複数センサ群による協調的なインタラクションの記録

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[†,¶]
†ATR メディア情報科学研究所　‡ブリティッシュコロンビア大学　¶ATR 知能ロボティクス研究所　§名古屋大学
sumi@atr.co.jp

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、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手法を提案する。提案手法の特徴は、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 ID タグシステムを利用して、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ことができる。本稿では、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記録し、各人のビデオサマリを自動生成するシステムを紹介する。個人のビデオサマリを生成する際、本人のセンサデータだけでなく、インタラクションの相手のセンサデータも協調的に利用される。

Collaborative Capturing of Interactions by Multiple Sensors

Yasuyuki Sumi[†]　Sadanori Ito[†]　Tetsuya Matsuguchi[†]　Sidney Fels[‡]　Akira Utsumi[†]　Noriko Suzuki[†]
Atsushi Nakahara[†]　Shoichiro Iwasawa[†]　Kiyoshi Kogure[¶]　Kenji Mase[§,†,¶]　Norihiro Hagita[†,¶]
†ATR Media Information Science Laboratories　‡The University of British Columbia
¶ATR Intelligent Robotics and Communication Laboratories　§Nagoya University

We are exploring a new medium in which our daily experiences are recorded using various sensors and easily shared by the users, 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 prototype to capture and summarize interactions among exhibitors and visitors at an exhibition site.

1　はじめに

近年、コンピュータは我々の生活に浸透し、家電やオフィス機器など、あらゆる電子機器に埋め込まれている。それらの多くは、従来のような、キーボード、マウス、ディスプレイを備えた典型的なコンピュータの形態を持たない。そして、近い将来、それらの電子機器は互いにネットワークでつながって連携動作し、我々の生活空間を包み込むようになるであろう。そうなったとき、従来の WIMP パラダイムで培ってきたような GUI やデスクトップメタファをベースにしたインタフェースだけでは不十分であり、もっと身体全体を利用した空間的インタフェースが求められると考える。

また、従来は単体としてのひとつのコンピュータが 1 人のユーザと 1 対 1 でインタラクションすることを基本としてきたが、我々の生活空間を包み込むようなコンピュータは、我々の社会生活、つまり、人と人のインタラクションを見守り参加する社会的要素として再設計されるべきであろう。

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）を理解してもらわなければいけない。そこで、我々は、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を大きな目標とする [1]。

そのための第一歩として、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手法を提案する。提案手法の特徴は、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。

本稿では、筆者らの所属する研究所の研究発表会におけるデモ展示会場において、展示者と見学者のインタラクションを記録するために試作したシステムを紹介する。また、蓄積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として、各ユーザの展示見学のビデオサマリを自動生成するシステムを紹介する。個人のビデオサマリを生成する際、本人のセンサデータだけでなく、インタラクションの相手のセンサデータも協調的に利用される。

2 複数センサ群によるインタラクション・コーパスの構築

人と人、人と人工物のインタラクションを広く捉えるために、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る。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。

我々の試みの特徴をまとめると以下のようになる。

- 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 [2]。
- 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 [3] 等）を利用する。

図 1にインタラクション・コーパス収集のためのシステム構成を示す。システムは基本的に、身につける携帯型の記録用クライアント、部屋に埋め込まれる据え置き型の記録用クライアントで構成される。それぞれ、カメラ、マイク、IR トラッカからのセンサデータを記録用サーバに中継する。携帯型クライアントのいくつかについては、生体データを記録するセンサも利用する。

記録データは、基本的にはカメラとマイクによるビデオデータである。また、それらのビデオデータのインデクスとして、記録開始時刻、記録時間といった基本的データの他に、IR トラッカが検出した LED タグの ID、生体データが刻一刻とデータベースに記録されていく。

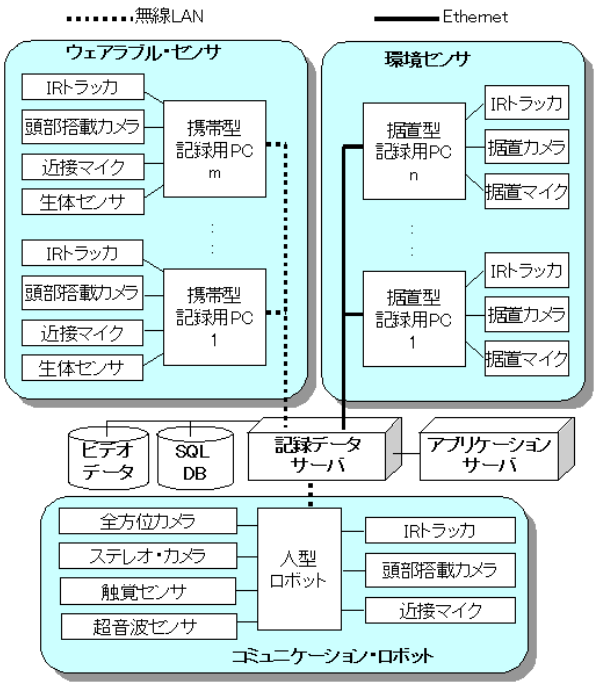

図 1: インタラクション・コーパス収集システムの構成

また、協創パートナーとしてのコミュニケーションロボットも、人とインタラクションする度に、自らのビヘイビア実行のログと、人との身体的な接触によるセンシングデータをサーバに記録する。

3 関連研究

これまでにも、環境側にカメラを埋め込み、部屋の中の人の行動を支援する試みが多くなされてきた（例えば、 Smart rooms[4], Intelligent room[5], AwareHome[6], EasyLiving[7] など）。これらは、コンピュータビジョン技術により人の存在や動きを認識し、ユーザの移動や行動意図を識別しようという試みであった。それに対し我々は、人の存在や移動の識別については赤外線 ID システムを使って楽をし、そのかわり、インタラクションのもう少しミクロなレベル、つまり、人と人の視線の一致や対話のプロソディに興味がある。また、インタラクションのデータを記録し、そのデータをさらなる創造的な協調活動に再利用することを目的とする。

ウェアラブルなビデオ収集システムを利用して、個人の記録を行う試みもなされてきた（例えば、 [8] や [9]）。しかしこれらは基本的に単体としての知的システムの構築を目指すものである。それに対して我々の試みは環境に埋め込まれたセンサ群や複数人のウェアラブルシステムの統合を利用し、人と人のインタラクションを協調的に記録し利用する枠組みを提案するものである。そのことで、一人の視点だけでは記録しきれないようなデータを相補的に記録することが可能になるであろうし、それと同時に、ユーザ

個人個人の主観的な視点を顕在化させることも可能になると考える。

生体データを利用して個人の記録にインデクスをつける試みもいくつかなされてきた（例えば、 StartleCam[10]）。しかしこれらの多くは、生体データ単独で人の心的状態を推定しようとするため、短絡的な解釈が多く、発展に行き詰まっている。元々、生体データはその人のおかれた状況（いつ、どこで、誰と、何をしているのか）に強く依存するものであると考えられる。したがって我々は、生体データ単独で心的状態を解釈しようと考えず、インタラクションに関する多角的なデータと共に生体データを記録し、横断的なパターンの変化から生体データ解釈のためのモデリングを行おうと考えている。

部屋の中での人の動きを認識するために、無線 [11] やウェアラブルな航行システム [12] を使って個人の位置を知る技術があった。それに対し、本研究では [13, 14] のように、赤外線を高速点滅された LED タグを用いて、見ている対象の ID を自動認識する技術を利用した。ただし [13, 14] ではセンサの認識スピードが十分速くなかったり、高価で携帯不可能という問題があった。そこで我々は、市販のビジョンチップとマイコンを利用してハードウェアレベルで高速の画像認識を行い、安価で携帯可能な赤外線 ID システムを開発した。

本稿では、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として、個人のビデオサマリを自動生成するシステムを紹介するが、それと関連する試みとして、ミーティングを記録したビデオをシーンごとに分割するシステムが提案されてきた（例えば [15]）。しかしそれらの多くは、固定カメラで捕らえられた画像の変化量に応じてシーンの切り替えを行うものであり、人のインタラクションのセグメンテーションを目指すものではない。我々のシステムは、複数人の視点による協調的なインタラクションのセグメンテーションを提供し、そこから自然にミーティングのメリハリやハイライトシーンの抽出を行える。

4 システム実装

筆者たちが所属する ATR 研究所の研究発表会が 2002 年 11 月 7,8 日に開かれた。それにあわせて、デモ展示会場の一部を「体験キャプチャルーム」と名付け、システムの試作を行った。

4.1 記録用サーバ – クライアントシステム

対象となる「体験キャプチャルーム」には 5 つの展示ブースが設置され、それぞれについて正面と背面に、つまり合計 10 台の据え置き型の記録用クライアントを設置した。また、展示者と見学者の希望者が身につけるための、携帯用の記録用クライアントを 15 台用意した。これらはすべて Windows パソコンである。複数センサ間の対応をとるには、時刻が重要な基軸になる。そこで、各クライアントは NTP を用いて 10ms 以上ずれないように設定した。

記録されるビデオデータは samba サーバを経由して UNIX のファイルサーバに記録される。また、ビデオデータに対するインデクス情報を記録するために、 Linux 上で動作する MySQL サーバを用意した。その他に、ビデオサマリを生成するために Linux ベースのアプリケーションサーバを用意し、そこでは MJPEG Tools を使ってビデオのカット編集プログラムを実行した。

4.2 ビデオデータ（映像と音）の記録

ビデオカメラは、据え置き用には SONY 製 CCD-MC100（41 万画素 1/4 インチ CCD）、携帯用には KEYENCE 製 CK-200 （25 万画素）を用いた。いずれも NTSC で記録用クライアントにデータを送る。マイクは、据置用クライアントには無指向性のマイク、携帯用には接話用のヘッドセットマイクを利用した。

ビデオについては、各記録用クライアントで Motion JPEG（解像度 320 × 240、 15 フレーム／秒）をリアルタイムエンコーディングした。音は PCM 22KHz 16bit モノラルで記録した。これらの値は、ネットワーク負荷と、全体のコーパスサイズを押さえることと、コンテンツとして再利用する際の品質のトレードオフで決めた。

一度のセッションをひとつの膨大なビデオファイルにするのは現実的ではないので、ビデオデータは内部的には 1 分ごとに別々のファイルにした。ただし、コーパスを利用する際にファイルが 1 分ごとにわかれていることを意識しなくて済むように、インデクスデータを SQL サーバで管理し、内部構造を隠蔽した。

4.3 赤外線 ID システム

LED タグは、 LED を高速に点滅させ、その点滅パターンで ID を発信し続ける。 IR トラッカは、 2.5 メートル先程度の LED タグを認識し、認識され次第、その ID と XY 座標を記録用クライアントを通して SQL サーバに書き込み続ける。

図 2が試作した LED タグと IR トラッカの外観である。 IR トラッカは、 LED の点滅を読み取る CMOS カメラとそれを制御するマイコンで構成される。見える範囲、 LED タグが送るデータビット数、 IR トラッカの認識スピードは、互いにトレードオフするが、今回は、 6 ビットの ID （つまり、 64 個のタグ）を扱い、秒あたり数回程度読み出せるものを試作した。そのため、視界の中で少々動いているタグの ID も読み取れる。なお、ヘッドマウント用のセンサには、同一ケースにコンテンツ撮影用の KEYENCE CCD カメラも入れ、 IR トラッカと光軸を合わせた。

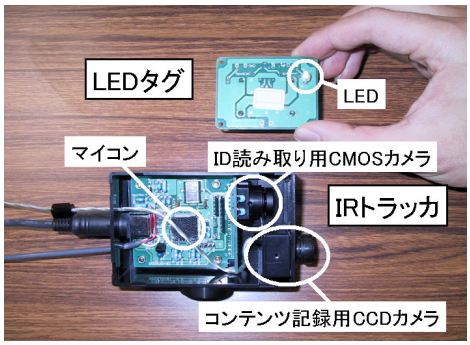

図 2: 試作した赤外線 ID システムの外観

4.4 生体データ

携帯型記録クライアントのうち 3 台については、生体データ記録用モジュール（Procomp+）を統合した。これは、リアルタイムに生体データを AD 変換してコンピュータに送る機器で、今回は脈拍、手の表面の伝導性（発汗）、温度の 3 つのセンサを使用した。これらのセンサはどれも指に付けることができるので、ちょうど片手が埋まった。数秒ごとにそれぞれの平均値を計算し、その値を記録用クライアントを通して SQL サーバに書き込み続ける。

4.5 クライアントシステムとタグの設置

図 3: センサ類と LED タグの設置

部屋には 10 台の据え置き型記録用クライアントを設置した。図 3にあるように、カメラ、マイク、 IR トラッカを天井と壁に設置し、各展示ブースごとに正面と背面から捉えた。また、人のインタラクションの焦点となるような点、つまり、ポスタやデモ用のディスプレイなどに、各展示ブースごとに 5 つ程度の LED タグを設置した。

展示員全員と見学者の希望者が身につけるセンサ用に、ウェアラブルな記録用クライアントを 15 台用意した。接話マイクを持つヘッドセットを利用し、それに、コンテンツ用のカメラ、 IR トラッカと LED タグをひとつにまとめたモジュールを固定した。パソコン（と 3 台については Procomp+ の本体）はバッグに納めて、背負うこととした。

4.6 記録データの形式

合計 25 台のクライアントのビデオデータ（映像と音）がインタラクション・コーパスの基本的なコンテンツとなる。画像については、今後オフラインの画像処理を行うかもしれないが、オンライン処理はしなかった。接話マイクからの音に関しては、ボリュームの変化量から発話タイミングを図る。これは、あとで会話シーンを解釈する際に利用される。

2 日間の研究発表会の間、 1 日あたり約 7 時間ずつシステムを稼働し続け、その間に合計 80 人のユーザが我々のウェアラブルセンサシステムを利用した。そのうち、説明員が 16 人、ロボットが 1 体含まれるので、来客者は 63 人であった。 2 日間の記録で、合計 300 時間近くのビデオデータが記録され、ビデオデータが 480GB、オーディオデータが 57GB に達した。また、 IR トラッカによる ID 検出の総数は約 38 万回に及んだ。

IR トラッカのデータは、生データは単なる検出結果の羅列なので、まずそこから時間方向の塊にまとめる。つまり、何が何時何分何秒に視界に入って何時何分何秒に視界からはずれたか、といった情報にまとめる。元データでは XY 座標も得られているので、それが右から左に通り抜けた、といったような情報も得られるが、今回はそこまでは利用していなかった。

生体データは、今回は収集のみ行い、オンラインで利用することはしなかった。今後オフライン処理により、他のモダリティ情報（声のボリュームやユーザの状況など）との関連性を探ることをしたい。

据え置き型記録用クライアントについては、システム起動時に、展示ブースの ID とクライアント ID を関連づけた。携帯型記録用クライアントについては、ユーザに貸し出す際にメールアドレスを ID としてユーザ登録を行い、それとクライアント ID、 LED タグの ID を関連づけた。

また、一部のユーザについては LED タグのみをバッジのようにして利用することを許した。その場合は、やはりメールアドレスでユーザ ID を発行し、それとバッジの ID を関連づけた。環境側に埋め込んだ LED タグについても、タグをつける対象（ポスタやデモ機器）を登録し、その ID と関連づけた。

4.7 協創パートナーの役割

「体験キャプチャルーム」のひとつの展示ブースはロボットに関するものであり、そこで自動走行させた人型ロボッ

トにも、展示者や見学者と同じように、携帯型の記録用クライアントを身につけさせた。つまり、我々のシステムの中では、ロボットもまったく人と同じ扱いをした。

ただ違う点としては、ロボットの場合は、動作の内部状態（話しかける、移動する、手を挙げる等）を逐一 SQL サーバに記録することができる。そのデータは、ロボットのインタラクションの相手であるユーザのエピソードの切り出しに利用することができる。

また、SQL サーバに逐一問い合わせることで、目の前にいるユーザの名前を参照したり、そのユーザのそれまでの行動履歴を得ることができるので、より個人化されたインタラクションを演出することが可能である。

5 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例：ビデオサマリの自動生成

我々のインタラクション・コーパスの主な利点は、計算コストの高い（映像・音声の）信号処理をすること無しに、インタラクションの切り出しやそれに参加している人の特定ができることである。

ここでは、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、赤外線 ID システム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。

まず次の用語を定義しておく。

イベント 同一のカメラとマイクの組み合わせによって記録されたビデオから、特定の LED タグが視界に入り続けている部分を切り出したクリップ。

シーン 例えば、展示ブース滞在シーンとか、会話シーンといったように、ある意味のある単位で、複数のイベントを組み合わせて生成されるビデオストリーム。

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 4に、いくつか基本的なイベントの解釈を図解する。

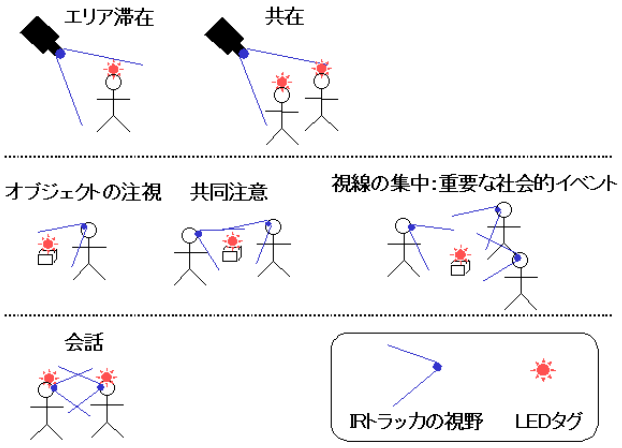

図 4: 様々なイベントの解釈

- 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 ID センサに、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。

前述した通り、IR トラッカによって出力される生データそのものは、断続的な LED タグの検出結果の羅列に過ぎない。したがって、そこから、特定の LED タグが視野に入り続けていたと判定する期間（interval）を特定し、それからそのときの IR トラッカと LED タグの担い手次第で上記のいずれかの解釈を当てはめて、それをひとつのイベントとする。

断続的な ID 検出列から interval を判定するにあたっては、ある IR トラッカに、maxInterval 以上の間隔を空けずに、LED タグが minInterval 以上の間検出され続けた

場合をイベントとして採用した。今回の試作では、 max-Interval、 minInterval 共に 5 秒とした。つまり、イベントの最小単位は 5 秒であり、また、同一 LED タグが検出されてもその間が 5 秒以上あいてしまった場合は、別のイベントに切り替わったものと判定した[1]。

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すぎてひとつの意味のあるシーンとは言えない。そこで、複数のイベントをボトムアップ的に連結させることでシーンを構成する戦略をとった。

例えば、今ユーザ A のためのシーンを構成しようとしている場合を考える。そのとき、ユーザ A の IR トラッカが何か LED タグを認識しているイベント、もしくは逆にユーザ A の LED タグが他のユーザや環境に付与している IR トラッカに捕えられているイベントがある程度の時間内で連続しているのであれば、それらを連結させて、ユーザ A にとって意味のあるシーンと解釈することとした。複数イベントが連続しているかどうかを判定するにあたっては、少なくともそれら 2 つのイベントが minInterval / 2 （2.5 秒）以上重なっていること、という指標を用いた。

また、空間的な同時性を有するイベント同士も、同一のシーンを形成するリソースとして連結させることとした。つまり、ユーザ A がユーザ B と会話している状態であると判定されるイベントがみつかったとき、ユーザ A の LED タグが認識されていなかったとしても、ユーザ B の LED タグが天井からの IR トラッカに捉えられていた場合には、その天井からの IR トラッカに対応するカメラにユーザ B と一緒にユーザ A も撮影されている可能性が高いので、そのカメラ映像もユーザ A のシーンを構成するリソースとして採用される。

開放的な空間における複数人の任意のインタラクションを捕えようとするとき、通常 occlusion の問題が起きるため、単一のカメラが同時に全員の LED タグを捕えることは稀である。したがって、上記のように、空間共有性を利用した複数のカメラリソースの連結を許すことが、あるインタラクションの塊全体を捕えるには重要な戦略になると考える。

極端な場合、空間共有性によるカメラリソースの連結を多段階繰り返すと、部屋全体のすべてのユーザがひとつのインタラクションに属する、と解釈されてしまうこともあり得よう。したがって今回は、空間共有性によるイベントの連結は 1 段階のみ許すこととしたが、このことは、どのようなサイズのインタラクションを観測したいか、目的に合わせて使い分けるパラメータであると考えられる。

[1]実際は maxInterval が 5 秒だと短すぎてイベントが細切れになりすぎてしまったので、デモ終了後、ビデオを目視しながら我々の直観に合う程度のイベント切り出しになるように maxInterval を調整し直した。その結果、固定カメラについては 10 秒が適当で、身につけたカメラについては（動きが激しいので） 20 秒が適当であることがわかった [16]。

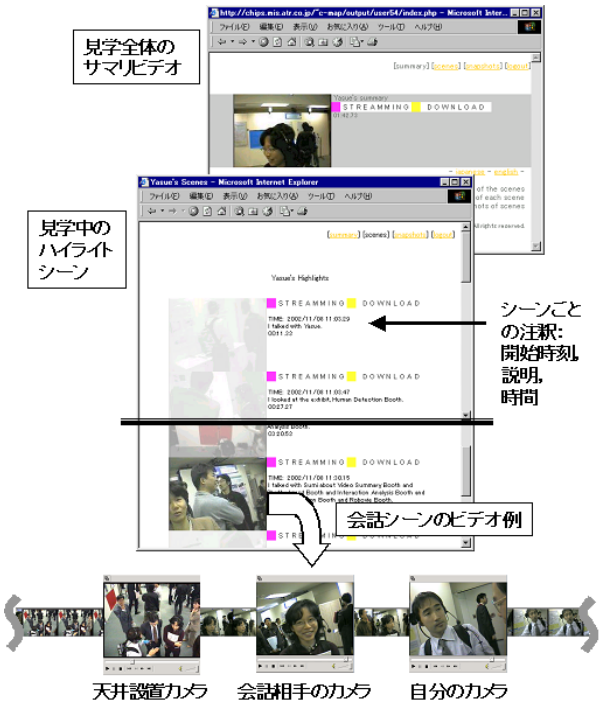

図 5: ユーザ個人のビデオサマリを表示するページの例

あるユーザのために生成された複数のシーンを時間順に並べると、そのユーザの展示見学のビデオサマリができる。図 5は、あるユーザ（見学者）のために実際に自動的に切り出されたシーンを時間順に並べてビデオサマリを表示したページの例である。

シーンのアイコンは各シーンのサムネイルを利用した。このアイコンをクリックすると MediaPlayer が起動し、対応するシーンのビデオクリップを見ることができる。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、以下の 3 種類のテンプレートを用意した。

TALK WITH I talked with [someone].

WAS WITH I was with [someone].

LOOKED AT I looked at [something].

これらは TALK WITH > WAS WITH > LOOKED AT の順に優先順位が高く、つまり、シーンの中に対話イベントが認識されれば、シーン全体の注釈としては TALK WITH が採用されるようにした。

また、展示会場での滞在時間が長くなるとシーンの数が多くなってくるので、クイックレビューが可能なように、シーンの時間的長さに応じてアイコンの濃淡を変えた。つまり、長い時間のシーンは色が濃くなるので、全体を見渡したときに目にとまりやすくなる。

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別途提供した。

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、さらには、二人を撮影している環境側のカメラのクリップが順々につながる可能性がある。また、音に関しては、ペアとなっているカメラと一致するとは限らない。例えば会話シーンでは、カメラ（クリップ）は切り替わっても、音は常に、会話者 2 人のマイクの音をミックスしたものを利用することとした。

以上で述べてきた通り、シーンは、時間の共有性と空間の共有性によって複数のイベントを集めて形成される。したがって、同じ時刻に複数のビデオリソース（イベントに対応したビデオクリップ）が存在することがある。そこで、簡単なカメラの切り替えルールを用意した。例えば、会話シーンの場合は、発話しているユーザの顔（実際は LED タグ）が写っているカメラの映像が採用されるようにした。それを実現するために、マイクのボリュームを見ることとした。そして、どちらのボリュームも低いとき（会話に沈黙があったとき）は天井もしくは壁からのカメラの映像に切り替わるようにした。

ビデオサマリは、記録されたデータをすぐに処理して出力することができるので、デモ当日に「体験キャプチャルーム」の最後のデモブースで、実際にユーザ本人のビデオサマリを見てもらうことができた。また 2003 年 1 月現在（デモの 2ヶ月後）、研究発表会後のアフターサービスとして、自分のビデオサマリを閲覧できる Web サービスを開始した。今後、ユーザのフィードバックを集めて、シーン切り出しのパラメータ調整や、ビデオサマリ表示の技法を再検討していきたい。

6　今後の課題

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開発や、インタラクションの原理の理解が今後の課題である。以下に、今後取り組みたい技術課題を列挙する。

会話シーンの分析 膨大な量のインタラクション・コーパスの中から会話シーンだけを抽出し、そこから会話における社会的インタラクションのパターンを分析したい。例えば、会話に参加するユーザの発話ボリュームの変化を可視化して、発話交代のパターンを可視化するツールを準備したい。

個人ユーザの移動のトラッキング 今のシステムの枠組みの中でも、展示ブースごとのエリア内であれば、理想的には各ユーザの存在とその滞在時間を認識することができる。しかし、展示エリア間の移動中は、環境側のカメラの配置が十分では無いのでトラッキングし続けることはできない。しかし、環境側のカメラの視界からはずれても、たまたま他のユーザ B の視界にユーザ A が入っている可能性があり、そのユーザ B がたまたまどこかの環境側カメラに認識されていれば、間接的に、どの辺りにユーザ A がいるのかがわかるので、ゾーン間の移動を補完できる可能性がある。そのためには、要所要所では 3 次元位置を特定できるホットスポットを用意する必要があるであろう。

ユーザ個人にとってのハイライトシーンの識別 同一ゾーンにおける滞在時間、生体データの変化量、音声のボリュームなど、様々なモダリティを統合させることで、ユーザ個人のハイライトシーンを識別するモデルを確立したい。今回の実験結果からも、例えば、見学者にとっては一カ所に滞在することが（興味のある展示ブースに滞在しているという意味で）ハイライトシーンとして認識されることは適当であったが、説明員にとっては、同じブースに滞在していることは当たり前なので、特に意味のない一人だけのシーンまでがハイライトシーンとして切り出されてしまった。シーン切り出しのルールは、ユーザの属性や状況に依存して変化すべきであると考える。

社会的なハイライトシーンの識別 上記の統計量や、同一対象（LED タグ）に同一時刻に視線が集まった瞬間を機械的に得ることができるので、展示時空間の中でのハイライトシーンを特定できるはずである。

協創パートナーの活用 ロボットが人とインタラクションした際の内部タスクのログから、ロボットが想定していたインタラクションシーン（挨拶する、個別の展示ポスタを見ることを促す等）を特定することができる。その結果と、これまでに述べてきたような、「見ること」によるインデクスを利用してボトムアップ的に形成されたシーンの識別結果との間で、どの程度、一致、不一致があるかを見極めたい。

ビデオサマリの表現技法 個別のシーンに対応したビデオクリップを作成する際には、複数のイベントクリップがリソースとして利用される。そこには、カメラの切り替えや音の扱いなど、映像技法に関わる課題がある。また、ビデオサマリのアイコンを並べて個人の展示見学サマリのページを提供する際には、アイコンの並べ方や注釈の付け方次第で、様々な利用形態を提案できると考えている。例えば、今回は個人に関わるシーンを時間順に並べたものだけを提供したが、本来、我々のシステムの枠組みでは、個別のイベントやシーンは複数のユーザやオブジェクトに帰属する。したがって、イベントやシーンに対応したビデオクリップをハイパーリンクで連結してユーザに提供することも可能であろう。そうすることで、あ

る時空間を共有した複数のユーザの間で、体験データを協調的に記録しそれらを共有し合えるような枠組みを提供できると考える。

社会的フィルタリングによる展示ガイド システムの運用にともない、どの展示がどういうユーザに人気あるか、といった統計情報を得ることができる。したがって、新たな来訪者に対して、社会的フィルタリングによる展示推薦や、見学のルートプランニングを提供することが可能である。

簡易的な拡張現実感の提供 今回試作した赤外線 ID システムは、コンテンツ撮影用のビデオカメラと光軸を合わせて利用することで、視界に入っているものや人の ID とその視界中の 2 次元座標を得ることができる。したがって、簡単に、見えているものや人の情報を見えている映像に重畳表示することができる。ユーザが身につけるヘッドセットのヘッドホンを利用し、さらにヘッドマウントディスプレイなども併用すれば、例えば、目の前にいる人の名前を表示したり、その人がそれまでにどの展示をどの程度見て廻ったのか、といったような情報を表示することができる。また、過去の情報を集約して表示することで、例えば、ある展示物に今までどのくらいの視線が集まったのかを定量的に表示することで、展示空間中の人気スポットが一目瞭然になるであろう。

3 次元仮想空間の再構成 「体験キャプチャルーム」の 3 次元 CG モデルを作り、そこに複数カメラでとられた映像データをレンダリングすることで、リアルな 3 次元仮想空間を再構築し、オンラインサービスとして提供したい。その際、固定カメラで得られた映像だけでなく、ユーザが身につけたカメラによる映像をコラージュ状に貼り詰めることで、「体験キャプチャルーム」内の人のインタラクションの営みのメリハリを顕在化させたい。つまり、多くのユーザの視線を集めたオブジェクトや人は多くの映像リソースに登場するであろうし、誰も目もくれないような場所（部屋の片隅や天井など）は映像のレンダリングがなされない。そういう不均一さをあえて強調することで、ある時空間を共有する人の営みを可視化する枠組みを構築したい。

7 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、インタフェース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

システム実装に協力頂いた山本哲史氏に感謝の意を表する。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, 間瀬, 萩田. 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] R.Stiefelhagen, J.Yang, and A.Waibel. Modeling focus of attention for meeting indexing. In *ACM Multimedia '99*, pp. 3–10, 1999.

[3] T.Kanda, H.Ishiguro, M.Imai, T.Ono, and K.Mase. A constructive approach for developing interactive humanoid robots. In *2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002)*, pp. 1265–1270, 2002.

[4] A.Pentland. Smart rooms. *Scientific American*, Vol. 274, No. 4, pp. 68–76, 1996.

[5] R.A.Brooks, M.Coen, D.Dang, J.De Bonet, J.Kramer, T.Lozano-Pérez, J.Mellor, P.Pook, C.Stauffer, L.Stein, M.Torrance, and M.Wessler. The intelligent room project. In *Proceedings of the Second International Cognitive Technology Conference (CT'97)*, pp. 271–278. IEEE, 1997.

[6] C.D.Kidd, R.Orr, G.D.Abowd, C.G.Atkeson, I.A.Essa, B.MacIntyre, E.Mynatt, T.E.Startner, and W.Newstetter. The aware home: A living laboratory for ubiquitous computing research. In *Proceedings of CoBuild'99 (Springer LNCS1670)*, pp. 190–197, 1999.

[7] B.Brumitt, B.Meyers, J.Krumm, A.Kern, and S.Shafer. EasyLiving: Technologies for intelligent environments. In *Proceedings of HUC 2000 (Springer LNCS1927)*, pp. 12–29, 2000.

[8] S.Mann. Humanistic intelligence: WearComp as a new framework for intelligence signal processing. *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2123–2125, 1998.

[9] T.Kawamura, Y.Kono, and M.Kidode. Wearable interfaces for a video diary: Towards memory retrieval, exchange, and transportation. In *The 6th International Symposium on Wearable Computers (ISWC2002)*, pp. 31–38. IEEE, 2002.

[10] J.Healey and R.W.Picard. StartleCam: A cybernetic wearable camera. In *The 2th International Symposium on Wearable Computers (ISWC'98)*. IEEE, 1998.

[11] A.Ward, A.Jones, and A.Hopper. A new location technique for the active office. *IEEE Personal Communications*, Vol. 4, No. 5, pp. 42–47, 1997.

[12] S-W.Lee and K.Mase. Incremental motion-base location recognition. In *The 5th International Symposium on Wearable Computers (ISWC2001)*, pp. 123–130. IEEE, 2001.

[13] 青木. カメラで読み取る赤外線タグとその応用. インタラクティブシステムとソフトウェア VIII (WISS 2000), pp. 131–136. 近代科学社, 2000.

[14] 松下, 日原, 後, 吉村, 暦本. ID Cam：シーンと ID を同時に取得可能なイメージセンサ. インタラクション 2002, pp. 9–16. 情報処理学会, 2002.

[15] P.Chiu, A.Kapuskar, S.Reitmeier, and L.Wilcox. Meeting capture in a media enriched conference room. In *Proceedings of CoBuild'99 (Springer LNCS1670)*, pp. 79–88, 1999.

[16] T.Matsuguchi, Y.Sumi, and K.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.