3E1-02

イベント空間におけるインタラクションの支援から理解へ

Supporting and Understanding of Interactions between People Situated in Event Spaces

角 康之
Yasuyuki Sumi

京都大学／ ATR メディア情報科学研究所
Kyoto University / ATR Media Information Science Laboratories

We have constructed an interaction corpus in order to understand verbal/non-verbal mechanisms of human interactions. Our approach employs *wearable* sensors, such as: video camera, microphone, physiological sensors, and tracking tags, and *ubiquitous* sensors, including: video camera, microphones, and tracking tags, to capture all events from multiple viewpoints simultaneously. This paper presents a prototype for capturing and summarizing interactions among presenters and visitors in an exhibition room.

1. はじめに

ユビキタスコンピューティングという概念の提唱者の一人と言われている Weiser は、「これからのコンピュータは、我々の生活のあらゆる所に浸透し、本来のタスクのためにその姿を消すべきである」というビジョンを提案した [1]。博物館での展示閲覧という文脈で考えると、本オーガナイズド・セッション「イベント空間支援」でも様々な提案があるように、携帯情報端末、キオスク、ロボットなどを利用した展示閲覧支援システムが利用されるであろう、というような未来像が議論されている。

実際、我々も過去の数年間、C-MAP (Context-aware Mobile Assistant Project) と呼ばれる、展示見学ガイドシステムに関する研究プロジェクトを進めてきた。そして、1999 年から 2001 年までの 3 年間に、人工知能学会の全国大会の参加者の会議参加の支援や、参加者同士のコミュニケーションの支援を目指したサービスを提供する運用試験を重ねてきた（図 1参照）[2]。

しかし、これらの研究事例で議論されている未来像ではまだ、博物館での本来の目的、つまり展示物の閲覧行為のために、支援システムがその姿を消すには至っていない。むしろ、展示物よりも目立っているのではないだろうか。それでは、本末転倒である。

究極の閲覧支援システムは、普段はその姿を消し、必要に応じて音声や展示物への重畳映像によって付加情報を提示するものであろうし、また、閲覧者が意識的に情報提示を要求するのではなく、閲覧者の無意識な動作（展示物を凝視する、側面を覗き込む、触ってみる等）から閲覧者の状況を自動的に認識し、的確な情報が提供されるべきであろう。

そういったシステムを実現するには、従来の、デスクトップメタファや GUI（Graphical User Interface）を基本コンセプトとした HCI（Human-Computer Interaction）設計パラダイムをこえて、身体全体を利用した新しい HCI 設計パラダイムが必要である。そのために、今後コンピュータには、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（人ならば無意識に理解しているような約束ごと）を理解してもらう必要があると考える。したがって、そういったインタラクションのプロトコルを機械可読にした、インタラクションの辞書を構築することが重要な課題であると考える。

そのための第一歩として、我々は、人と人のインタラクションにおける社会的プロトコルを分析・モデル化するために、複数人のインタラクションを様々なセンサ群で記録し、蓄積された大量のデータに緩い構造を与えてインタラクションのコーパスを構築することを目指している [3]。我々の試みの特徴は、環境に遍在するカメラ／マイクなどのセンサ群に加えて、インタラクションの主体となるユーザが身につけるカメラ／マイク／生体センサを利用することで、同一イベントを複数のセンサ群が多角的に記録することである。また、赤外線 LED を利用した ID タグ（LED タグ）と、それを認識する赤外線センサ（IR トラッカ）を利用して、各カメラの視野に入った人や物体の ID を自動認識することで、蓄積されるビデオデータに実時間でインデクスをつける。

複数ユーザの IR トラッカのデータや音声のボリュームを協調的に処理することにより、ユーザ同士のインタラクションの意味を解釈することが可能となり、大量のビデオデータの利用性が高まる。ここでは、インタラクション・コーパスの効果を示すために、展示見学における各見学者のビデオサマリを自動生成するシステム [4] を紹介する。

連絡先: 角 康之，京都大学，京都市左京区吉田本町，電話: 075-753-5381， FAX: 075-753-4961， sumi@i.kyoto-u.ac.jp

2. 複数センサ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試みを紹介する。テストベッドとして我々が所属する ATR 研究所の研究発表会を題材とし、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。

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 [5]。
- 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

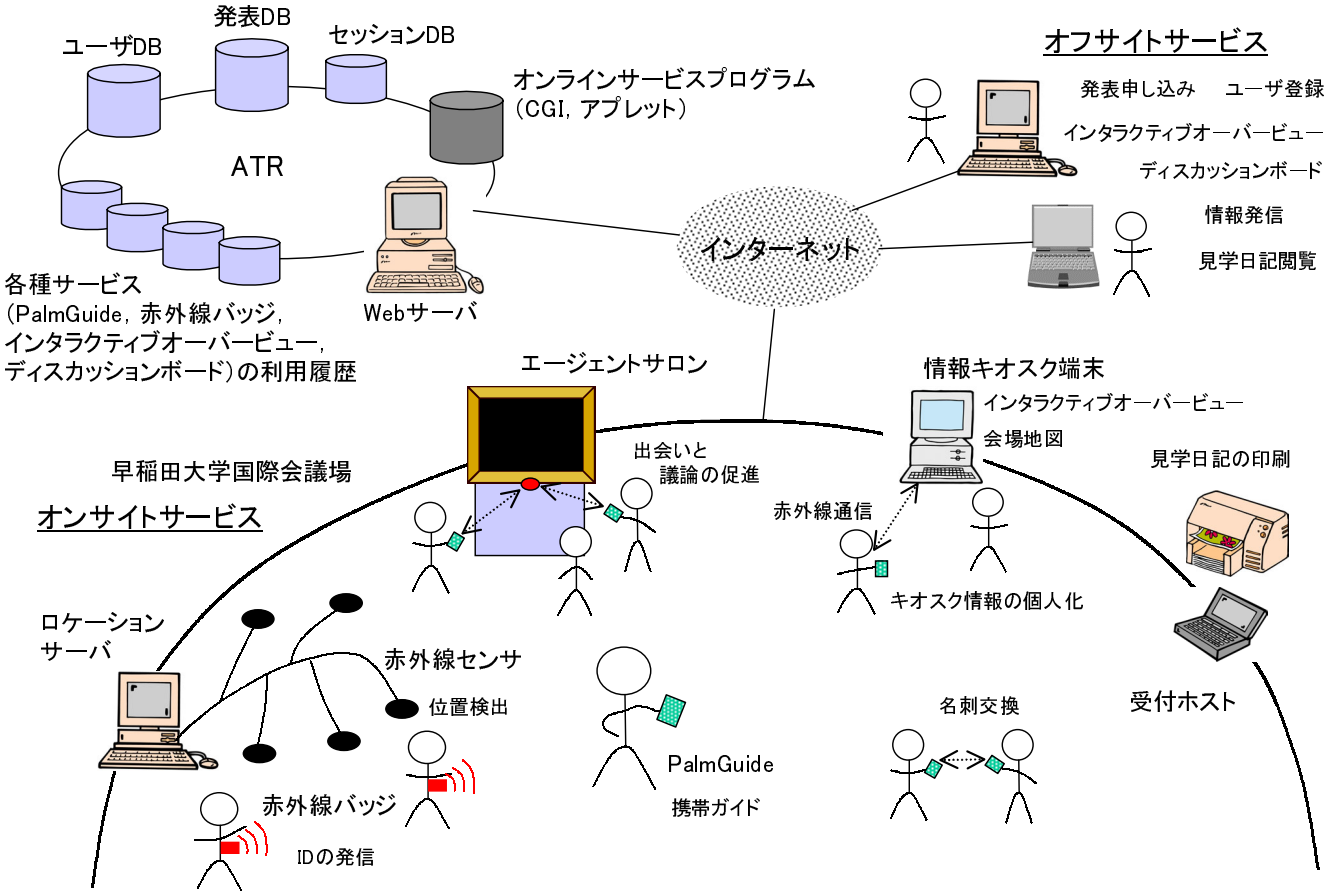

図 1: C-MAP システムの構成

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 2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

ここでは、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を取り上げる。ビデオサマリは、インタラクション・コーパスを利用して社会学的／認知科学的研究を行おうとする研究者の道具として重要であろうし、講演会、授業、普段のミーティングの記録の閲覧や、博物館の来訪者行動の分析など、エンドユーザが利用する道具としても利用価値が高いと考える。

ビデオサマリを自動生成する基本的な方針として、 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 3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

図 2: ユビキタス・センサ・ルームの様子

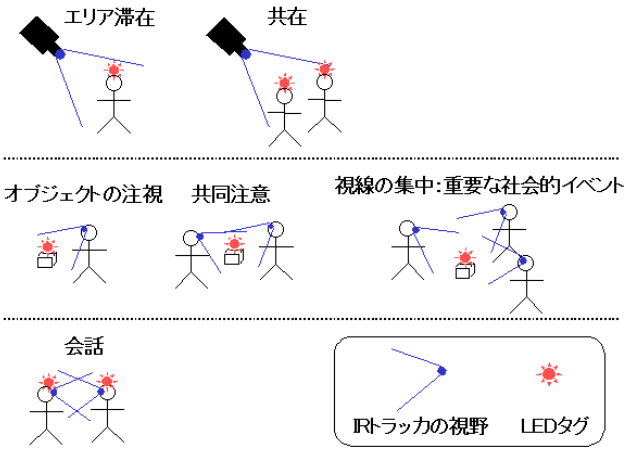

図 3: インタラクションのプリミティブ

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 4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。

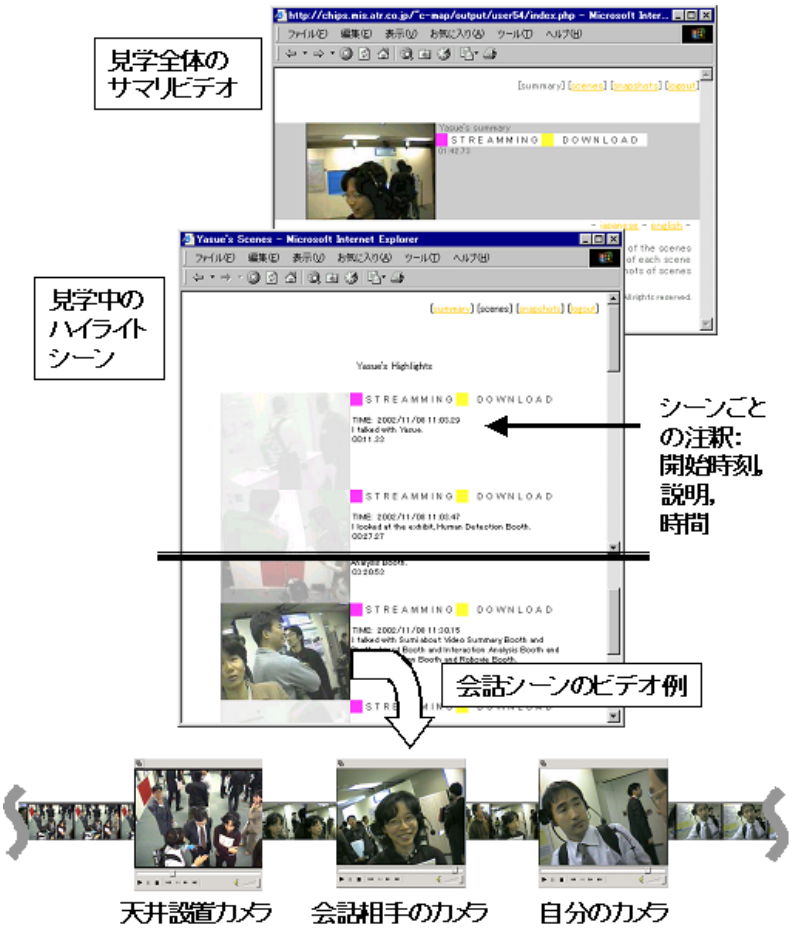

図 4: ビデオサマリ

5. おわりに

複数センサを利用したインタラクション・コーパス構築の試みを紹介した。提案手法は、ビデオデータ記録と同時に IR トラッカによる ID 付与を行うことが特徴である。試作システムによる 2 日間のデモを行い、そこでは、各ユーザの見学サマリをその場で提供することができた。今後は、 HCI デザインや社会心理学に興味を持つ研究者が簡単な操作でインタラクション・コーパスを利用できるように、システムを改善していきたいと考えている。

謝辞

本研究を共に進めている、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、山本哲史の諸氏に感謝する。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] Mark Weiser. The computer for the 21st century. *Scientific American*, Vol. 265, No. 30, pp. 94–104, 1991.

[2] 角康之. JSAI2000 デジタルアシスタントプロジェクトの報告. 人工知能学会誌, Vol. 15, No. 6, pp. 1012–1026, 2000.

[3] 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.

[4] 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.

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