協調的なインタラクションの記録と解釈

角 康之[†1,†2] 伊藤禎宣[†2] 松口哲也[†2,☆]
シドニー フェルス[†3] 間瀬健二[†2,†4]

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 and Interpretation of Interactions

Yasuyuki Sumi,[†1,†2] Sadanori Ito,[†2] Tetsuya Matsuguchi,[†2,☆]
Sidney Fels[†3] and Kenji Mase[†2,†4]

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（Windows，Icons，Menus，Pointing device）パラダイムで培ってきたような GUI（Graphical User Interface）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 1 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，筆者らは，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 1 歩として，人と人のインタラクショ

†1 京都大学情報学研究科
Graduate School of Informatics, Kyoto University
†2 ATR メディア情報科学研究所
ATR Media Information Science Laboratories
†3 ブリティッシュコロンビア大学
The University of British Columbia
†4 名古屋大学情報連携基盤センター
Information Technology Center, Nagoya University
☆ 現在，カリフォルニア大学サンフランシスコ校
Presently with University of California, San Francisco

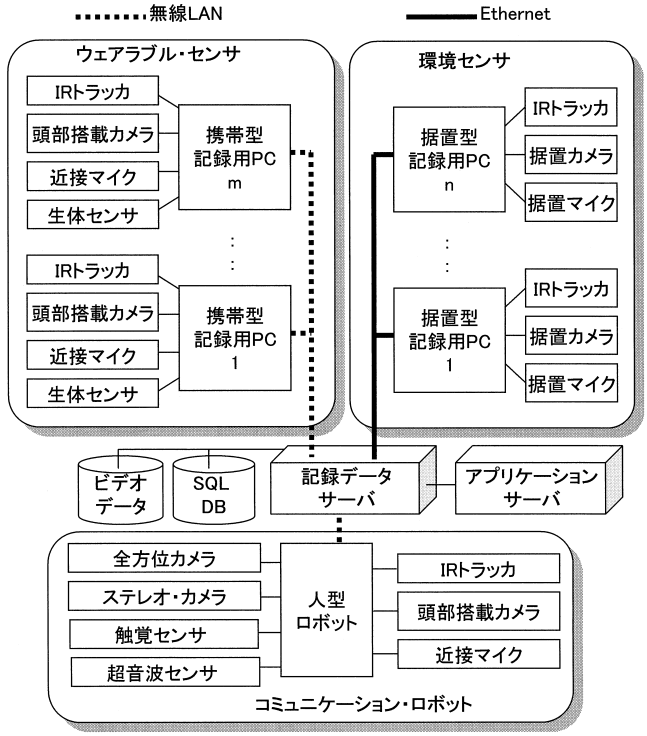

図 1 インタラクション・コーパス収集システムの構成
Fig. 1 Architecture of the system for capturing interactions.

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（Light-Emitting Diode）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とらえるために，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．

筆者ら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．

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たびに，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（たとえば，Smart rooms[4]，Intelligent room[5]，Aware-Home[6]，Kidsroom[7]，EasyLiving[8]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．それに対し筆者らは，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（たとえば，文献 9) や 10)）．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．それに対して筆者ら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，1 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（たとえば，StartleCam[11]）．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．したがって筆者らは，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とも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[12]やウェアラブルな航行システム[13]を使って個人の位置を知る技術があった．それに対し，本研究では文献 14), 15) 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．ただし文献 14), 15) 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．そこで筆者らは，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（たとえば文献 16)）．しかしそれらの多くは，固定カメラでとらえられた画像の変化量に応じてシーンの切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．筆者らのシステムは，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めりはりやハイライトシーンの抽出を行える．

個人の行動文脈の変化を記録してエピソード記憶の強化を試みた先駆的な試みとして，Lamming らによる Forget-me-not[17]がある．これは赤外線を用いた位置検出システム[18]を利用している点でも，筆者らの試みと関連が深い．しかし，彼らが用いた赤外線位置検出システムは部屋単位でユーザ検出をするものであったため，記録できるインタラクションは「他のユーザ

と同じ時間帯に同じ部屋に滞在した」といったような粒度の大きなものに限られる．それに対し，筆者らの赤外線 ID システムはカメラ視野の中に入っている ID タグの座標を出力するので，もっと粒度の細かいインタラクション，つまり，特定の対象物をしばらく注視しているイベントや，2 人のユーザが向かい合っているイベントを検出したり，さらには，視線（正確には頭）の移動を検出したりすることもできる．

4. システム実装

筆者らが所属する 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（Network Time Protocol）を用いて 10 ms 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 SQL サーバ（MySQL を利用した）を用意した．そのほかに，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 22 KHz 16 bit 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おさえることと，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．

☆ MPEG データの編集ツール．http://mjpeg.sourceforge.net

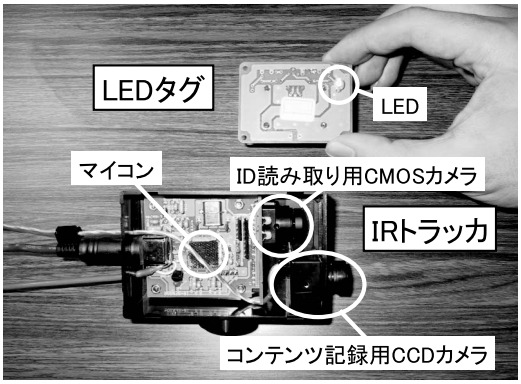

図 2 試作した赤外線 ID システムの外観
Fig. 2 IR tracker and LED tag.

一度のセッションを 1 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．

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

図 3 センサ類と LED タグの設置
Fig. 3 Setup of the sensor room.

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

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とらえた．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 1 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，音声信号のパワーの変化量から発話しているタイミングを推測した．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 人のユーザが筆者ら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 480 GB，オーディオデータが 57 GB 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 トラッカが出力する生データは ID 検出結果の不連続な羅列なので，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（発話音声のパワーやユーザの状況など）との関連性を探りたい．

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 1 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，筆者ら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，手をあげる等）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たりす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

5.1 インタラクションの解釈

筆者ら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なしに，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 1 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合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 たとえば，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）をとらえ続けるビデオクリップであり，筆者ら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とらえることができる．

すべてのイベントは，IR トラッカが LED タグをとらえる，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せに応じて，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．

- IR トラッカが環境側に設置されたものであり，とらえられた 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 タグが同時にとらえられた場合は，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 トラッカが同時にとらえている場合は，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 タグをとらえ，同時に，B の IR トラッカが A の LED タグをとらえている場合は，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よいであろう．

前述したとおり，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すぎない．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）

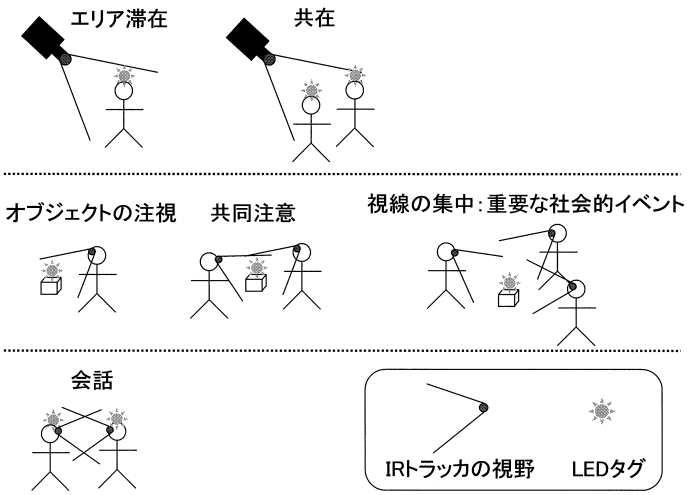

図 4 様々なイベントの解釈
Fig. 4 Interpretation of various events.

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あてはめて，それを 1 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 以上の時間検出され続けた場合をイベントとして採用した．最初の試作では，maxInterval，minInterval ともに 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．しかし，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筆者ら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分かった[19]．

上記のとおり，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 1 つの意味のあるシーンとはいえない．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．

たとえば，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 トラッカにとら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どうしも，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 トラッカにとら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とらえようとするとき，単一のカメラだけでは複数人の像の重なりが生じるため，単一カメラで同時に全員の LED タグをとら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とら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 1 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えよう．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

図 5 ユーザ個人のビデオサマリを表示するページの例
Fig. 5 Automated video summarization.

であると考えられる．

5.2 ビデオサマリの表示

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，1 つ 1 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．各シーンからの切り出しを最大 15 秒としたのは，サマリビデオ全体が長くなりすぎず，また，各シーンの内容を想起できる程度に短くなりすぎない妥協点を，経験的に調整した．また，切り出し部分は各シーンの開始からの 15 秒である．

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，2 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，組になっているカメラと一致するとは限らない．たとえば会話シーンでは，カメラ（クリップ）は切り替わっても，音はつねに，会話者 2 人のマイクの音を混合したものを利

用することとした．

以上で述べてきたとおり，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，簡単なカメラの切替えルールを用意した．つまり，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，マイクから得られる音声信号のパワーを見ることとした．発話中の（つまり，音声信号のパワーが大きい）ユーザの顔を写しているカメラ映像が見つからない場合は，会話に参加しているユーザたちが滞在しているブースに設置された環境側のカメラの映像を採用した．

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．またデモの 2 カ月後（2003 年 1 月），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. 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社会的インタラクションの分析にインタラクション・コーパスを利用できるように，対話的に任意のインタラクション・シーンを抽出し閲覧できるようなシステムを開発していきたいと考えている．

謝辞 システム実装や試用実験にご協力いただいた山本哲史，岩澤昭一郎，中原淳，内海章，鈴木紀子，小暮潔，萩田紀博の各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 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) Stiefelhagen, R., Yang, J. and Waibel, A.: Modeling focus of attention for meeting indexing, *ACM Multimedia '99*, pp.3–10, ACM (1999).
3) Kanda, T., Ishiguro, H., Imai, M., Ono, T. and Mase, K.: A constructive approach for developing interactive humanoid robots, *2002 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS 2002*), pp.1265–1270 (2002).
4) Pentland, A.: Smart rooms, *Scientific American*, Vol.274, No.4, pp.68–76 (1996).
5) Brooks, R.A., Coen, M., Dang, D., Bonet, J.D., Kramer, J., Lozano-Pérez, T., Mellor, J., Pook, P., Stauffer, C., Stein, L., Torrance, M. and Wessler, M.: The intelligent room project, *Proc. 2nd International Cognitive Technology Conference* (*CT'97*), pp.271–278, IEEE (1997).
6) Kidd, C.D., Orr, R., Abowd, G.D., Atkeson, C.G., Essa, I.A., MacIntyre, B., Mynatt, E., Startner, T.E. and Newstetter, W.: The aware home: A living laboratory for ubiquitous computing research, *Proc. CoBuild'99* (*Springer LNCS1670*), pp.190–197 (1999).
7) Bobick, A.F., Intille, S.S., Davis, J.W., Baird, F., Pinhanez, C.S., Campbell, L.W., Ivanov, Y.A., Schütte, A. and Wilson, A.: The KidsRoom: A perceptually-based interactive and immersive story environment, *Presence*, Vol.8, No.4, pp.369–393 (1999).
8) Brumitt, B., Meyers, B., Krumm, J., Kern, A. and Shafer, S.: EasyLiving: Technologies for intelligent environments, *Proc. HUC 2000* (*Springer LNCS1927*), pp.12–29 (2000).
9) Mann, S.: Humanistic intelligence: WearComp as a new framework for intelligence signal processing, *Proc. IEEE*, Vol.86, No.11, pp.2123–2125 (1998).
10) Kawamura, T., Kono, Y. and Kidode, M.: Wearable interfaces for a video diary: Towards memory retrieval, exchange, and transportation, *The 6th International Symposium on Wearable Computers* (*ISWC2002*), pp.31–38, IEEE (2002).
11) Healey, J. and Picard, R.W.: StartleCam: A cybernetic wearable camera, *The 2th International Symposium on Wearable Computers* (*ISWC'98*), IEEE (1998).
12) Ward, A., Jones, A. and Hopper, A.: A new location technique for the active office, *IEEE Personal Communications*, Vol.4, No.5, pp.42–

47 (1997).
13) Lee, S.-W. and Mase, K.: Incremental motion-base location recognition, *The 5th International Symposium on Wearable Computers* (*ISWC2001*), pp.123–130, IEEE (2001).
14) 青木 恒：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, 日本ソフトウェア科学会，近代科学社 (2000).
15) 松下伸行，日原大輔，後 輝行，吉村真一，暦本純一：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).
16) Chiu, P., Kapuskar, A., Reitmeier, S. and Wilcox, L.: Meeting capture in a media enriched conference room, *Proc. CoBuild'99* (*Springer LNCS1670*), pp.79–88 (1999).
17) Lamming, M. and Flynn, M.: "Forget-me-not" Intimate computing in support of human memory, *Proc. International Symposium on Next Generation Human Interface '94*, FRIEND21, pp.150–158 (1994).
18) Want, R., Hopper, A., Ao, V.F. and Gibbons, J.: The active badge location system, *ACM Trans. Inf. Syst.*, Vol.10, No.1, pp.91–102 (1992).
19) Matsuguchi, T., Sumi, Y. and Mase, K.: Deciphering interactions from spatio-temporal data, 情報処理学会研究報告，ヒューマンインタフェース，Vol.HI102 (2003).

(平成 15 年 4 月 15 日受付)
(平成 15 年 9 月 5 日採録)

角 康之（正会員）

1990 年早稲田大学理工学部電子通信学科卒業．1995 年東京大学大学院工学系研究科情報工学専攻修了．同年（株）国際電気基礎技術研究所（ATR）入所．2003 年より京都大学大学院情報学研究科助教授．博士（工学）．研究の興味は知識処理システムとヒューマンインタフェース．

伊藤 禎宣（正会員）

2003 年北陸先端科学技術大学院大学博士（知識科学）課程修了．同年より（株）ATR メディア情報科学研究所客員研究員．知識処理システムやユビキタスデバイスの研究に従事．博士（知識科学）．人工知能学会，ACM 各会員．

松口 哲也

2002 年マサチューセッツ工科大学理学部生物学科，化学科卒業．2002 年〜2003 年 ATR 学外実習生．2003 年よりカリフォルニア大学サンフランシスコ校大学院理学研究科生物化学専攻．

Sidney Fels

Sidney has been in the department of Electrical & Computer Engineering at the University of British Columbia since 1998. Sidney received his Ph.D. and M.Sc. in Computer Science at the University of Toronto in 1990 and 1994 respectively. He received his B.A.Sc. in Electrical Engineering at the University of Waterloo in 1988. He was a visiting researcher at ATR in Kyoto, Japan, from 1996 to 1997. He also worked at Virtual Technologies Inc. in Palo Alto, CA in 1995. His research interests are in human-computer interaction, neural networks, intelligent agents, new interfaces for musical expression and interactive arts.

間瀬 健二（正会員）

1979 年名古屋大学工学部電気工学科卒業．1981 年同大学大学院工学研究科修士（情報）課程修了．同年日本電信電話公社（現在 NTT）入社．以来，コンピュータグラフィックスおよび画像処理，そのヒューマンインタフェースへの応用の研究に従事．1988 年〜1989 年米国 MIT メディア研究所客員研究員．1995 年〜2001 年（株）ATR 知能映像通信研究所第二研究室室長．2001 年〜2002 年 ATR メディア情報科学研究所第一研究室室長．2002 年より名古屋大学情報連携基盤センター教授．博士（工学）．