システムソフトウェアと
オペレーティング・システム 69－1
（１９９５．６．２２）

プロトコル解析のためのテキスト／ビデオ統合システム

高田敏弘　鷲坂光一　尾内理紀夫
NTT 基礎研究所

プロトコル解析の実験模様を収録したビデオテープと，その実験中の発話を書き起こしたテキストデータとを双方向にリンクしたプロトコル解析支援システムを，意味計算研究の応用の一つとして試作した．テキストからビデオへのリンクは，HTML でテキストを記述し，ハイパーテキスト化することにより実現した．また，テキストに対しては曖昧なキーワードをもとに全文検索することができる．ビデオからテキストへのリンクは，二種類が実現されている．一つは，ビデオ映像の時間軸上のある地点から対応するテキストへのリンクであり，もう一つは，ビデオ映像の空間領域から対応するテキストへのリンクである．後者の機能により，ビデオ映像中のある領域をクリックすると，その領域で発生したイベントに対応するテキストとビデオを表示することができる．今回はその一例として，映像中の碁盤上の碁石から，その碁石を打った時点の発話テキストへのリンクを実現した．なお，本システムのハードウェアは，市販のビデオ装置と計算機から構成されている．

Text/Video Linkage System for Protocol Analysis

Toshihiro Takada　Mitsukazu Washisaka　Rikio Onai
NTT Basic Research Laboratories

This paper describes the text/video linkage system for protocol analysis. Text data is hyper-structured using HTML, and a linkage from text to video is implemented using this hyper text structure. Also fuzzy retrieval using keywords for the text data is possible. There are two kinds of linkage methods from video to text. One is from video frames specified by timecode to text, the other is from a specific area in a video image to text. Using the latter method, when some area in the video image is clicked, the text data and video image corresponding to the area event are displayed on the monitor. This system is comprised of computer and video equipment currently on sale.

1 はじめに

我々は，「意味」[1]を意識したコンピューティングである意味計算（Semantic Computing）の研究を展開している [1]．このような研究の方向を模索しようと思い立ったのは，二つの技術的背景がある．一つはWWW(World Wide Web)に代表される分散化された巨大データベースの出現であり，もう一つは急速な進展を見せるマルチメディア技術 [2] である．

例えば，WWWでは，既存の検索サービスを利用しても，完全一致するテキストデータがあればよいが[2]，完全一致するデータがない場合には目的地に行きつくことは難しい．テキストデータに対してもそうであるから，質問をすること自体が難しい画像や音の入ったマルチメディアデータベースに対しては検索はさらに困難なものとなる．即ち，マルチメディアデータベースへの検索では，完全一致は無力であり，意味を理解した上での言い換えや説明が可能でなければ所望のデータを取り出すことは難しい．分散化された巨大マルチメディアデータベースが台頭してきた時には，それに対する意味計算，意味検索は，“Information on Demand” の実現のために，きわめて重要となる．

我々は，このような意味検索研究を展開する上でのいくつかの応用例の一つとして，プロトコル解析を選んだ．プロトコル解析については次節で説明するが [3]，これを選んだ理由は，以下のとおりである．

- 閉鎖系である．即ち，データ収集中の実験室に突然，予期せぬ人が入ってくる可能性がある開放系に比べ，画像 / 音声認識処理が簡単になる．
- テープ起こしされたテキストデータが存在する．現在の画像 / 音声認識技術の現状では，マルチメディアデータベースに対して，テキストの手がかりなしに多様な意味検索をかけることは難しい．
- 完成し，それが快適なシステムならば，使用する人が多数いる．これは好み，立場の問題かもしれないが，我々は多くの人々に使われることを良しとする．

本稿で述べるシステムは，意味検索へ向かう第一歩であり，このシステムを利用することにより意味検索の研究を展開していく予定である．

[1]「意味」とは，それを理解していれば言い換えや説明ができるもの

[2]実際はハイパーテキスト構造の中で堂々巡りをしてしまい，適切なガイド機構があればと思ってしまうことも多い．つまり，WWWに対する検索ガイド機構の研究も重要である．

2 プロトコル解析とは

プロトコル解析とは，心理学，認知科学等の研究手法の一つであり，その応用には，ユーザによる工業製品評価も含まれる．

プロトコル解析のためには，まず，解析のためのプロトコルデータを収集する．通常は実験室があり，そこには，被験者（一人あるいは複数）と場合によっては，実験を主催する人（実験者）がいる．そして，実験者から被験者に，実験課題が説明され，時には練習をし，実験に入る．

実験では，課題（例えば，パズル）を実行しながら，課題に関連して，その時，考えていることをできるだけ声に出して語ってもらう．この時，被験者が，自発的に自然に発話してくれることが重要である．そして，それをオーディオテープ / ビデオテープに記録する．ビデオテープを使用するのは，主に発話以外の非言語行動（ジェスチャ等）を記録するためである．

次に記録テープを元に，解析可能な形のデータへと「書き起こし」をする．通常は，発話の言語データ化（テープ起こし）である．この作業のコストが高いことがこの手法の問題の一つであると言われており，オーディオテープからの書き起こしや，記録テープやデータへの検索を容易にするための効率的ツールが開発されている．

そして，書き起こされたデータが解析の対象となる．書き起こされたデータをトレースし，被験者の頭の中の動きを追い，頭の中の様子に関する種々の推論を行なうが，データ解析にコストがかかることも問題となっている．ここで，書き起こされたデータとは，音声を元にした言語データだけでなく，課題を実行中の指差し，首振り等の非言語行動もデータとなりうる．

実験はビデオテープにも記録されているが，これのデータ化はオーディオテープからの書き起こしに比べてはるかにコストがかかる．よって，例えば，被験者の動作の内，指定された非言語行動を自動的に書き起こすことができ（被験者が指差しをした時のビデオフレーム番号等），さらに，指定された非言語行動 / 言語行動をビデオテープ / オーディオテープから自動的に検索できれば，プロトコル解析手法のコスト低下と高度化に寄与すること大である．また，言語データとビデオテープの間にリンクをはり，ある発話をした時の被験者の様子を即座に見たり，ある動作をした時の対話データを即座に表示することができるだけでも便利である．

我々の目指すプロトコル解析システムはこのような支援システムであり，本報告はその第一歩である．

ただ，我々の研究においては，プロトコル解析法の妥当性，即ち，方法論として，プロトコル解析が頭の中の処理プロセスを明らかにする上で有効かどうかということは，スコープ外であることを付け加えておく．

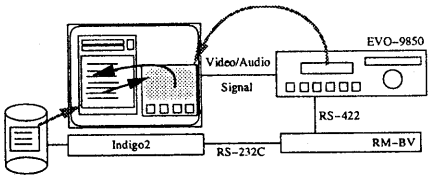

図 1: ハードウェア構成

3 システムの全体構成

はじめに今回試作したシステムのハードウェアとソフトウェアの構成を説明する.

3.1 ハードウェア構成

現在のシステムのハードウェア構成を図 1に示す. ビデオ側の機材には 8mm ビデオデッキ (EVO-9850) とそれを RS-232C 経由で制御するためのプロトコル変換装置 (RM-BV) を用い, 計算機側には SGI 社の Indigo2 を使用している. ビデオ機材と計算機との間はビデオ / 音声信号入力とビデオ制御用信号 (RS-232C) の 2 系統で接続されている. 今回使用したビデオデッキ (およびプロトコル変換装置) は, RS-232C 経由でタイムコード (ビデオテープ上の時分秒とフレーム番号) によって指定された時点へ自動的にテープを送り, そこから再生を開始するなどのコントロールが可能なものである.

3.2 ソフトウェア構成

計算機上に構築されたソフトウェアの構成は図 2の通りである. これらのソフトウェアの役割を以下簡単に説明する.

テキストブラウザ: テキストの表示を行う. 現在は既存の WWW ブラウザ (Mosaic) をテキストブラウザとして使用している.

ビデオブラウザ: ビデオ映像の表示, および, 再生・停止・早送り・巻戻しなどの ビデオデッキを操作するための GUI, 後述するビデオからテキストへのリンクを辿る機能などを実現している.

制御用デーモン: テキストブラウザ／ビデオブラウザ間の情報交換およびビデオデッキの制御を行う.

HTTP サーバ: テキストは HTML を用いてハイパーテキスト化されてディスクに納められている. テキストブラウザは HTTP サーバを経由してこれらのテキストにアクセスする.

WAIS サーバ: テキストの検索を実現する.

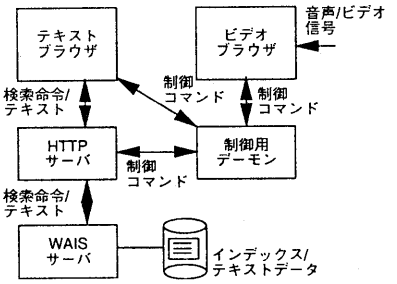

図 2: ソフトウェア構成

4 テキストの検索

プロトコル解析では, 解析の対象となる事柄に関連してどのような発話が行われたかを, テープ起こしされたテキストから検索して解析を進める必要がある. しかし, テープ起こしされたテキストは話し言葉であるため, 完全一致型の検索では目的の場所を見つけることは非常に難しい. このため本システムでは, 書き言葉と話し言葉の違いをある程度吸収でき, あいまな検索文からも目的の場所を見つけることができるように検索サーバとして日本語 WAIS を使用している.

日本語 WAIS は, WAIS Inc. が開発した全文検索システム WAIS (Wide Area Information Servers)[4] に, 京都大学が開発した形態素解析プログラム juman[5] を組み込んだもので, 日本語テキストを全文検索することができる. 日本語 WAIS では, 形態素解析を行い切り出した各単語を終止形に変換した後に, テキスト中にその単語が出現する頻度を調べて検索インデクスを作成する. そして, 検索文を単語に分割し, 検索文中の単語がテキスト中にどの程度含まれているかをスコアで表すことができるため, ある程度あいまいな検索文を指定しても, 目的とするテキストを得ることができる.

日本語 WAIS の検索結果はファイル単位でしか得られないため, 検索単位となるファイルがあまりにも大きいと, 使いにくいものになってしまう. このため, 本システムでは, テープ起こし時に日本語テキストに埋め込まれたタイムコードをもちいて, あらかじめテキストをタイムコードで挟まれた小区間ごとにファイルに分割した後に, 検索インデクスを作成した.

実際の検索は, GUI に WWW ブラウザの FORM 形式をもちいており, 検索文と検索する範囲を指定し, WAIS サーバを呼び出して行う. 検索の結果, WAIS サーバから返されるスコアをもとに, 検索文に最も一致するタイムコードの区間から順に, テキストとその区間の要旨をあらわす文章がリストされる. リストされた項目には対応するテキストへのリンクが張られており, 項目をク

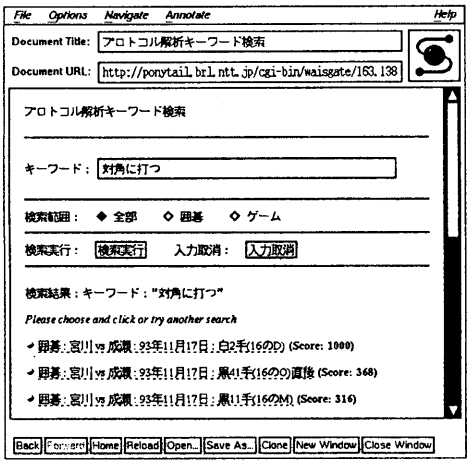

図 3: テキストの検索

リックすることで，そのテキスト部分を表示することができる．「対角に打つ」で検索した例を図 3に示す．

5 テキストからビデオへのリンク

5.1 テキストのハイパーテキスト化

2節で述べたように，プロトコル解析に用いられるビデオテープの発話部分は，あらかじめ書き起こされてテキストデータとして存在する．今回はこれらのテキストを HTML (HyperText Markup Language) を用いてハイパーテキスト化することにより，テキストからビデオへのリンクを実現した．

テキストからビデオのリンク付けにあたっては，ビデオテープ上のタイムコードを軸として採用した．すなわちテキストの要所要所からは，そのテキストに書かれている発話に対応する映像へリンクが張られている．このようなハイパーテキストの例を図 4に示す．

5.2 軸となるタイムコードの選択

上で述べたように発話を書き起こしたテキストから映像へのリンクを張る際に問題となるのが，どのような粒度 (頻度と言ってもよい) でリンクを張るか，そしてその粒度がある程度粗い場合には，どのような基準でリンク付けを行うか，という点である．

まず前者のリンクの粒度の問題であるが，今回はかなり粗めのリンクを実現することにした．これは，4節で述べたように，2つのリンク (タイムコード) で囲まれた部分を単位としてテキストの検索を行うためであり，ある程度大きなテキストの集まりを単位とする方が検索システムとして使いやすいものになる，という理由からである．

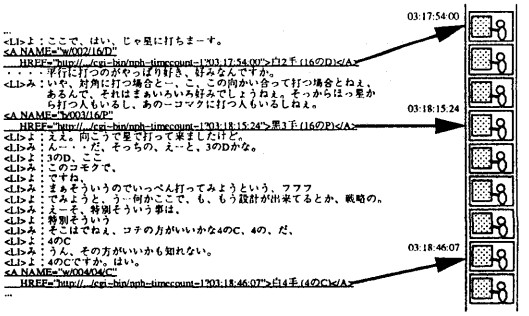

図 4: ハイパーテキスト化されたテキスト

次に後者のリンクを選ぶ基準について述べる．本試作システムの題材となるテキストおよび映像は，囲碁の対局の模様を収録したテープである．囲碁とその対局中に交わされる発話の特性から，今回は以下の2点をタイムコードの選択地点として採用した．まず第1に対局者が石を打った時点であり，第2には，長考などがあった際にそれを補完する意味で発話の途中でいくつかの点を選んでリンク付けを行っている．

5.3 テキストブラウザの実現

上記のようなハイパーテキストを閲覧するためのブラウザとして，今回は既存の WWW ブラウザ (具体的には NCSA が作成した Mosaic) を利用することにした．現時点では Mosaic 自体に改造を施すことはせず，ほぼそのままで使用している．

テキストブラウザにテキストが表示されている様子を図 5に示す．この図において下線が引かれている (更に実際の計算機のモニタ上では地のテキストとは別の色で表示されている) 部分には上述のリンクが付加されており，この部分をマウス等で選択してクリックすることにより，ビデオブラウザ上にその発話が交わされた時点の映像が表示されるようになっている．

6 ビデオからテキストへのリンク

本システムでは前述のテキストからビデオへのリンクに加えて，逆方向のビデオからテキストへのリンクも実現されている．ビデオからテキストへのリンクは，ビデオ中の時間的領域を始点とするものと空間的領域を始点とするものの2種類が実現されている．以下それらについて順に説明する．

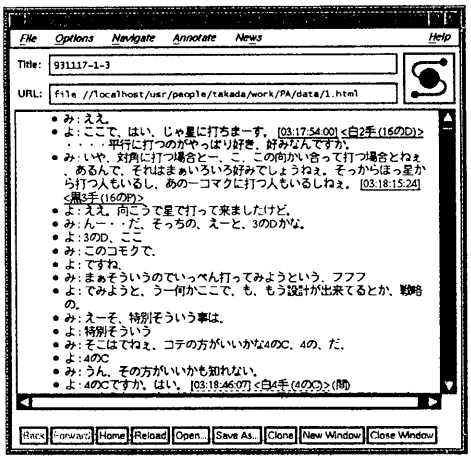

図 5: テキストブラウザ

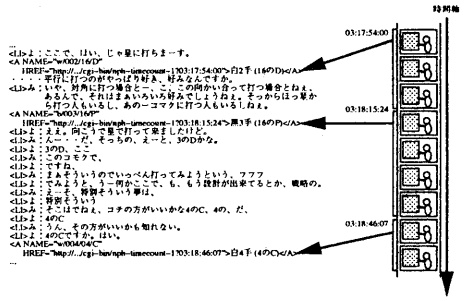

図 6: ビデオの時間的領域を始点とするリンク

6.1 ビデオの時間的領域を始点とするリンク

ビデオブラウザ上の GUI を用いて映像を早送り (巻戻し) した後に再生を開始すると，テキストブラウザ上では自動的にその映像に対応する部分の表示が可能になっている．これは別の表現をすれば，ビデオ映像の時間軸上のある地点から対応するテキストへのリンクが張られており，ビデオブラウザでビデオを表示 (再生) した際にそのリンクを辿ってテキストを表示すると言うことができる．

時間的領域を始点とするリンクの概念を図 6に示す．また，このリンクの粒度は前節で述べたテキストからビデオへのリンクと同じものが使用されている．

6.2 ビデオの空間的領域を始点とするリンク

そもそもビデオ映像というものは，2 次元の画像と 1 次元の時間を持ったデータである．したがって，一般にビデオ映像を始点とするリンクというものを考えると，

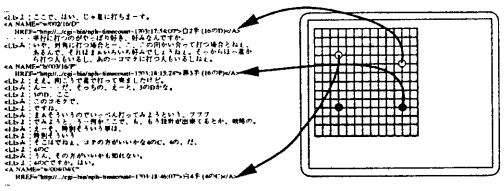

図 7: ビデオの空間的領域を始点とするリンク

当然前述の時間的領域だけではなく，映像中の空間的領域を始点とするリンクの実現も必要不可欠となるだろう．

既に述べたように，本試作システムで題材とされるのは囲碁の対局シーンを収録したものである．ここで使用されるビデオの映像は，対局中の碁盤を固定カメラで録画したものである (そしてその碁盤上の場所等を指し示す対局者の手なども収録されている)．そこで今回はこのような限定的な画像の状況を利用して，ビデオの空間的領域を始点とするリンクとして，以下のような機能を実現した．それについて以下説明をする．

ビデオブラウザ上に表示されるビデオ映像中のある領域，具体的には映像における碁盤上の石を指定すると，その石を打った時点の発話を書き起こしたテキストがテキストブラウザ上に表示される．これは言い換えれば，ビデオ映像の空間的領域から対応するテキストへのリンクが張られており，ビデオブラウザでそのリンクを選択すると，そのリンクを辿ってテキストを表示することを意味する．図 7にこの空間始点リンクの概念を示す．

また，この空間領域リンクと前述の時間的領域リンクとは直交した概念であり，空間領域リンクの選択は任意の時点で行うことが可能である．

6.3 ビデオブラウザの実現

上記の 2 種類のリンク機能，および，ビデオ表示の基本的な機能を実現するビデオブラウザを SGI/Indigo2 上に実現した．このビデオブラウザはビデオ映像の表示を行うウインドウの他に，再生・停止・早送・巻戻などのビデオデッキを操作するための GUI を持っている．

ビデオブラウザを図 8に示す．ビデオからテキストへの時間的領域を始点とするリンクの操作は主にブラウザ下部の操作ボタンによって行われる．また空間的領域を始点とするリンクの操作は，ブラウザ中央部の画像表示部内でマウス操作を行うことによって実現されている．

7 まとめと今後の課題

今回試作したシステムは，プロトコル解析を支援するための，テキストとビデオを統合し双方向のアクセスを可能にしたブラウジング・ツールである．

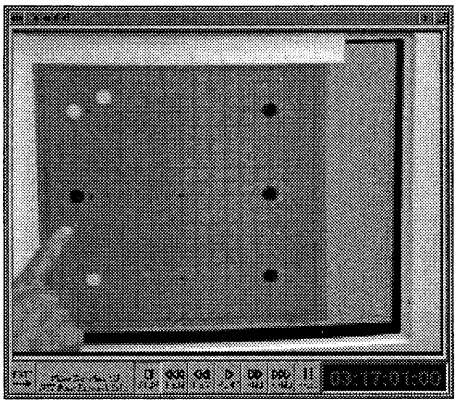

図 8: ビデオブラウザ

本システムの特徴をまとめると以下のようになる.

- ビデオテープの再生・操作を可能にするビデオブラウザ
- 実験内容を書き起こしたテキストをハイパーテキスト化し閲覧することを可能にするテキストブラウザ
- 上記のビデオとテキストの間を双方向に結ぶリンクとそれを利用するツール
- テキストに対して曖昧なキーワードを基に全文検索を行うシステム

現時点でのシステムは試作的なものであり, 数多くの問題点が残されている. それらの中で代表的なものについて, 以下簡単に述べる.

1. MPEG の利用
   今回の試作システムではビデオ素材を納めるメディアとして 8mm ビデオを使用している (Betacam ビデオ等でも同様のことが可能). しかしこのようなテープ方式のメディアは, ビデオに対するランダムアクセス性が非常に劣っており, システムの利用性を著しく不便なものにしている. この点に関しては今後のビデオ・メディア技術の発展などを考慮しつつ, プロトコル解析に十分な画質とランダムアクセス性を両立でき得るメディアを利用する方向へと進めていく予定である. 現時点では, 次段階として MPEG 等の利用を検討中である.
2. より多様なリンクの実現
   現時点では, テキストからビデオへのリンクの数も石を打った時点とそれを補完するもの程度に限られており, またビデオからテキストへのリンクに関しても, 今回題材とした「囲碁の対局場面」というものに強く依存したシステムとなっている. 今後は, 前者のテキストからのリンクに関しては, 利用者の目的に合わせてその粒度を選択可能にするなどの改良が必要となるであろう.
   後者のビデオからのリンクについては, それをより一般的概念として整理し直す必要があるだろう. さらにこの作業過程では, 映像を始点とするリンクとはどのようなものであるのか, あるいは, そもそも映像というものは何なのであろうかといった根本的再考察が必要になる可能性もある [6].
   またこのようなリンクを含むテキストやビデオを作成するための支援ツール (エディタ) の作成も必要不可欠になると考えられる.

今後は, 本システムをプロトコル解析を行っている研究者に試用してもらうことによりフィードバックを受け, その問題点や改良点などを明らかにしていく予定である. またこれをひとつのツールとして利用することにより, 今後, 意味検索の研究を展開していくつもりである.

謝辞

本研究を見守っていただく石井健一郎情報科学研究部長, ご討論いただく分散コンピューティング研究グループの方々, プロトコル解析に関してご教授いただき, 本システムを試用していただく斉藤康己メンタルプロセス研究グループリーダとグループ員の方々に深謝する.

参考文献

[1] 尾内理紀夫, 高田敏弘, 鷲坂光一: "意味コンピューティングに向けて", 第 36 回プログラミング・シンポジウム報告集, 情報処理学会, 1995.

[2] Matthew E. Hodge and Russell M. Sasnett: "Multimedia Computing", Addison-Wesley, 1993. (翻訳書: 尾内理紀夫, 竹内彰一, 原田康徳: "MIT のマルチメディア", アジソン・ウェスレイ・パブリッシャーズ・ジャパン, 1994.)

[3] 海保博之, 原田悦子: "プロトコル分析入門", 新曜社, 1993.

[4] Brewster Kahle: "Wide Area Information Servers Concepts", Thinking Machines technical report TMC-202, 1989.

[5] 松本裕治他: "日本語形態素解析システム JUMAN 使用説明書 version 1.0", 1993.

[6] Theodor H. Nelson: "Literary Machines", Mindful Press, 1991. (翻訳書: 竹内郁雄, 斉藤康己 (監訳): "リテラリーマシン", アスキー出版局, 1994.)