成です.

図 C.1: ［作業］ノートパソコンへの IO アダプタの取付け／創作


1. それでは車の模型の組立て方について説明します.


2. この車の組立てには,


3. このパワーツールという道具を使います. …

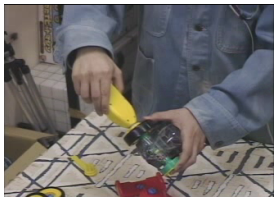

4. それではボディにタイヤを付けていきます.


5. こちらの前タイヤを付けていきます.

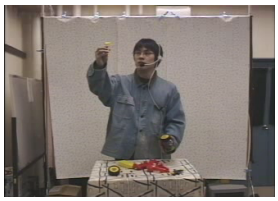

6. 表面に模様がついていることがわかると思います.


7. これが前タイヤです.


8. このようにして, 前のほうに被せて


9. パワーツールで絞めていきます. カチカチと音がするまで…


10. 取り付けが終わりましたら,


11. こちらの手絞めのレンチで,


12. 更に増し締めをしていきます.


13. では後ろタイヤをつけていきましょう.


14. こちらが後ろタイヤですが,


15. 表面に凹凸がないのがわかると思います.


16. このように後ろ側の穴に取り付けて,

図 C.2: ［作業］車の模型の組立て・前半／創作


17. パワーツールで取り付けていきます.


18.


19. 次に車にボディを取り付けていきます.


20. これが車のボディです.


21. 真ん中にボルトが一つ付いています.


22. これをこのように,


23. 車の真ん中のところに取り付けて…


24. 次に,


25. このウイングと


26.


27. こちらのバンパーを


28. 取り付けていきましょう.


29. 後ろの穴にウイングを, 前の穴にバンパーを…


30. 同じように締めていきます.


31. これで車が完成しました.


32.

図 C.3: ［作業］車の模型の組立て・後半／創作


1. はい，できました.（あとはこれを）


2. 上に乗せます．このときは混ぜないでください.


3. …もうこんなに小さくなりました.


4. （ぐっと量が減りましたね）


5. これを皿の上に


6. 乗せていきます.


7.（もうこのお皿に載るぐらいまで量が減って，…）


8. （あ，先生，美味しそうなスープも出てきていますね）


9. そうなんです，このスープも贅沢なほど


10. 美味しいスープなんです.


11. （このしんなりした葉はとろけるようですし…）


12. 二度手間だと思うけど，この一手間が


13. 本当に美味しくていいですね.


14.（蒸し汁もキャベツのスープも全部加えまして）


15. …みんなが大好きなチーズもたっぷりかけて…


16. （…オーブンで 10 分，こんがり焼けたら出来上がりです.）

図 C.4: ［模擬料理］キャベツとチーズのオーブン焼き／キューピー 3 分クッキング（括弧内はアシスタントの台詞）


1. （早速作っていくんですけど，豚肉ですね.）


2. 今日は豚肉を用意しました．…これにお酒を少し


3. 振りかけていきますよ．…


4. これは豚肉をやわらかくするんですよ．


5. こうしてしばらく


6. 置いておきます．


7. じゃあ炒めてまいりましょうね．


8. まずはフライパンを温めて，油を少し入れます．


9. ここに豚肉を入れて，


10. 炒めてまいりますね．


11. （豚肉ですから，完全に火を通さないといけませんね）


12. まず，こうして強火で炒めてまいりますね．


13. 色が変わることを目安になさるといいですよ．…


14. そしたら,（竹の子ですね）はい，竹の子．


15. いま美味しい竹の子でしょう．こうして，


16. 入れますね．

図 C.5: ［模擬料理］豚肉と竹の子の炒めもの料理／今日の料理（括弧内はアシスタントの台詞）


1. 台所にあるもので電池が作れますから，やってみましょう.


2. …台所にあるものという前提ですから，これは？


3. アルミホイルですよね.（はい）これが二枚要ります.


4. あと，これはクッキングペーパーですが，紙を一枚用意します.


5. そして，一番大事なのが，


6. これです.


7. これ何だかわかりますか？（冷蔵庫にある臭い取りですよね？）


8. そうです，脱臭剤です. で，この中身を開けてみます.


9. こんな.


10. これ，開けたことあります？（ないです）


11. …この炭，活性炭を使ってこれから電池を作ります.


12. （これをどうしたら炭になるんでしょうか?）


13. まずアルミを敷いてください.


14. この上に，紙を一枚乗せてください. 後で電球を繋ぎますから，…


15. さっき話にあった電気を流す水ですね.

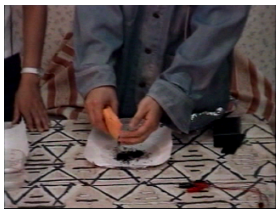

16. 今の場合は，台所にある一番ありふれた食塩水，

図 C.6: ［科学実験］炭とアルミホイルで電池を作る・前半／やってみようなんでも実験（括弧内はアシスタントの台詞）


17. これを紙の上に．染み渡る程度でいいです．


18. そうですね，そんなものでいいです．


19. それで，この上に炭を，

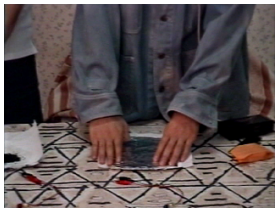

20. 脱臭剤の炭を撒いてください．薄く撒けばいいです．


21. （これでよろしいでしょうか？）そう，そのくらいで．…


22. では本当に電池になっているか確かめてみましょう．…


23. ただこれ，電池なんですが，


24. 炭がバラバラでコードが繋げませんね．


25. だから，炭から電気を集めるためにアルミを乗せます．…


26. （先生これでいいですか？）いいです．


27. そして，下のアルミから線を一本．…


28. （先生の光っていませんね）


29. これは接触が悪いんです．


30. 下の炭と上のアルミの接触が悪いわけです．


31. だから，ちょっとこう．（あ，光ってますね）


32. …こんな電球なら何時間でも光ってますね．

図 C.7: ［科学実験］炭とアルミホイルで電池を作る・後半／やってみようなんでも実験（括弧内はアシスタントの台詞）


1. 簡単な紙のブーメランを作ってみたいと思います.


2. まず，こういう厚紙，


3. …極厚の紙を用意します.


4.


5. ここのところを，ハサミで


6. （真ん中の部分ですね）１～２ｃｍくらいの筋を入れて，


7.


8. このように三枚用意します.


9. 三枚用意したら，


10. その切れ目同士を合わせるように入れてください.


11. …ちょうどこう，Ｙの字を見るように三等分…


12. そして，


13. このステイプラで，まあ糊でもいいんですけど，


14. 簡単にやるときはこういうものを使います.


15. 一応，これだけで出来上がりなんです.


16. では，飛ばしてみましょう.

図 C.8: ［工作］紙で作ったブーメラン／やってみようなんでも実験（括弧内はアシスタントの台詞）

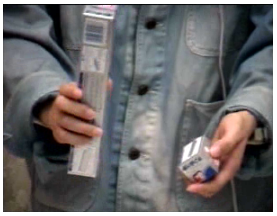

1. （この輪っかで何するの？）


2. これは，こんな電車が


3. 走ります.（わぁ，モノレールだ）


4. ガタンガタン…


5. これはね，こんな細長い箱の先っぽを


6. こういう風に切ります.


7. それでね，


8. この，横のところに


9. レールが入る溝を両方作って，


10. ここに乗せるんだ.


11. そして，じゃん. ガタンゴトン.


12. あ，そうだ. このレールはね，


13. こうやって輪っかで作るんだけども，そのときに，


14. こう離れたりズレたりしないで平らにしてください.


15. （そうか，平らじゃないと引っかかちゃうんだね）


16. じゃあ，このモノレールで遊ぼう！

図 C.9: ［工作］アルミホイルの筒で作ったモノレール／つくってあそぼ（括弧内はアシスタントの台詞）


1. あ，飛んだ.（飛んだ）


2. そうだ，空飛ぶこいのぼりを作ってみよう.


3. 材料はね，これでいきましょう.


4. このレジの袋と，ラップの芯のような紙の筒を使おう.


5. まずはレジの袋のこの手のところ，ここは要りません.


6. まずはここを切っちゃってください.


7.


8. で，切ったここに紙の筒をちょっと差し込んで，


9. セロハンテープで止めます.


10.（空気が漏れないように貼るんだよね）そのとおり.


11. しっかりと貼りましょう. こんな風にね！


12.（これで筒を吹いて袋を膨らませるんだ）


13. そのとおり.


14. （そしたら封筒こいのぼりを筒に差して,）そうそう…


15.（こうやって，スポンと！） あぁ！


16. 自分だけいいところやっちゃうんだから！

図 C.10: ［工作］封筒で作った空飛ぶこいのぼり／つくってあそぼ（括弧内はアシスタントの台詞）