2003－UBI－2　(41)
2003／11／19

複数センサを利用したインタラクション・パターンの自動抽出

高橋 昌史[†‡], 伊藤 禎宣[‡], 角 康之[†‡], 間瀬 健二[‡§]

筆者らのグループは，複数のセンサ群を用いて人と人，人と物，人と環境の間のインタラクションを自動で抽出することで，インデックス付きデータで構成される機械可読性の高いインタラクション・コーパスを構築する手法を開発している．本稿では，ユーザの注視状況や移動情報，発話状況に関してセンサ群から得た情報をもとに，ボトムアップによる処理によってインデックスとなり得るインタラクションの要素を抽出する方法を提案する．構築されたインタラクションのインデックス情報をデータベースに蓄えることで，コーパスの利用者が必要なデータを柔軟に問い合わせることが可能となり，蓄積情報の可用性を高めることができる．

Automatic Extraction of Interaction Patterns with Multiple Sensors

Masashi Takahashi[†‡], Sadanori Ito[‡], Yasuyuki Sumi[†‡], Kenji Mase[‡§]

We are developing a machine readable interaction corpus that consists of indexed data by automatically extracting human interactions from stored data using various sensors. In this paper, we propose a method to infer interaction primitives that can be significant indexes from user's conditions of gazing, wandering, staying and utterance. Storing raw data in databases with these indexes, we can develop a more useful corpus and enable corpus users to acquire appropriate data easily.

1　はじめに

近年，コンピュータはさまざまな形態で人間社会に浸透しており，我々はそれらのコンピュータに取り囲まれた環境の中で生活している [1][2]．しかし，コンピュータに対して何か要求がある場合，明示的に入力を行う必要があり，人間とコンピュータがうまく共生できているとは言い難い．我々の生活空間を包み込むコンピュータが人間社会に参加し，人と人のインタラクションを見守るような存在となるために，人間の身体動作などからコンピュータがその意図を自動で汲み取ることが目標とされる．その実現のためには人と人，人と物，人と環境の間のインタラクションをモデル化する必要があり，その研究インフラとしてインタラクション・コーパスが有効であると考えられる．そのため，我々は映像，音声，視線情報，生理情報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きた [3]．インタラクションの構造を体系化し，記録された生データに対してインデックスをつけることで，さらに有用性の高いコーパスを構築することができる．そこで，本研究ではユーザの注視状況や移動情報，発話状況に関してセンサ群から得た情報をもとに，インデックスと成り得るインタラクションの要素を抽出する．

解釈の時間的空間的広がりに応じた階層を設け，より広範囲の参照が必要な抽象的インデックス（イベント）を上位層，狭範囲で即時的抽出が可能なインデックス（プリミティブ）を下位層とする．下位層から上位層へボトムアップ的にインデックスの抽象化とデータベースへの記録を行う．これにより，インタラクション・パターンの即時的利用が必要なアプリケーションと，抽象度の高い分析的アプリケーションの双方にとって利用性の高いインタラクション・コーパスを提供することが可能になった．

2　インタラクション・パターンの自動抽出

我々は，開放的な空間における複数人の自由なインタラクションをさまざまなセンサ群で記録する試みを進めている．そのテストベッドとして，2003 年 11 月 6 日，7 日に開催される ATR 研究所の研究発表会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てインタラクション・パターンを抽出し，自動でインデックスを付与するシステムの試作を行った．ここでは，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

† 京都大学:Kyoto University
‡ ATR メディア情報科学研究所:ATR Media Information Science Laboratories
§ 名古屋大学:Nagoya University

く，自律的に動作するロボットを導入し，積極的に人とのインタラクションを演出する．また，さまざまな種類のアプリケーションが蓄積情報を参照できるように，汎用性の高いインデックスを付与する必要がある．

これまでにも，人と人，人と物のインタラクションにインデックスの自動付与を行う研究が行われてきた．会議場内で発話者の音源位置から映像の自動切替えを行う [4] では，会議の場面で有意とされるインタラクションを抽出し，蓄積された映像の可用性を高めている．しかしインデックスの付与ルールが，座席などが固定された会議場におけるインタラクションに限られるため，別の状況に適用することが困難である．例えば，開放的な空間では多数の人が一つの空間を共有するが，集団でインタラクションを行う際には，いくつかの意味的な単位のグループに分割できると考えられる．本研究ではこのような単位をイベントとして抽出し，場面を特定するのに有用なインデックスとして付与する．このように，蓄積されたデータに汎用性の高いインデックスの付与を行うことで，コーパスとしての有用性を高めている．

センサとしてはカメラ，マイクに加え，視覚内の対象物の認識・位置測定を行うために，赤外線 ID タグと，それを認識する赤外線 ID トラッカ [5] を利用している．ユーザの顔の向きに一致させてビデオカメラと赤外線 ID トラッカを装着することで，視野内のどこに何が映っているかを実時間で記録することができる．また，マイクは周りの雑音を容易に拾ってしまうため装着者が発話を行ったかどうかを知ることができない．そのため，人に対しては発話のボリュームを測定するスロートマイクを取り付けた．これにより，閾値処理を施すことで装着者が 発話した/発話していない の判定を行うことが可能である．現在，人に対しては装着型センサセット (マイク，カメラ，ID トラッカ，スロートマイク) を，ロボットやユビキタスに対しては設置型センサセット (マイク, カメラ，ID トラッカ) を用意している．このように，インタラクションの参加者だけでなく，環境側に対してもそれぞれ複数のセンサを設置して多角的にデータを記録する．

3 ボトムアップによるインタラクション・インデックスの抽象化

以上で述べたセンサ群によって記録されたデータを用いて，ボトムアップによる抽象化を行うことでインタラクションを自動抽出し，生のデータに対して段階的にインデックスを付与する手法について説明する．本研究では，注視 (gazing) が人のインタラクションに対してインデキシングを行うのに有効な手段であるとしている [6]．すなわち，人が何かに対して働きかけを行う際には，その相手を自分の視界の中に捕らえているものとする．従って，基本的にインタラクションは赤外線 ID トラッカが赤外線 ID タグを捕らえるという単純な要素により抽出されることになるが，赤外線 ID トラッカと赤外線 ID タグの装着対象の組み合わせや他のセンサによる情報次第でさまざまなインタラクション状況を推察することが可能である．

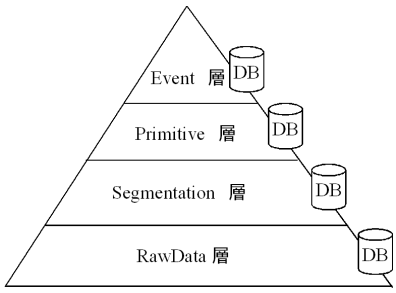

図 1: インタラクション・インデックスの抽象化

ここで，次の用語を定義しておく．

- **オブジェクト** センサやタグを装着するすべての対象物を指す．
- **プリミティブ** 1 体のオブジェクトと 1 体の相手オブジェクトの間のインタラクション．
- **イベント** 時間的・空間的に共有される複数のプリミティブを連結して意味のある単位にしたもの．

本研究では，図 1 のような階層的モデルを設定し，ボトムアップによる処理を行うことでインタラクションを抽出する．このモデルでは，各階層にそれぞれ図 2 のようなデータベースを保持し，上位層になるほど抽象度の高い情報を得ることができる．しかし，抽象度の高い情報を得るには時間的，空間的に幅の広いデータが必要になるため，上位層のデータベースほど，更新されるタイミングに遅延が発生することとなる．そのため，より即時性が求められる場合にはより下位層へとアクセスを行い，より抽象度の高い情報が必要な場合はより上位層へとアクセスを行うことで，柔軟に問い合わせを行うことができる．

オブジェクトには一意の ID(オブジェクト ID) とその型 (人，ロボット, ユビキタス など) が指定される．さらに，センサセットに対してはセンサ ID を，赤外線 ID タグに対してはタグ ID を一意に割り付け，これらを装着しているオブジェクトの ID と関連付ける．これ

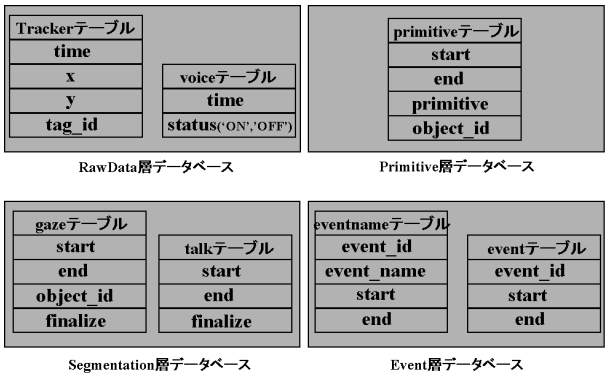

図 2: 各階層に設定されるデータベース群

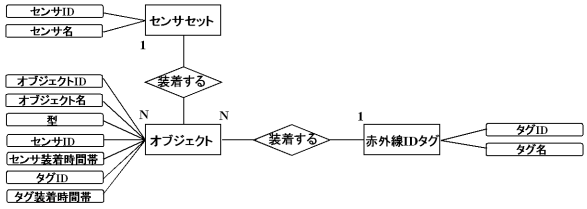

図 3: オブジェクトとセンサ，タグに関する ER 図

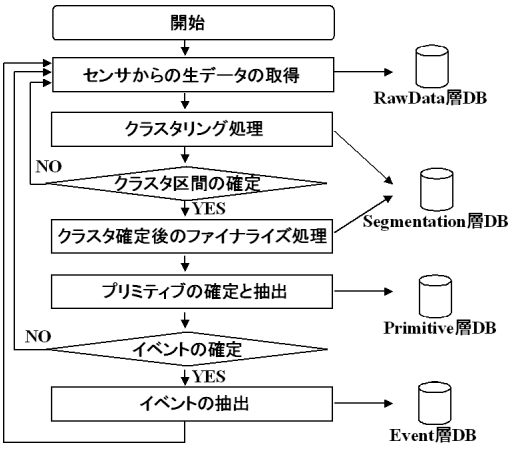

図 4: 処理のフロー

らの関係を図 3 の ER 図に示す．階層構造を用いた本システムにおける処理のフローは図 4 のようになる．

次に，各階層の詳細について，簡単に説明を行う．

3.1 RawData 層

最下層である RawData 層では，センサによって記録された生のデータを格納する．これらのデータは断続的に入力され，時刻と観測値のペアという形式で格納される．赤外線 ID トラッカによるデータに対しては，検出されたタグの ID と時刻，2 次元座標が tracker テーブルに逐次書き込まれる．また，スロートマイクによるデータに関しては，発話ボリュームに対して閾値処理を行った後，発話の開始時刻と終了時刻が voice テーブルに書き込まれる．これらのテーブルはセンサ ID ごとに用意される．現在は以上のセンサを用いてシステムの実装を行っているが，観測値が断続的に入力される形式であれば同様に扱うことが可能であるため，今後生体データ記録用モジュールなどの利用を考えている．

RawData 層では，センサがデータを検出すると即座にデータベースの更新を行うため最も即時性が高いが，検出された生データが加工されずに格納されるためデータ量が多く，利用用途が限定される．

3.2 Segmentation 層

RawData 層におけるデータはセンサの特性上，断続的なものである．例えば赤外線 ID トラッカの場合，データ間の間隔が最低でも 100 ミリ秒，スロートマイクの場合は最低でも 3 秒空いてしまう．従ってこれらのデータの間を埋める必要がある．

Segmentation 層では，RawData 層の生データに対して時間でクラスタリングを行い複数の区間に分割することで，動作の主体となるオブジェクトが注視を行った区間を推定する．さらにセンサ ID やタグ ID に対してそれらの装着対象オブジェクトの ID との対応付けを行い，動作の主体となるオブジェクトとその相手となるオブジェクトを決定する．

まず，赤外線 ID トラッカによる断続的な生データ入力に対してクラスタリングを行い，gaze テーブルに格納する．ここでは，図 5 のように定数 Max_Interval を設定し，赤外線 ID トラッカに Max_Interval 以上の間隔を空けずに赤外線 ID タグが検出され続けた区間を抽出する．この結果，断続的に得られるセンサデータの間を埋めることができ，さらにトラッカの認識頻度が多少低くても妥当な結果が得られる．図 5 では，センサから得られた多数の離散的データが 6 個の区間に分割されていることがわかる．

また，クラスタの区間が確定してからデータを書き込むと時間的な遅延が発生するため，注視の状況を随時提供することができるように完了フラグ finalize を導入し，クラスタに属する最初のデータが記録されてから区間が確定するまでの間は完了フラグを偽に設定す

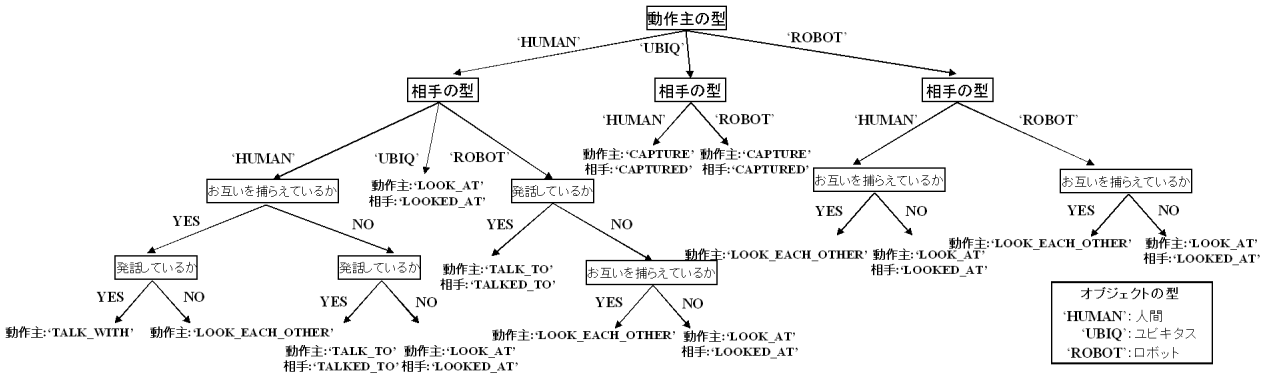

図 6: オブジェクトのプリミティブの決定木

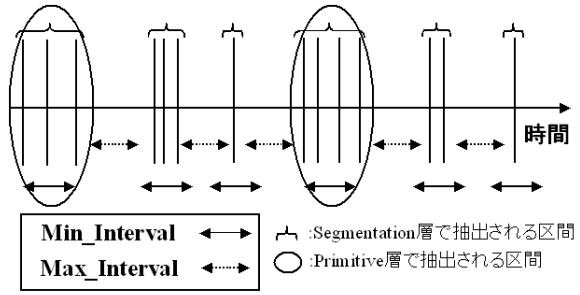

図 5: Max_Interval と Min_Interval

ることで，オブジェクトが注視を行っている状態であることを表すことができる．この場合，区間の確定には完了フラグが真に設定されるまでの Max_Interval だけの遅延が生じることになる．スロートマイクによるデータに対しても同様にクラスタリングを行い，talk テーブルに格納する．これらのテーブルはオブジェクト ID ごとに用意される．

3.3 Primitive 層

続いて Primitive 層では，Segmentation 層においてクラスタリングが完了した区間に対して，オブジェクトのプリミティブを推定する．それぞれの区間では，動作の主体となるオブジェクトが他の 1 体のオブジェクトを注視しており，お互いのオブジェクトの型や発話状況などにより，その区間において両者にどのようなインタラクションのプリミティブがあるかを推定する．

ここでは，図 5 のように定数 Min_Interval を導入し，クラスタリングが完了した区間のうち，長さが Min_Interval 以上のものだけを抽出する．これにより，一瞬だけ視界に入ったようなオブジェクトは意味のないものとして注視対象から排除することができる．図 5 では，Max_Interval の導入によって分割された 6 個の区間のうち，長さが Min_Interval 以上の 2 区間のみが抽出されている．

続いて，長さが Min_Interval 以上の区間に対して，図 6 のような決定木を用いてオブジェクトのプリミティブを推定する．例えば動作主である人 A が身につけている赤外線 ID トラッカが，別の人 B が身に着けている赤外線 ID タグを捕らえた場合，同じ時間帯に相手 B の赤外線 ID トラッカも動作主 A の赤外線 ID タグを捕らえていて，さらに A と B の少なくとも一方が発話を行っていれば，両者の間で会話が交わされているものと推定され，動作主のプリミティブは 'TALK_WITH' となる．primitive テーブルはオブジェクト ID ごとに作成し，プリミティブの種類と時間帯，相手オブジェクトの ID が格納される．

Primitive 層のデータベースに対しては，クラスタリングが完了すると即座に更新される．従って，実際にインタラクションを行ってからデータベースに反映されるまでには Max_Interval だけ時間の遅延が生じることになる．

3.4 Event 層

最上位層である Event 層では，時間的・空間的な共有性を有する複数のプリミティブを連結することで，インタラクションを行っているオブジェクトのグループを特定し，イベントとして抽出する．抽出されたイベントはそれぞれ一意の ID(イベント ID) が割り付けられ，その時間帯とともに eventname テーブルに格納される．また，event テーブルはオブジェクト ID ごとに用意され，該当オブジェクトが参加したイベントの ID と，それに参加した時間帯が格納される．図 7 に，いくつか

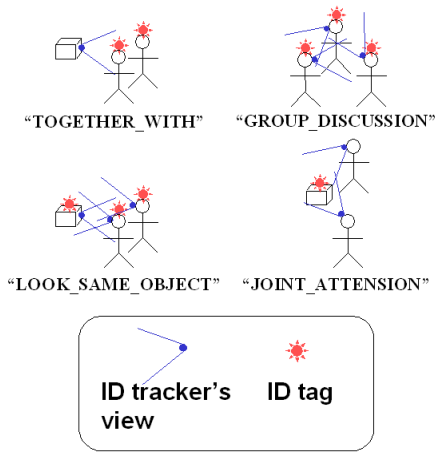

図 7: さまざまなイベントの解釈例

のイベントの解釈例を図解する．例えば人 A と人 B が会話を行っている場合，その時間帯の近くで人 B と人 C が会話を行っていれば，3 人はグループ討論を行っているとして，イベント "GROUP_DISCUSSION" が発生する．

イベントの検出にはプリミティブの蓄積を待つ必要があるため，Event 層におけるデータベースの更新に関してはすべての階層の中で最も遅延が発生することになる．

4 各アプリケーションからの利用

ボトムアップにより段階的に抽象化を行うことで，さまざまな抽象度のインデックスが付与されたコーパスを構築できることを示した．利用者がその目的に応じてインデックスを使い分けられる具体例として，ATR 研究所の研究発表会におけるポスター展示会場で実際に用いられるアプリケーションから本システムのデータベースへのアクセス例について紹介する．

- **見学者に対する HMD への情報の提供**
  見学者に HMD(Head Mount Display) を装着してもらい，各ブースの盛況状況や別の見学者に関する情報を提示するシステムを開発した．ここでは，見学者が見たものと関連する情報をただちに提示する必要があるため，主に時間的遅延の少ない Segmentation 層へアクセスを行い，人の注視対象を特定した．

- **見学者の体験履歴によるビデオサマリの生成**
  ハイライトシーンを自動的に切り出したり，同一シーンのビデオクリップの中でカメラの切り替えを自動化することで，見学者の記録をその場で短いビデオ（ビデオサマリ）に自動要約するシステムの開発を行った．ここでは，抽象度の最も高い Event 層へアクセスしてシーン長と参加者を抽出し，Primitive 層を参照することでインタラクションの状況に沿ったカメラの切り替えを行った．

- **ロボットのビヘイビアへの適用**
  人と積極的にインタラクションを行うロボットの動作を決定するビヘイビアの開発を行った．ロボットは目の前にいる人のこれまでの行動履歴を参照し，それに関する問いかけなどを行う．ここでは，Segmentation 層により注視対象を特定し，Primitive 層によりその行動履歴を参照した．

現在，Primitive 層や Event 層は社会学者や認知学者の知見を反映してインデックスの構成を決定している．しかし，より適切なインタラクションコーパスの構築には，RawData 層や Segmentation 層に蓄積されたデータを参照し，新たなイベントインデックスを考えることも必要である．このように，コーパスの利用者目的に応じたアクセスを行うことにより，時間的・情報量的に無駄の少ない処理が可能となる．

5 実験と考察

研究発表会に先がけて，小規模な環境下における人のインタラクションを自動抽出する実験を行った．オブジェクトを 10 体 (うち人間 4 体，ユビキタス 6 体) 導入し，2 種類のブース (各ブースに 3 体のユビキタスを設置) を設けて次のようなシナリオでそれぞれ行動を行った．

Scene1 ユーザ A がブースを閲覧している．
Scene2 ユーザ B が近寄りユーザ A と会話を始める．
Scene3 ユーザ C が会話に加わる．
Scene4 ユーザ D が会話に加わる．
Scene5 各人が別々の展示物を閲覧し始める．
Scene6 ユーザ A とユーザ B が会話を始める．
Scene7 ユーザ D が会話に加わる．
Scene8 ユーザ C が会話に加わる．

全 2737 秒に渡る一連の行動の結果，蓄積されたデータ件数は RawData 層で 9950，Segmentation 層で 1059，Primitive 層で 640 に至り，最終的に 7 つのイベントが検出された．そのうち，イベント "GROUP_DISCUSSION" が検出された区間を図 8 に示す．ここでは，横軸は時間 (秒) を表す．この結果，3 人以上で会話を行っている時間帯に集中して

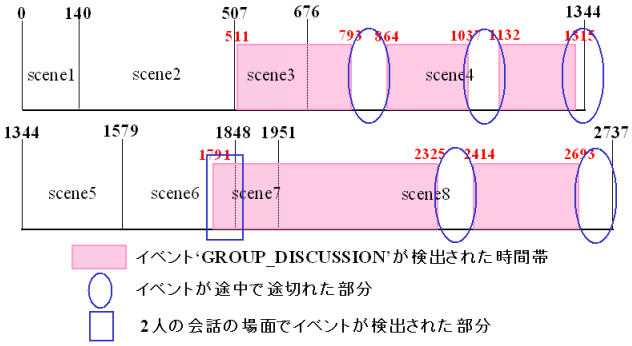

図 8: イベント “GROUP_DISCUSSION” の抽出結果

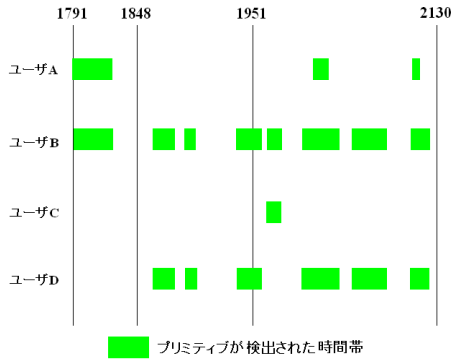

図 9: プリミティブの抽出結果例

イベントが検出されているため，ほぼ妥当な結果が得られたと言える．途中でイベントが途切れる部分が幾つか存在するが，これらは人が皆ポスターに集中していた区間である．このことから，環境側センサなどを利用してユーザ位置情報を得，イベント “GROUP_DISCUSSION” の継続判定に付加的に利用する必要があると考えられる．一方，ユーザ A とユーザ B が 2 人で会話を行っている区間に対してイベント “GROUP_DISCUSSION” が検出されている部分がある．この区間の一部を抽出し，各人オブジェクトに対して検出されたプリミティブのうち，相手が人間であるもの (‘TALK_WITH’,‘LOOK_EACH_OTHER’ など) の発生時間帯を図 9 に示した．この場面では，ユーザ A とユーザ B が話していたところにユーザ C とユーザ D がやってきて会話に参加する場面であるが，発生した複数のプリミティブが時間的に近かったため，それらのシーンが同一のイベントとして抽出されている．しかし，イベントの発生時間帯の中で各人が参加した時間が別の情報として保持されているため，イベント全体へのインデックス付与という目的では問題ないと思われる．

6 おわりに

インタラクションを自動で抽出してインデックスを付与することで，利用性の高いコーパスを構築する手法を提案した．さらに試用実験を行い，妥当なインデックス付きコーパスを構築できることを確認した．現在は，展示会場におけるインタラクションを想定したシステムの実装を行っているが，ボトムアップによる抽象化を行うことで生データにインデックスを付与する本手法は，場所やセンサの種類に依存することなく利用できるため，将来的にはさまざまな環境においてインタラクションのパターン収集と抽出を行いたい．

謝辞

本研究を進めるにあたり，多分のご意見，ご協力を賜りました中原淳氏，坊農真弓氏をはじめとする ATR メディア情報科学研究所の皆様に感謝いたします．また，この研究の機会を与えて頂いた，片桐恭弘所長，萩田紀博所長に感謝します．本研究は通信・放送機構の研究委託により実施した．

参考文献

[1] Cory D. Kidd, Robert Orr, Gregory D. Abowd, Christopher G. Atkeson, Irfan A. Essa, Blair MacIntyre, Elizabeth Mynatt, Thad E. Starner , and Wendy Newstetter. The aware home: A living laboratory for ubiquitous computing reserch. In Proceedings of CoBuild’99 (Springer LNCS1620), pp. 190-197, 1999.

[2] Barry Brumitt, Brian Meyers,John Krumm, Amanda Kern, and Steven Shafer. EasyLiving:Technologies for intelligent environments. In Proceedings of HUC 2000 (Springer LNCS1927), pp. 12-29, 2000.

[3] 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，情報処理学会，2003.

[4] Ross Cutler, Yong Rui, Anoop Gupta, JJ Cadiz, Ivan Tashev, Li-wei He, Alex Colburn, Zhengyou Zhang, Zicheng Liu, and Steve Silverberg. Distributed Meetings: A Meeting Capture and Broadcasting System, In Proceedings of ACM Multimedia 2002, 2002.

[5] 伊藤 禎宣，角 康之，間瀬 健二．赤外線 ID センサを用いたインタラクション記録装置，情処研報，ヒューマンインタフェース，vol.HI104, 2003.

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