2003－HI－104 (7)
2003／7／11

インタラクション・コーパス分析ツールの試作

角 康之[†‡], 岩澤 昭一郎[†], 間瀬 健二[¶†]
sumi@acm.org
†ATR メディア情報科学研究所, ‡京都大学, ¶名古屋大学

人のインタラクションを理解し次世代ヒューマンインタフェースの実現に役立てるため、映像、音声、視線情報、生理データなどのマルチモーダルなデータで構成されるインタラクション・コーパスの構築を進めている。本稿では、記録された大量のインタラクション・データからインタラクションの構造をモデル化するための分析作業を支援するためのツール試作について報告し、今後の研究ロードマップを提案する。

Prototyping a Tool for Analysing Interaction Corpus

Yasuyuki Sumi[†‡], Shoichiro Iwasawa[†], Kenji Mase[¶†]
†ATR Media Information Science Laboratories, ‡Kyoto University, ¶Nagoya University

We propose to build an interaction corpus in order to understand the verbal/non-verbal mechanism of human interactions. Our approach is to employ wearable sensors (camera, microphone, physiological sensors) as well as ubiquitous sensors (camera, microphone, etc.); and to capture events from multiple viewpoints simultaneously. This paper presents a tool to interactively browse and analyse the huge amount of captured data.

1 はじめに

GUI(Graphical User Interface) やデスクトップメタファに象徴される現在の HCI(Human-Computer Interaction) パラダイムの転換を目指して、マルチモーダル・インタフェース、PUI(Perceptual UI)、エージェント型のインタフェースといったものが提案されている。これらはコンピュータと人のインタラクションを、より人間同士の自然なインタラクションに近づけようという共通目的をもっているので、その実現のためには、人と人、人ともの、人と環境の間のインタラクションのプロトコルを理解しモデル化するか、少なくとも、それらのプロトコルを機械可読にすることが必要であると考える。

言語コーパスを利用することで自然言語理解（特に機械翻訳）の研究が加速したように、また、音声コーパスを利用することで音声認識/生成の研究が加速したように、人のインタラクションを扱う HCI 研究にも研究インフラとしてのインタラクション・コーパスが有効であると我々は考えている [1]。そのとき重要となる視点は、言語情報に限らないマルチモダリティを対象とすることと、人と人の間で無意識にやり取りされる社会的なインタラクションのプロトコルを扱うことであると考える。

そういった基礎データを収集するために、我々は複数のセンサ群を用いて、多くの人のインタラクションを記録することから始めた [2]。本稿では、記録された大量のインタラクション・データからインタラクションの構造をモデル化するための分析作業を支援するためのツール試作について報告し、今後の研究ロードマップを提案する。

2 複数センサ群によるインタラクションの記録

開放的な空間における複数人のインタラクションを様々なセンサ群で記録することを試みた。そのためのテストベッドとして、筆者らが所属する ATR 研究所の研究発表会や学会のデモ会場を題材にして、デモ展示会場における展示者と見学者のインタラクションを対象としたインタラクション・コーパス収集システムを試作した [2]。以下、2002 年 11 月の ATR 研究発表会における試みを例に説明する。

我々の試みの特徴は以下の通りである。

- 人のインタラクションを構成している様々なモダ

図 1: インタラクション記録を試みたデモ展示会場の様子

リティを記録する。

- ユビキタスなセンサや主体となるユーザが身につけたセンサを利用して、同一のインタラクションを多角的に記録する。

- すべてのビデオカメラに対応させて IR トラッカを設置することで、視野に何/誰が映っているのかを実時間で記録する。このことは、注視（gazing）が人のインタラクションをインデクスする手段として利用できるであろう、ということを仮定している [3]。

- 人のインタラクションをただ受動的に記録するだけでなく、積極的にインタラクションを演出して意図的に人間のインタラクションパターンを記録するために、自律的に動作する人工物（ロボット等）を利用する。

図 1 は、インタラクション・コーパス構築のためにセットアップしたデモ展示会場のスナップショットである。展示会場には 5 つの展示ブースを用意した。各ブースの天井には前後 2 セットのセンサ群（ビデオカメラ、マイク、IR トラッカ）を設置した。またポスタやデモディスプレイそれぞれに LED タグを取りつけた。各展示ブースに立つ説明員は、ウェアラブルなセンサセット（カメラ、マイク、IR トラッカ、LED タグ、生体センサ）を身につけた。カメラと IR トラッカは側頭部に固定されるようにヘッドセットに取り付け、頭の向いている方向の映像の記録と、ユーザの前方に存在する LED タグの信号を認識できるようにした。見学者のうち希望者には説明員と同じウェアラブルセンサシステムを身につけてもらった。

2 日間のデモで 80 人のユーザが我々のシステムを利用し、300 時間近いビデオデータを収集することができた。

3　インタラクションの解釈

収集されたインタラクション・コーパスを利用したアプリケーションのひとつとして、ビデオサマリの自動生成システムの試作を行った [4]。

ビデオサマリを自動生成する基本的な方針として、

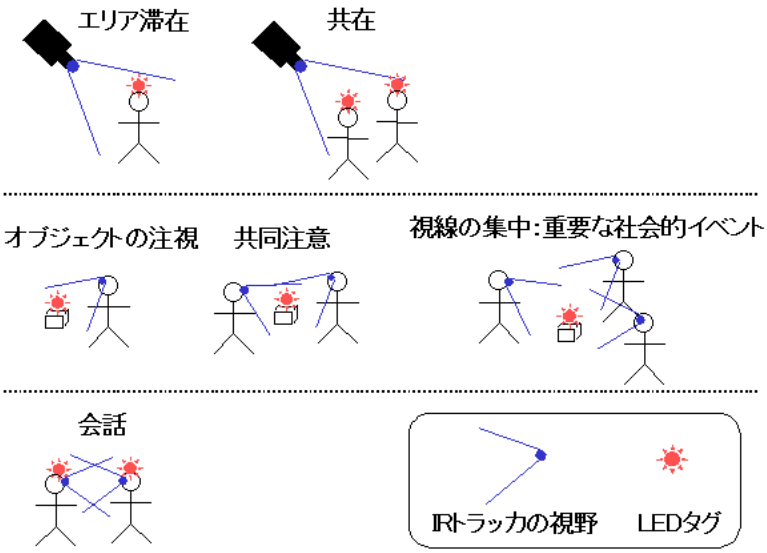

図 2: インタラクションのプリミティブ

IR トラッカによって与えられたインデクスを利用し、ボトムアップ的にインタラクションのシーンを切り出していくこととした。

イベントは、同一のカメラが同一の対象（人やもの）を捕え続けるビデオクリップであり、我々が扱うインタラクションの最小単位、つまりインタラクションのプリミティブと捉えることができる。すべてのイベントは、IR トラッカが LED タグを捕える、という意味では、これ以上単純化できないくらい単純な要素であるが、IR トラッカと LED タグの付与対象の組合わせ次第では、様々な意味を解釈することが可能となる。

図 2 に、いくつか基本的なイベントの解釈を図解する。

- IR トラッカが環境側に設置されたものであり、捕えられた LED タグが人に付与されたものである場合は、それはすなわち、その人があるエリアに滞在していることを意味する。また、同一の環境設置 IR トラッカに、複数の人の LED タグが同時に捕えられた場合は、それはすなわち、それらの人々が同じエリアに共在する状態を意味する。

- 人が身につけている IR トラッカが、あるものに付与された LED タグをとらえている場合は、それはすなわち、その人があるものを注視していることを意味する。また、同一の対象物を複数の人の IR トラッカが同時に捉えている場合は、それらの人々が同じものに対して共同注意を向けている状態であると考えられる。さらに共同注意に参加している人の人数が増えた場合、それはすなわち、注意を向けられている対象物は重要な社会的イベントを担っていると考えられる。

- ある人 A の IR トラッカが他の人 B の LED タグを捕え、同時に、B の IR トラッカが A の LED タグを捕えている場合は、それはすなわち、A と B が対話している状態であると解釈して良いであろう。

上記の通り、イベントはインタラクションのプリミティブであり、それに対応するビデオストリーム自体は短かすぎてひとつの意味のあるシーンとは言えない。そこで、時間的／空間的な共有性を持つ複数のイベントを連結させることでシーンを構成する戦略をとった。

4 ビデオサマリ

図 3 は、あるユーザのために集められたシーンを時間順に並べてビデオサマリを表示しているページの例である。シーンのアイコンは各シーンビデオのサムネイルであり、ビデオの時間長にサムネイルの濃淡を対応させた。

各シーンには、シーンの開始時刻、シーンの説明、シーンの時間を注釈として自動付与した。シーンの説明の生成には、*I talked with [someone]*、*I was with [someone]*、*I looked at [something]* といったテンプ

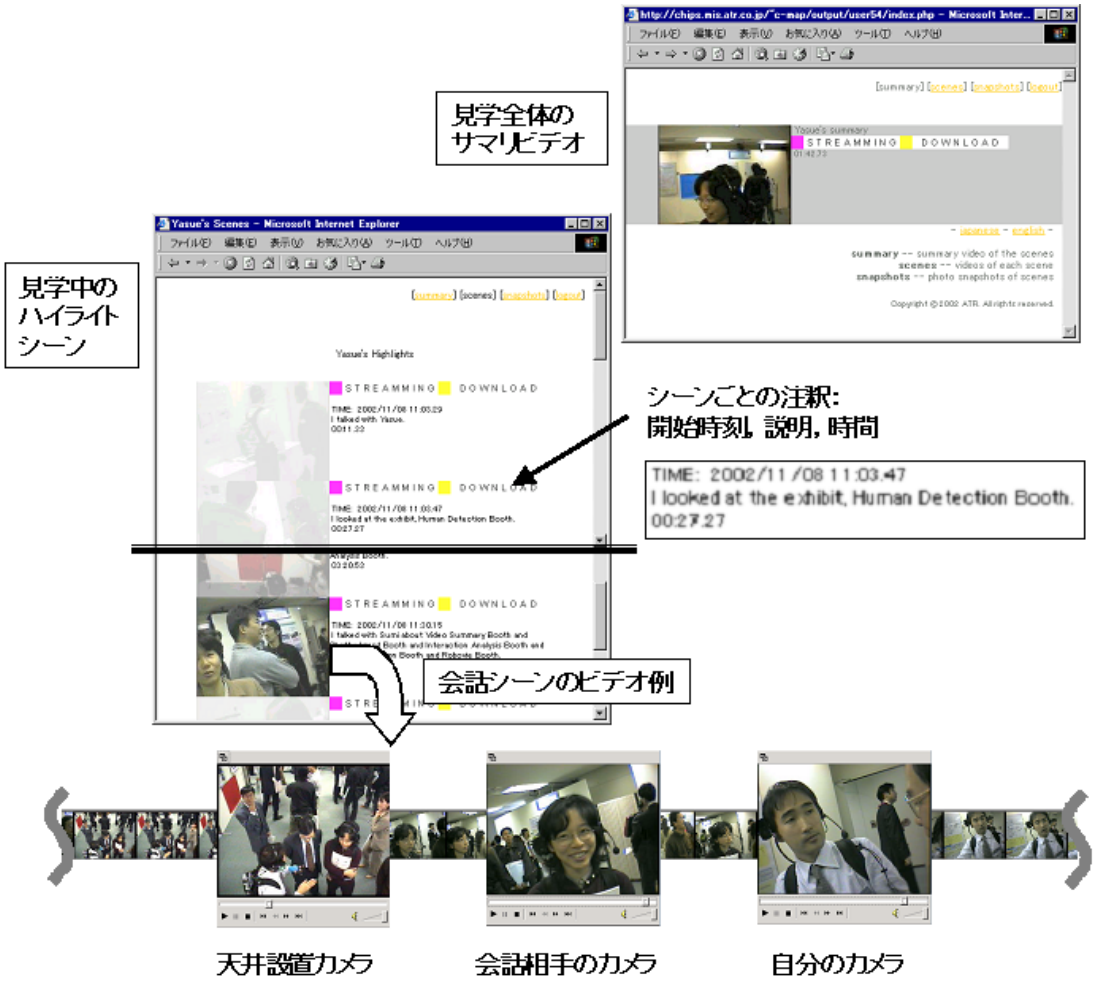

図 3: ビデオサマリ

レートを利用した。さらに、一つ一つのシーンを見ることすら面倒なユーザのために、各シーンを最大 15 秒ずつ切り出し、それらを fade-in, fade-out で連結して 1 本のクリップにまとめたサマリビデオも作成した。

シーンを構成するイベントは、単一のカメラとマイクの組合わせから撮られたものだけとは限らない。つまり、会話シーンであれば、自分のカメラだけでなく相手のカメラで記録されたクリップと、二人を撮影している環境側のカメラのクリップも利用される。マイクのボリュームを見ることで、発話しているユーザの顔（LED タグ）が映っているカメラの映像が採用されるようにした。

5 Corpus Viewer の試作

前節で紹介したビデオサマリは、展示見学に訪れた訪問者の見学記録をその場で短いビデオに自動要約するもので、訪問者一人一人が気軽に利用することを指向したものであった。それに対し、最初に述べた通り、インタラクション・コーパスを構築している本来の理由は、その膨大なデータからインタラクションのパターンを見出し、それをモデル化したり、パターンの辞書化を試みることである。前節のビデオサマリは、ヒューリスティクスを用いてハイライトシーンを自動的に切出したり、同一シーンのビデオクリップの中でのカメラの切り替えを自動化した。エンドユーザ向けのアウトプットとしては、そういった要約の自動化の基本方針は正しいと考えるが、分析者にとっては、特定のシーンのビデオを特定の視点でじっくりと観察したいであろう。したがって、そういった分析的な研究を進めるために、

膨大なデータで構成されるインタラクション・コーパスを、分析者の要求に動的に応えながら閲覧可能にするツールが必要であり、その第一歩として我々はCorpus Viewerと呼ばれるコーパス閲覧ツールの試作を始めた。

我々のインタラクション・コーパスは、大量のビデオや音声の生データにあわせて、SQL サーバに蓄積されたインデクス・データで構成されている。SQL に記録されるデータは大きく分けて以下のようなデータで構成される。

- 登録ユーザに関するテーブル。各ユーザの個人プロファイルに加えて、システムを利用した時間帯（利用セッション）と利用した機器セットの ID の対応関係を記録したテーブルである。

- 各キャプチャマシンが記録しているビデオや音の元データのインデクス・データ。元ビデオのクリップは 1 分ごとに分割されているので、それらのスタート時間とクリップ ID の対応テーブルとなる。

- IR トラッカデータ。各ビデオに対応して、その視野に移っている LED タグの ID を検出した結果が時系列に並ぶ。

- 生体データ。Procomp+から送られるデータを時分割した値を時系列に並べたデータである。どの生体データがどのユーザのものであるかは、上記のユーザに関するテーブルを参照して対応させる。

上記のような、インデクス・データが SQL サーバに格納されているので、SQL を利用できるプログラマであれば、自ら問合せ文を記述して、様々な分析を行うことができる。実際、前節に示したビデオサマリのアプリケーションも、主に IR トラッカのデータに着目して複数の SQL 問合せ文を組み合わせて、その結果に基づいてビデオクリップを自動編集するプログラムである。

しかし、インタラクション・コーパスを分析に利用することを想定する認知心理学者やインタラクティブ・システムのデザイナが誰でも SQL プログラムを自ら書けるとは考えづらいし、また、いくつかの基本的な問合せ文は再利用性が高いであろうから、そういったものを提供することは有益であろうと考える。そこで、そういった基本的な分析研究に利用可能であろうと考えるCorpus Viewerを試作した。

基本的な方針は以下のようにした。

- GUI を提供する。GUI としては、汎用性を考えWeb アプリケーションにすることにした。

- php による CGI の問合せと、Java Script を利用した対話性を統合した。

- まずは IR トラッカのデータを利用してインタラクションの「密度」を可視化することを試みた。生体データやロボットの行動データなどの他のインデクス・データを指標とした可視化も可能であろうが、それは今後の課題とした。

- ユーザ（分析者）が着目する部分を選択すると、そこに対応したビデオデータを簡単に閲覧できるようにした。

試作したCorpus Viewerを利用している様子を図 4 に示す。なお、混乱を避けるため、コーパス記録のデモ展示においてキャプチャ・システムを利用したユーザのことを、以下では「イベント参加者」と呼ぶこととする。

グラフはある一人のイベント参加者のセッション中のインタラクションの様子を可視化したものである。縦軸が時間軸で、上から下に向かって時間が進んでいる。図 4 の例ではある日の 10 時過ぎから 17 時過ぎまでの 7 時間のセッションの全体像を表している。縦に伸びた帯は、一本ずつがそれぞれ、他のイベント参加者や天井に備え付けた IR トラッカと LED ID を表している。その帯の上には、今着目しているイベント参加者との「インタラクション」があった瞬間をマークしてある。つまり、IR トラッカを表す帯の上には、その IR トラッカがイベント参加者の LED ID を捕らえた瞬間をマークし、逆に、他者の LED ID の帯の上には、着目しているイベント参加者の IR トラッカにその LED ID が捕らえられた瞬間をマークしている。したがって、マークが密集している部分は、その対象物（他者）と着目しているイベント参加者が「密に」インタラクションしていることが直感的に理解できる。なお、これらの帯は、選択されたセッション中に少なくとも一度でもインタラクションがあったものだけが表示されており、また、インタラクションの回数（IR トラッカに LED ID が捕らえられる回数）が多かったものを左から順に並べて表示した。

このような可視化結果を、以下のような手順で利用することを想定している。

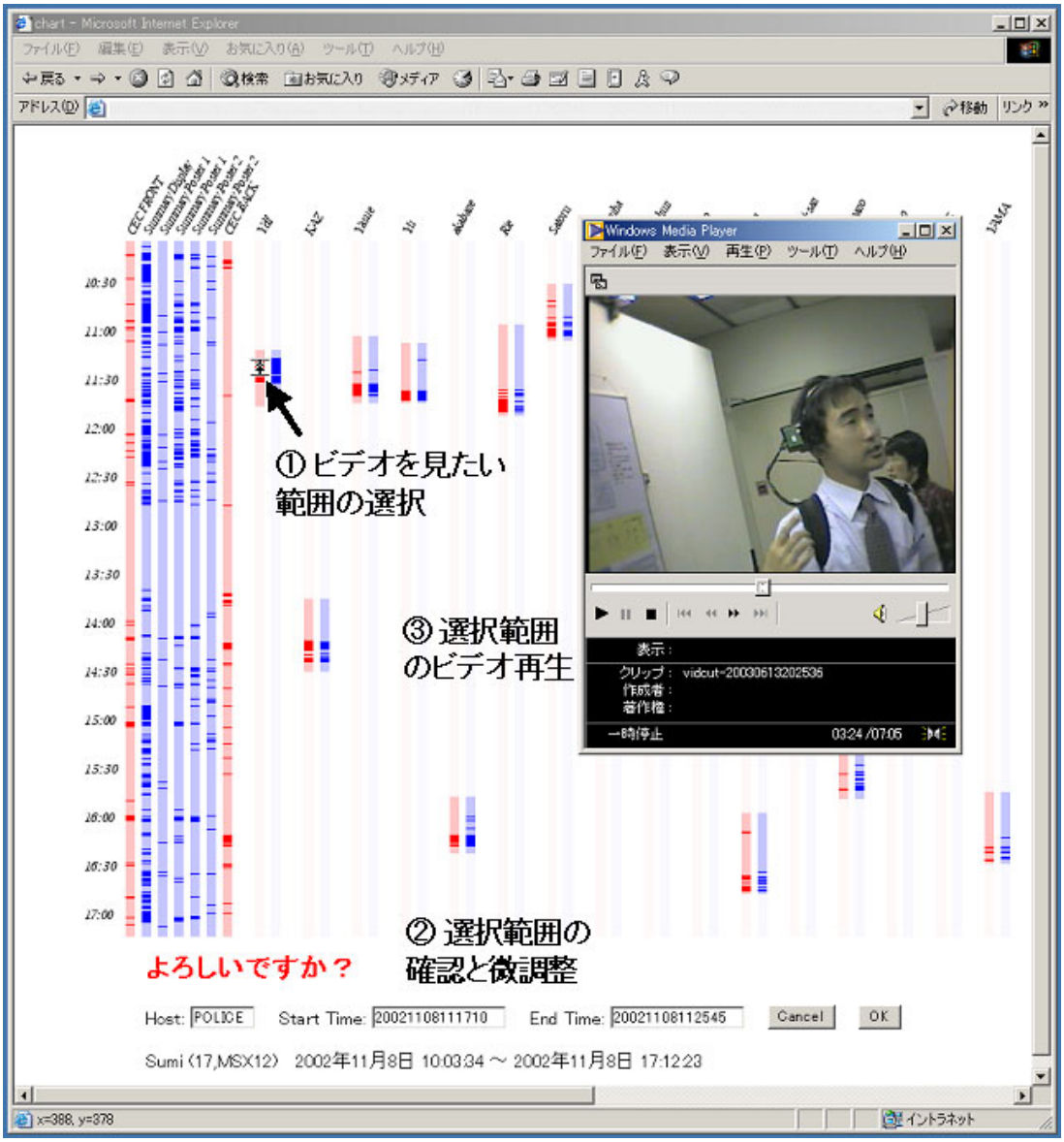

図 4: Corpus Viewer を利用している様子

1. コーパスに記録されているイベント参加者のリストから、着目しているイベント参加者を選択する。

2. すると、そのイベント参加者のインタラクション記録セッションがリストアップされる。通常のイベント訪問者ならばセッションはひとつであるが、イベントにおけるデモの説明者の場合は複数のセッションがリストに出る可能性がある。分析者はそこからひとつを選ぶ。

3. すると php プログラムが選択されたイベント参加者が「インタラクションした」他のイベント参加者や天井備え付けのセンサセットをリストアップして、そのインタラクションの様子を図 4 にあるようなグラフで可視化する。

4. 分析者は、グラフから、そのイベント参加者のインタラクションのパターン（他の参加者との対話や、ある展示ブースへの滞在など）を見出す。

5. そして、個別のインタラクションの様子を閲覧したい場合、その該当部分を選択し、図 4 にあるようにオンデマンドでビデオクリップを見ることができる。その際、サーバ側では、選択された範囲に入る複数のビデオデータを集め、それらを時間順につなげて 1 本のクリップにする作業が内部的に行われている。

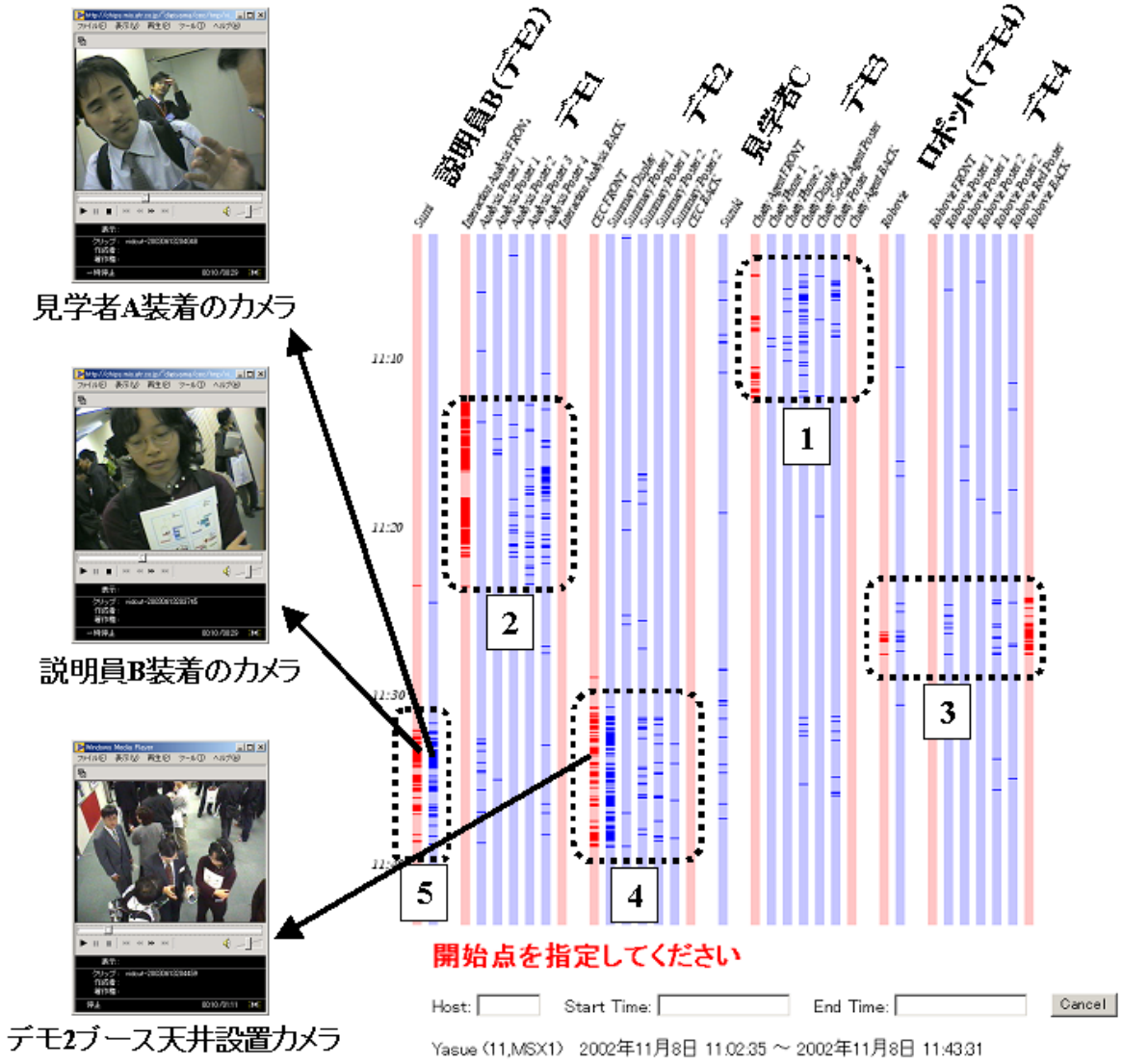

図 5: 見学者 A のインタラクション・データの閲覧例

6 Corpus Viewer の利用例

現在、実際に Corpus Viewer を認知心理学者に利用してもらいながら、彼らのビデオプロトコル分析に堪えられるものかを評価してもらい、また、他に有益なツールが何であるかのフィードバックをもらっている。

Corpus Viewer を用いてどのような分析が可能であるかを示すために、簡単な例を図 5 に示す。これは、典型的なデモ展示見学者 A を選択して、彼女のインタラクションの様子を可視化した例である。

グラフを見ると、容易に「インタラクションのクラスタ」を見出すことができる。それを時間順に見ていくと、見学者 A は 40 分間程度のセッション（見学）中に、デモ 3、デモ 1、デモ 4、デモ 2 という順番で展示ブースを廻覧していることがわかる[1] 。ちなみに、デモ 4 のブースにはロボットが展示されており、ロボットとも「会話」インタラクションをしたことがわかる。また、見学者 C がセッション全体を通して断続的に見学者 A の視界に入っていることから、見学者 C は見学者 A と一緒に会場を回っていたらしいことが想像できる[2] 。

見学者 A が最も多くインタラクションした相手

[1] 各デモ展示ブースには、2 組のビデオカメラと IR トラッカのセットがあり、5 つ程度の LED ID が展示品やポスタに貼られていた。したがって、デモ展示ブースに対応する部分は、7 本程度の帯がセットになって表示される。

[2] 見学者 C はカメラと IR トラッカのヘッドセットを利用せず、LED ID のみをバッジとして装着していたので、帯が 1 本だけになっている。

は、デモ2のブースにいた説明員Bである。そのときの様子を実際にビデオで見るには、各帯の対応部分を選択して、それぞれのビデオをオンデマンドで生成すれば良い。そうすることで、同じ時刻の同一シーンを、複数の視点から（この例では、見学者Aと、会話の相手の説明員Bと、2人をとらえている天井の3つのカメラ）の映像で見比べることができる。このことは、インタラクションの分析を行う認知心理学者にとって、人の身振り手振り、視線の動き、マクロ的なフォーメーション（立ち位置や周りの状況）を複数の角度から調べることができ、有益であると考える。

この例で示したように、インタラクションを理解するための分析作業において、Corpus Viewerは、着目すべきシーンを「見極める」ためのツールとして利用できる。大量のビデオデータの中から、インタラクションのクラスタが存在している範囲を対話的に選択してビデオを閲覧できるのは、便利であると考える。

また、そういった個別シーンの見極めだけでなく、インタラクションのマクロ的なパターンを見つけるのに、本ツールは役立つのではないかと期待している。つまり、図5からは、インタラクションのクラスタのサイズやその時間的な推移に何らかのパターンが見出せるであろうし、また、同一シーンに関与している他のイベント参加者とのインタラクションの密度や、そのシーンへの参加や脱退のタイミングが直感的に見出すことができる。こういった対話的な可視化ツールが、個別インタラクション要素のミクロ的な構造や、また、複数のインタラクション要素間のマクロ的な出現パターンを発見するための支援ツールになると期待しており、人間のインタラクションを理解するための研究が加速されることを望んでいる。

7　おわりに

人のインタラクションを記録したインタラクション・コーパスのデータを分析するための支援ツールCorpus Viewerを紹介した。現在は、数あるインデクスのうちIRトラッカによるデータのみでインタラクションの構造を可視化することを試みた。今後、他のインデクス（生体データなど）を指標にした可視化も試みたい。

また、現状のCorpus Viewerは、ある一人のイベント参加者を中心にした1対多の関係を可視化するものである。が、実際の分析では、まず先に誰かに着目するよりは、「3人以上の会話シーンを分析したい」とか「誰かが10分以上、同一の場所にとどまっているシーンを見たい」といったように分析対象シーンを絞り込むことがあると考える。こういったことも基本的には、上記の言葉をSQLの問合せ文に翻訳すれば良いわけであるが、分析者本人にそれを求めるのは現実的ではない。したがって、いくつかのプリミティブな問合せ文を定義して、それらの組み合わせから簡単に様々な問合せ文を生成できるようなツールを用意することが今後の課題である。

謝辞

本稿で紹介したCorpus Viewerの実装にご協力頂いている山本哲史氏、インタラクション・コーパス構築について日頃から議論頂いている伊藤禎宣氏、中原淳氏、坊農真弓氏をはじめとするATRメディア情報科学研究所および知能ロボティクス研究所の各氏に感謝する。本研究は、通信・放送機構の研究委託「超高速知能ネットワーク社会に向けた新しいインタラクション・メディアの研究開発」により実施したものである。

参考文献

[1] 角康之, 間瀬健二, 萩田紀博. 人と人工物の共生を実現するためのインタラクション・コーパス. 第16回人工知能学会全国大会, 2002.

[2] 角康之, 伊藤禎宣, 松口哲也, Sidney Fels, 内海章, 鈴木紀子, 中原淳, 岩澤昭一郎, 小暮潔, 間瀬健二, 萩田紀博. 複数センサ群による協調的なインタラクションの記録. インタラクション2003, pp. 255–262. 情報処理学会, 2003.

[3] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing. In *ACM Multimedia '99*, pp. 3–10. ACM, 1999.

[4] Tetsuya Matsuguchi, Yasuyuki Sumi, and Kenji Mase. Deciphering interactions from spatio-temporal data. 情処研報, ヒューマンインタフェース, Vol. HI102, , 2003.