講義の自動撮影系における音声・映像インデキシング

石塚 健太郎† 亀田 能成‡ 美濃 導彦‡

†京都大学大学院情報学研究科
‡京都大学総合情報メディアセンター
〒606-8501 京都市 左京区 吉田本町
京都大学総合情報メディアセンター開発支援部門 美濃研究室

あらまし

本稿では、事後利用のために講義を記録する「講義のアーカイブ化」のために、講義の自動撮影系において収録された講義映像に対し、自動的に講義映像を区切りインデクスを付与する手法について述べる。我々は従来、講義室内で発生する状況を「動的状況」により表現し、それに基づき講義を自動撮影する研究を行ってきた。この自動撮影系においては、動的状況に基づく撮影が行われているため、収録された講義映像に対し、自動的に動的状況の名称をインデクスとして付与することが可能となる。そこで本稿では、自動撮影系において用いる動的状況として被写体の位置および音声レベルに基づく動的状況を導入することで、視覚的・音声的に適切なインデクスを講義映像に付与する手法を提案する。また、本手法を応用することにより、講義映像を含む複数メディアの自動アーカイブ化が可能となるので、これについても述べる。

キーワード 講義収録、アーカイブ、マルチメディア、映像インデキシング、ハイパーテキスト

Speech and Video Indexing on Automatic Lecture Recording System

Kentaro Ishizuka† **Yoshinari Kameda**‡ **Michihiko Minoh**‡

†Graduate School of Informatics
‡Center for Information and Multimedia Studies
Kyoto University, Kyoto 606-8501, Japan

Abstract

We propose a method for speech and video indexing on automatic lecture recording system for archiving lectures. We previously proposed an automatic imaging method for lectures based on "Dynamic Situation", which describes the situation in the lecture room. The imaging method can segment and index the video image based on the dynamic situation. The dynamic situation is estimated by the location and the speech volume of a lecturer and students. We apply the indexing method to the automatic archiving system for lectures.

key words Lecture Recording, Archive, Multimedia, Video Indexing, Hypertext

1 はじめに

事後利用のために講義を記録しておく「講義のアーカイブ化」を考える。このアーカイブ化において、講義の映像を利用できれば有用である。講義映像を事後利用する際には、利用者が映像を全て見ることなく、必要とする映像のみを見られることが望ましい。

このような、映像を計算機上で有効利用するための研究として、映像インデキシングや映像記述の研究が近年数多く行われているが[1][2][3][4]、現在これらの研究は主に編集済みの映像を対象としていることが多い。編集済みの映像からインデクスの抽出を行う際には、撮影対象の状況やそれを撮影する際のカメラワークを獲得することが有用とされている。

これに対し、我々はこれまで研究を行ってきた講義を対象とする自動撮影系を活用し、講義映像の取得とそれに対するインデクス付与を同時に行うことで、状況に即した映像とインデクスを共に獲得する。

特に講義は被写体の位置と音声によってその動的状況が規定できるので、視覚的・音声的情報に基づいて動的状況を推定する。この動的状況を利用して、映像とインデクスを同時に獲得する手法について述べる。

このようにして得られたインデクスが付与された映像は、アーカイブ化において有効である。これは、状況に即したインデクスは区切り情報として正確であり、他メディアからの区切り情報とも整合を取りやすいことによる。

そこで本研究では、本手法をアーカイブ化に適用し、テキスト・音声・映像の 3 種のメディアを構造化しハイパーテキストの自動生成を行った。

以下、2 節で対象とする講義形態について述べ、3 節で自動撮影系について述べ、4 節で自動撮影系における映像への区切りとインデクス付与について述べ、5 節で講義のアーカイブ化について述べる。6 節で自動撮影系を用いたアーカイブ化の実験について示し、7 節で結論を述べる。

2　対象とする講義形態

講義の形態にはさまざまなものがあるが、本研究でのアーカイブ化の対象とする講義形態を、

(1)　1 人の講師と 1 人以上の生徒によって講義が進行する
(2)　講師が事前に作成したオンラインスライドをブラウザによりスクリーンに提示する
(3)　スライドの提示・切り替えには音声操作プロジェクタ[6]を用い、講師が音声コマンドにより操作を行う
(4)　講師は一定領域内を自由に移動できる
(5)　生徒は着席したまま受講する

の 5 つの条件を満たすものとする。

3　講義のインデクス自動生成撮影系

本節では、インデクスと映像を同時生成する手法として、講義室内の動的状況に基づく自動撮影系について述べる。

3.1 動的状況に基づくインデクス生成と自動映像化

講義の撮影を行う際には、「どの被写体を、どのように撮影するか」を知る必要がある。これを知るために、講義室内に発生している状況を「動的状況」により記述する。この動的状況に対し一意にカメラワークを割り当てることによって、「どの被写体をどのように撮影すべきか」を表現する。例えば、「講師が教卓の前にいるときは、講師を撮影する」のような記述となる。このような表現形式のことを、「撮影ルール」と呼ぶ。

このような撮影を実行するためには、「どの種類の動的状況が発生したか」を知らなくてはならない。そこで、動的状況を一意に表現するのに十分な、講義室内の被写体の位置と音声レベルを用いる。これらを「状況特徴量」と呼ぶ。

なお、ここでいう「インデクス」とは、発生した動的状況の名称と、その撮影法を記述した撮影ルールのことを指す。

以上に従い、状況特徴量から動的状況を推定し、それを撮影ルールに対応付けることによって、講義のインデクス自動生成撮影を行うことができる。この手法における処理の手順を以下に示す。

1. 講義室内の状況特徴量を抽出
2. 状況特徴量を評価し、動的状況を推定
3. 動的状況から撮影ルールを選択
4. 撮影ルールに基づき被写体を撮影し、同時にインデクスを付加

3.2　動的状況の定義

動的状況の定義に被写体の 3 次元空間内の位置および音声レベルを用い、講義室内の動的状況を以下に述べる 2 種類の動的状況により表現する。

表 1:工学部 10 号館第 2 講義室における被写体の位置に基づく動的状況

動的状況	動物体	動物体の存在領域	領域内の静物体
教卓の前で講義中	講師	area 1	教卓
黒板を使って講義中	講師	area 2	黒板
前縁部で講義中	講師	area 3	なし
スクリーンを使って講義中	講師	area 4	スクリーン
生徒が活動中	生徒	area 5-15	なし

3.2.1 被写体の位置に基づく動的状況

自動撮影の対象とする講義形態においては、講義室内の動的状況は、講義を構成する動物体（講師と生徒）と静物体（教卓、スクリーンなど）の関係によって規定できる。

講義はその講義室内に存在する講義資料と講師、生徒をその構成要素とする。すなわち、講義で発生する動的状況はその動的状況を発生させた動物体とその位置する領域、およびそれが使用する静物体の位置関係によって規定できる。特に、動物体の存在する領域の情報はそれに対する撮影方法に大きな影響をおよぼす。

動物体である講師と静物体との関係は講師の位置と静物体との距離によって定まる。具体的には、講師の移動可能領域を静物体の位置に応じて複数領域に分割し、一つの領域には 0 または 1 つの静物体しか存在しないようにする。これにより、講師が特定の静物体が含まれる領域にいる場合は、その領域内の静物体との関係が強いと仮定する。この推定には、状況特徴量として講師の 3 次元空間内の位置を用いる。

一方、生徒についての動的状況は、質問などで生徒が活動的になっているかどうかが考慮されるべき要因となる。この推定には、状況特徴量として生徒の動きの度合いを表す生徒活性度を用いる。

ここで例として、本学工学部 10 号館第 2 講義室で行われる講義について定義した動的状況を表 1 に挙げる。領域は図 1 に示すように分割した。

図 1 中斜線部が講師の移動可能領域であり、講師の移動可能領域は area 1 から 4 までの 4 つに分割されている。各領域について、area 1 には教卓、area 2 には黒板、area 4 にはスクリーンが静物体として含まれている。area 3 には静物体は何も含まれていない。これにより、講師の存在する領域に応じて動的状況が規定される。

図 1 中網掛け部が生徒の存在領域であり、座席を area 5 から 15 までの 11 の領域に分割している。これら各領域の生徒について生徒活性度が高ければ、その領域内の生徒が何らかの動作をしているという動的状況であると規定する。

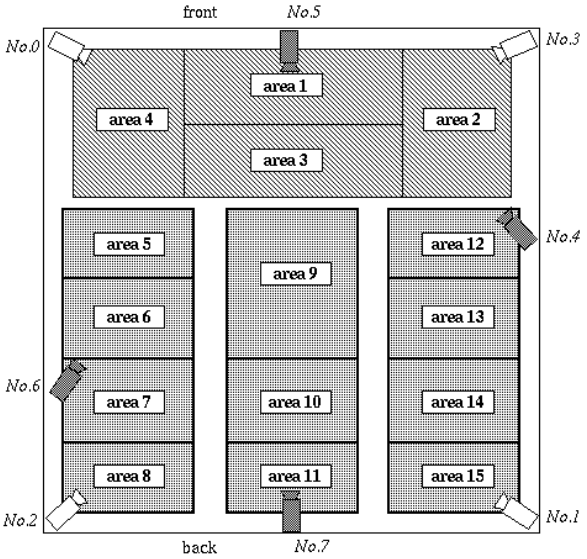

図 1:京都大学工学部 10 号館第 2 講義室における領域分割

3.2.2 被写体の音声レベルに基づく動的状況

講師の音声と撮影方法との協調を目的とし、映像化に用いる動的状況の定義に被写体の発話状況を導入する。

講演調の話し言葉に対する分析を行なった研究 [5]においては、講演調の発話が「話し言葉」と「書き言葉」の中間的性質を持つため、これに対する文法構築の単位としてポーズを用いることが考えられている。ポーズには機械的に検出できる音響的ポーズと、人間によって知覚される「間」に相当する知覚的ポーズがあり、この知覚的ポーズの間に発話される音声について、人間は意味的な単位を与えていると考えられる。しかし、知覚的ポーズは主観的なものであり個人差も大きいことと、知覚的ポーズが音響的ポーズの一種であることから、本研究では音響的ポーズに基づく状況特徴量により、被写体の音声レベルに基づく動的状況を推定する。

ここでは、講師が発話状態か否かを音声レベルによって判定する。講師の音声レベルに基づく動的状

況を表 2 に挙げる。音声レベルが高い状態にある区間を有音区間、一定時間以上低い状態にある区間を音響的ポーズとし、それに応じて動的状況を定義する。

表 2:被写体の音声レベルに基づく動的状況

動的状況	講師の音声レベル
発話状態	有音区間
非発話状態	音響的ポーズ

3.3　動的状況に基づく撮影カメラの自動制御と映像選択

講義室内を撮影するためのカメラを「撮影カメラ」と称する。撮影カメラとして、固定した位置に設置した首振りカメラまたは固定カメラを複数台利用する。

映像収録の際には、撮影に用いたカメラとそのカメラパラメータを、映像に対しインデクスとして付与する。

講義室の観測による動的状況の推定結果に基づき、撮影カメラの制御と映像選択を行う。ここでは、動的状況の発生した被写体の存在領域に基づき各撮影カメラの制御を行う方法と、その映像の中から最適なものを選択する方法について述べる。

撮影カメラに固定カメラないし首振りカメラを用いるという前提と、被写体の位置に基づく動的状況が被写体の存在領域に強く関係付けられていることから、それぞれの動的状況を特定の撮影カメラで撮影する場合、そのパン・チルト制御値は撮影カメラの空間内における位置によって定めることができる。このカメラ制御に関わる動的状況においては、音声レベルに基づく動的状況を考慮しない。

１つの動的状況に対する撮影ルールは、画角の問題を別に考えると、最大でその動的状況が発生している存在領域を撮影できる撮影カメラの数だけ用意できる。実際には、１つの動的状況に対応できる撮影カメラを、動的状況の存在領域とカメラ位置とから事前に決定しておく。

講義において、動的状況は同時に複数個発生するので、発生した１つ以上の動的状況に対して、どの撮影カメラを用いて撮影するべきかを選択決定する。この撮影カメラの選択決定のことを今後「映像選択」と称し、以下にその選択方法について述べる。

まず、複数発生しうる動的状況に対する選択基準を示す。これは映像化の目的に即してトップダウン的に決定する。

撮影するべき「被写体の位置に基づく動的状況」の選択基準

D1: ある動的状況に対する撮影を開始してから一定時間が経過していない場合は、その動的状況を継続して撮影

D2: 被写体の異なる動的状況が同時に発生した場合、事前に設定した優先度に基づきいずれかを優先

D3: 生徒を被写体とする動的状況が複数発生した場合、最も生徒活性度の高い動的状況を優先

D1 の基準を導入したのは、頻繁な映像選択の実行を抑制するためである。

以上の **D1** から **D3** の選択基準に基づき、撮影されるべき動的状況が決定された後、撮影カメラの選択を行う。その基準を以下に挙げる。

撮影カメラの選択基準：「被写体の位置に基づく動的状況」が変化した場合

Sa1: 現在利用している撮影カメラが利用可能ならば、再度選択

Sa2: そうでない場合は、新しい動的状況を撮影するのに最適な撮影カメラを選択する

撮影カメラの選択基準：「被写体の位置に基づく動的状況」が変化しない場合

Sb1: 一定時間以上、同一の撮影カメラで撮影していれば、他の撮影可能なカメラに変更

Sb1 を導入したのは、視聴者の映像の飽きを考慮したものである。

以上の基準に従い映像選択が行なわれるが、その実行の際には、被写体の音声レベルに基づく動的状況を考慮した、次の基準を導入する。

撮影カメラの選択基準：「被写体の音声レベルに基づく動的状況」が変化した場合

Sc1: 被写体の音声レベルに基づく動的状況の変化をトリガとし、映像選択を実行する。

Sc1 は、映像選択を音声と同期させることを目的とする。これにより、映像に音声的情報を付加できる。

4　映像・音声インデキシング

講義の映像を事後利用する際には、利用者が講義

映像を通して見ることなく、必要とする部分の映像のみを見られるようにすることが重要である。そのためには意味的情報に従い映像を区切り、インデクスを付与しておく必要がある。このような、連続メディアを区切るための意味的情報を以後「区切り情報」と称する。また、区切り情報により区切られた時間連続な映像を今後「カット」と称する。

インデクス自動生成撮影系においては、2 種類の動的状況を用いることで、映像と同時に区切り情報を獲得し、映像へのインデクス付与を行っている。

被写体の位置に基づく動的状況を、動物体とその位置する領域、およびそれが使用する静物体との位置関係によって規定した。被写体の位置に基づく動的状況は視覚的に規定されているため、カットに対応する動的状況の名称は、視覚的情報源に基づくインデクスとなる。このインデクスが変化する際は映像選択が実行されるため、これは映像の区切り情報として利用できる。

被写体の音声レベルに基づく動的状況は、音声的情報である講師音声の有音・ポーズに基づき規定している。ポーズに挟まれた一つの有音区間が一つの動的状況を表し、これは音声的情報源に基づくインデクスに対応する。このインデクスの変化によっても、映像を区切ることが可能である。

3.3 節に述べた映像選択基準においては、被写体の位置に基づく動的状況をもとに撮影カメラの選択を行う一方で、被写体の音声レベルに基づく動的状況の変化が映像選択実行のトリガとなるため、映像選択の実行が音響的ポーズの発生と時間的に同期し、映像選択実行が視覚的・音声的な区切りとなる。

この結果、一つのカットに対しては、被写体の位置に基づく動的状況の名称をインデクスとして付与する他に、有音区間において映像選択が実行されないことから、時間的に共起した音声を対応づけることができる。

以上のように、本稿で述べたインデクス自動生成撮影系を用いることで、映像を視覚的・音声的に区切ることができ、カットに対し視覚的・音声的なインデクスを映像獲得と同時に付与することができる。

5　講義の自動アーカイブ化

インデクス自動生成撮影系をアーカイブ化に適用し、複数メディアを対象とした自動アーカイブ化を行った。

この講義のアーカイブ化においては、映像・音声に対し、動的状況によるインデクスを付与するだけでなく、テキストメディアとの協調により、講義において使用された教材を映像や音声のインデクスとして利用する。

5.1　複数メディアを対象とした講義の自動アーカイブ化

本稿では講義を行う際に、講師が講義内容を生徒に伝達する際に行うプレゼンテーションのことを、「プレゼンテーション情報」と呼ぶ。講義のアーカイブ化においては、このプレゼンテーション情報を収録対象とする。

本研究における「アーカイブ化」とは、講義のプレゼンテーション情報を区切り、それに対しインデクスを付与し、映像・音声・テキストなどの複数メディアを構造化しハイパーテキスト化を行うことを指す。

ここでは、自動アーカイブ化システムの設計・構成と、それを用いた講義のプレゼンテーション情報の自動構造化について述べる。

講義のプレゼンテーション情報を扱うことを考えれば、講義に現れる有用な情報は全て記録しておく必要がある。そこで、講義におけるこれらの情報をすべて記録するためのシステムとして講義の自動アーカイブ化システムを設計・構成した。

講義に存在するプレゼンテーション情報伝達のためのメディアとして、2 節の講義形態のもとでは、

- 板書・OHP・教科書・オンラインスライドなどのテキストメディア
- 講師・生徒の発言などの音声メディア
- 講師・生徒の映像などの映像メディア

の 3 種のメディアが存在し、これらメディア上に現れるプレゼンテーション情報が自動アーカイブ化システムの収録対象となる。

最終的に生成されるハイパーテキストにおいては、これら 3 種のメディア上のプレゼンテーション情報が、利用者が参照しやすいように構造化されている必要がある。この構造化を行うため、本システムでは各メディア毎に処理を行い区切り情報を生成する。それぞれのメディアでは、プレゼンテーション情報を理解し予め定められたイベントを検出して、区切り情報を生成する。

この区切り情報を元に、各メディア上のプレゼンテーション情報を区切り、区切られたプレゼンテーション情報について複数メディア間での対応づけを

行うことで、自動構造化を行い、ハイパーテキストを生成する。

5.2 収録対象とするプレゼンテーション情報

2 節に述べた講義形態に従い、各メディアでのプレゼンテーション情報の収録対象を以下のように定める。

テキストメディアの収録対象

オンラインスライドのみとする。オンラインスライドは事前に講師によって作成され、また講師の意思で切り替えが行われることから、講義内容そのものを明示的に表していると言える。この意味でテキストメディアは他のメディアよりも重要な位置づけにある。

音声メディアの収録対象

講師音声のみとする。講師は講義の中心的存在であり、講義空間において最も多く講義内容に関わる発話を行うためである。また本研究で規定する講義形態において音声メディアはテキストメディアに対する解説であると位置づける。

映像メディアの収録対象

映像メディアとしては、動物体である講師・生徒を被写体とする映像を収録する。ここでは、オンラインスライドを用いた講義形態であることから、それが投影されている静物体であるスクリーンを映像メディアにより収録する必要はない。また、本研究での講義形態において映像メディアはテキストメディアの解説であると位置づける。

5.3 区切り情報の生成

各メディアにおいて講義を観察し、その状態変化の度に時刻と状態変化に関する情報を記録する。これを「区切り情報」とし、講義収録時に生成する。前述の講義形態に従い、各メディアでの区切り情報を以下のように生成する。

テキストメディア

オンラインスライドが切り替わった時点で生成

音声メディア

3 章に述べた音声レベルに基づく動的状況が変化した時点で生成

映像メディア

3 章および 4 章において述べた手法に基づきカット単位で区切り情報を生成

5.4 テキストメディアと音声メディアの対応

テキストメディアと音声メディアの対応づけを行なうために、音声操作プロジェクタを用いた講義音声とオンラインスライドの自動ハイパーテキスト化システム[6]を利用する。

この手法の利点は、講師が音声でプロジェクタへ命令を発することにより「スライドを切り替えながら発話する」といったことがなく音声とスライドの時間的な対応を完全に得ることができる点にある。また、映像を用いなくともスライドの切り替え時刻を得ることができる。

音声操作プロジェクタにおいては、操作コマンドを発話する際には前後に一定の無音区間（ポーズ）をとることが前提とされており、この結果、テキストメディアと音声メディアの区切り情報の発生が時間的に同期する。これによりテキストメディアと音声メディアの対応づけが可能となる。この結果、講演音声に対し、オンラインスライド 1 枚 1 枚がインデクスとして付与されることになる。

5.5 映像メディアと音声メディアの対応

映像メディアと音声メディアはともにテキストメディアの解説過程であるという位置づけから、互いに関係がある。すなわち、映像メディアと音声メディアについても区切り情報の時間的同期がとれることが望ましい。

本稿でのインデクス自動生成撮影系においては音声レベルに基づく動的状況を用いることで、音声メディアと映像メディアの区切り情報の同期をとることができる。音声レベルに基づく動的状況の変化は、前述した音声メディアにおける区切り情報の生成と等価である。従って、音声レベルに基づく動的状況を用いることで、映像メディアの区切り情報である映像選択の発生が音声メディアの区切り情報の発生と時間的に同期することになる。

これにより、映像の区切りをカット単位で行なうと、1 つのカットに対し、時間共起していた発話を対応づけることができる。

5.5 テキストメディアと映像メディアの対応

テキストメディアと映像メディアの対応づけについては、アーカイブ化においてテキストメディア

が最も重視されることから、テキストメディアの区切り情報が発生した時点で映像選択を実行し、映像メディアにおける区切り情報を発生させるようにする。ただし、映像選択を実行した時点で、被写体の位置に基づく動的状況が必ずしも変化しないことから、実際には映像が切り替わらない場合もある。

この結果、映像に対し、オンラインスライドをインデクスとして付与することができる。

5.6 講義の自動アーカイブ化システムの構成

自動アーカイブ化システムにおいては、講義映像・音声・スライドの 3 種のメディア上のプレゼンテーション情報を自動構造化し、テキストメディアを中心にハイパーテキスト化を行なう。

図 2 に講義の自動アーカイブ化システムの構成図を示す。入力される音声は音声レベルに応じて区切られ、音声操作プロジェクタへ送られる。音声操作プロジェクタは入力される音声を識別し、コマンド音声ならばスライド操作を行なう。自動撮影系は動的状況を観測により求め、それにより講師と生徒の映像を撮影する。テキストメディアの区切り情報と、2 種類の動的状況に応じて映像選択を行ない、

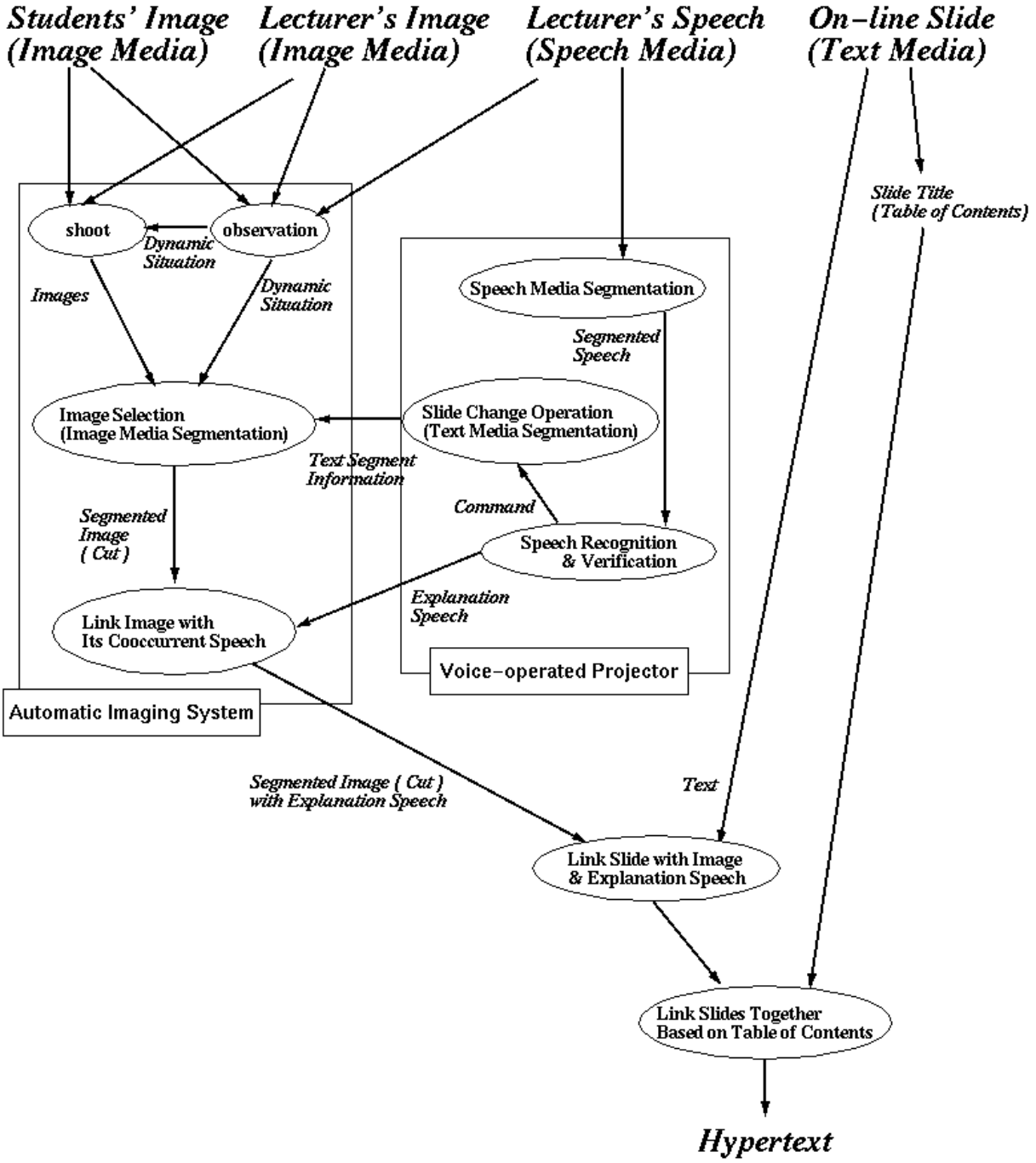

図 3:講義の自動アーカイブ化システム

映像メディアの区切り情報を生成し、カットに区切る。その後、カットに対し時間共起していた講演音声をリンクし、講演音声がリンクされた映像メディア（カット）とテキストメディア（オンラインスライド）を対応づける。最後に、テキストメディア間の構造をオンラインスライドの目次から求め、映像メディアがリンクされたスライドを構造化することにより、映像・音声・テキストの 3 種のメディアを自動構造化したハイパーテキストを生成する。

この結果生成されたハイパーテキストにおいては、映像・音声に対しオンラインスライドによるインデクスが付与されているため、オンラインスライド 1 枚単位で、それに対応する映像・音声を得ることができる。

また、映像については、インデクス自動生成撮影系により付与されたインデクスを元に、動的状況の名称を元に、特定の動的状況に相当する映像と音声を取得できる。

6　講義の自動アーカイブ化システムの実装および実験

前節で述べた講義の自動アーカイブ化システムを実装し、1999 年 11 月 30 日に本学で行なわれた大学院向けの講義を対象に実験を行った。

本講義において得られた区切り情報の数を表 3 に示す。また、この区切り情報を元にハイパーテキストを生成した（図 3）。

この結果、全講義映像を見ることなく、スライド単位でその解説音声と映像を取得することができ、有効な構造化ができたと考えられる。

今後、この自動アーカイブ化システムにより、多くの講義をハイパーテキスト化し、受講者の事後利用を通じて評価を行う必要がある。

7　結論

講義の自動アーカイブ化を行うために、インデクス自動生成撮影系を考案し、音声・映像インデキシングを行う手法を提案した。また、複数メディアを対象とした講義の自動アーカイブ化システムに提案手法を適用し、その実験を行った。今後は生成されたハイパーテキストに対する利用者の評価などを通じ、手法の有効性を検証する必要がある。

表 3:得られた区切り情報の数

メディア	区切り情報の数
テキストメディア	29
音声メディア	3810
映像メディア	223

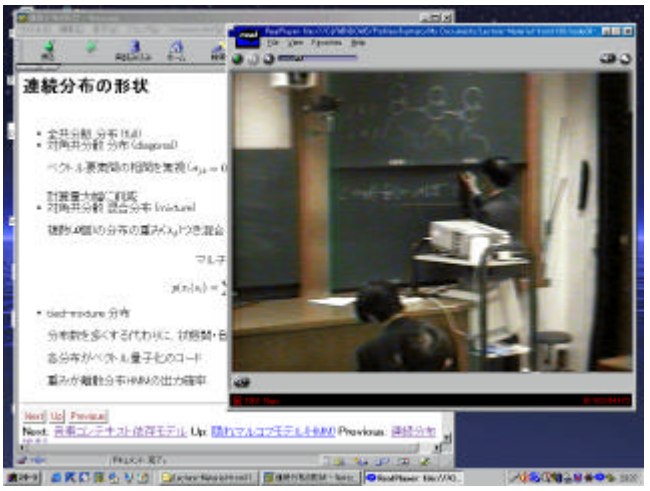

図 3:講義の自動アーカイブ化システムにより生成されたハイパーテキストの例

参考文献

[1] Davenport, G., Smith, T. A., Pincever, N.: Cinematic Primitives for Multimedia, IEEE Computer Graphics and Applications, Vol.11(4), pp.67-75, 1991.

[2] G.Hauptmann, A., A.Smith, M.: Text, Speech, and Vision for Video Segmentation: The Informedia Project, AAAI Fall Symposium, 1995.

[3] 柴田正啓: 映像の内容記述モデルとその映像構造化への応用，信学論，Vol.J78-D-II, No.5, pp.754-764, 1995.

[4] Satou, T., Akutsu, A., Tonomura, Y.: Video Corpus Construction and Analysis, IEEE ICMCS, pp.479-485,1999.

[5] 峯松信明，片岡嘉孝，中川聖一: 講演調の話し言葉に対する分析，情処研報，SLP-8, pp.39-46, 1995

[6] 河原達也，石塚健太郎，堂下修司: 発話検証に基づく音声操作プロジェクタとそれによる講演の自動ハイパーテキスト化，情処論，Vol.40, No.4, pp.1491-1498, 1999.