2B1-02

インタラクション解釈における階層構造の検討

A Layered Structure of Human Interaction Interpretations

高橋 昌史[*1*2] Masashi Takahashi　伊藤 禎宣[*2] Sadanori Ito　土川 仁[*2] Megumu Tsuchikawa　角 康之[*1*2] Yasuyuki Sumi　間瀬 健二[*2*3*4] Kenji Mase　小暮 潔[*4] Kiyoshi Kogure

[*1]京都大学大学院情報学研究科 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]名古屋大学情報連携基盤センター Information Technology Center, Nagoya University
[*4]ATR 知能ロボティクス研究所 ATR Intelligent Robotics and Communication Laboratories

Although computers have widely prevailed in our life in various forms recent years, it is still necessary to input the demand to the computers using such interfaces as mouses and keyboards if we want to do something with them. As we can't always use thses interfaces in our daily life, we need to develop a new interface using the entire body of person to appropriately coexist with these computers. To construct such an interface, it is necessary that computers understand person's demands or situations from his bodily movement. So we are developing a interaction corpus that consists of multimodal data in various situations using wearable and ubiquitous sensors in order to analyze person's interaction structurally. In this paper, we propose a bottom-up method to extract interpretation indices using multiple data about person's conditions of gazing, utterance, staying and wandering to make the corpus more useful.

1. はじめに

近年，様々な形態でコンピュータが我々の生活に浸透してきているが，人がこれらのコンピュータを利用して何かを行いたい場合，マウスやキーボードといった，人間の手先を使ったインタフェースを利用してその要求を明示的に入力する必要がある．しかし現実には常に手先を利用できる状態ではないため，これらのコンピュータとより適切な形で共存するためには，GUI といった現在主流のマンマシンインタフェースの見直しを行い，より直感的にコンピュータと触れ合うことができるようなインタフェースを構築することが求められる．こういったインタフェースを実現するためには，コンピュータが人間の身体動作などからその状況や意図を理解することが求められるが，まずは人と人，人と物のインタラクションについて分析を行うことが必要である．そこで筆者らのグループでは，人のインタラクションの分析を行うことを目的として，人が装着するウェアラブルなセンサユニットに加えて環境に遍在するセンサ群を利用することで人の行動を多角的に観測し，映像，音声，注視情報，生理情報など，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蓄積することでインタラクションのコーパスを構築する試みを進めてきた [1].

しかし，ただセンサ群を利用してデータを蓄積するだけでは再利用性に乏しいため，人の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，さらに可用性の高いコーパスを構築することができる．そこで，我々は展示会と会議，講義といった，複数の状況 (我々はドメインと呼んでいる) において日常的に複数人のインタラクションを記録し，有用性の高いインデックスの付与を行う試みを進めている．

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付与を行う研究が行われてきた．例えば，会議場内で発話者の音源の位置から映像の自動切換えを行う [2] 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会議場におけるインタラクションに限られるため，別の環境に適用することができない．本研究では，会議に限らず幅広い環境においてインデックスを付与できるようなシステムを構築することを目指している．また，講義の内容と受講者の視線の関係を構造的に調べた [3] では，遠隔講義における映像選択の指針を得るために，講義の内容の変化に伴って受講生の注視行動がどのように変化するのかを明らかにした．ここでは，映像を利用して受講者の注視状況を人手でタグ付けしているが，我々はこういったことを自動的に行うことを目指している．

本稿では複数のドメインで人のインタラクションに対してインデキシングを行うために，解釈の抽象度に応じた階層を有するモデルを設定し，センサによる生のデータを利用して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．まず，断続的なデータに対してクラスタリングを行って複数の連続区間に分割し，さらに「目線を合わせた」「話しかけた」といった，インタラクションの基本単位となるような解釈を行う．さらに，それらの解釈を組み合わせて「討論」「質疑応答」といった複合的なインタラクションの解釈を行うが，例えば「討論」というインデックスは，開放的な空間で人が自由に動き回れる展示会場では意味のある情報であるといえるが，固定された座席に座って絶えず一つの議題について話し合う会議の場では意味があるものとは言えない．従ってドメインの特徴やインデックス情報の利用用途について十分考慮した上で複合的な解釈を行わなければならない．そこでコーパス構築のテストベッドとしてポスター展示会と会議，講義を選び，各ドメインの特徴について洗い出しを行って複合的にインタラクションを抽出する規則について検討を行った．

連絡先: 高橋 昌史，京都大学大学院情報学研究科，京都市左京区吉田本町，takahashi@lab1.kuis.kyoto-u.ac.jp

2. 複数センサ群によるインタラクションの記録

センサ部は運用形態に応じて柔軟な構成の変更が可能であり，図 1 のように，ウェアラブルなヘッドセットタイプのものと，設置型の据え置きタイプのものから構成される．カメラ，マイクに加え，視野内の対象物の認識・位置測定を行うため

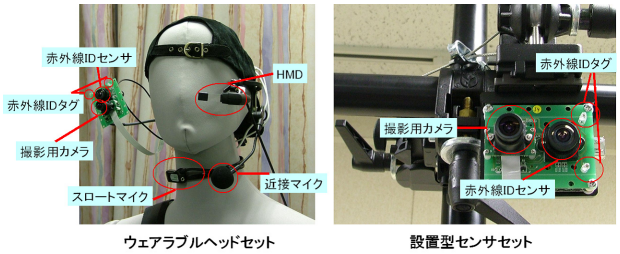

図 1: センサセット

に，赤外線 ID タグシステムを利用した．赤外線 LED の点滅パターンによって固有の ID を発信できる赤外線 ID タグを認識対象に取り付け，それを認識する赤外線 ID センサをユーザの顔の向きに一致させて装着することで，視野内のどこに何が存在するかを実時間で記録することができる．また，喉に取り付けて声帯の振動から発話のボリュームを測定することができるスロートマイクを利用した．これにより，閾値処理を施すことで装着者が発話しているか否かの判定を行うことが可能である．これらのセンサ群を協調的に利用することで，人のインタラクションを多角的に観測し，映像，音声，注視情報，発話情報からなるインデックス情報付きのインタラクション・コーパスの構築を行う．

3. インタラクションの階層構造

我々は，人の注視 (gazing) と発話 (utterance) 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考えており，これらの情報に基づいて人のインタラクション情報の抽出を行う．本稿では，図 2 のように階層的なモデルを設定し，センサによる断続的な生のデータに対してボトムアップアプローチを行うことで，インデックス情報を段階的に抽象化するといった手段を利用する．各階層では，その解釈の抽象度に応じた機械可読なインタラクションのインデックス情報を蓄積し，階層ごとに用意されたデータベースに記録される．

我々は，人のインタラクションがその解釈の抽象度に合わせた階層を有すると考えており，例えば，「討論」や「質疑応答」といった，各ドメインに依存するような解釈もあれば，「目線を合わせた」「話しかけた」といった，すべてのインタラクションの基本単位となるような解釈も存在する．しかし，「討論」といった解釈も，複数のインタラクションの基本単位から構成されるため，HAS-A 関係に基づく階層構造が存在することになる．そこで，こういったインタラクションの基本単位となる解釈 (Primitive) を下位層に，状況依存な複合的解釈 (Composite) を上位層に設定して階層的にモデル化を行い，下位層から上位層へボトムアップ的に解釈の抽象度を上げていくことで，ドメイン間のインタラクション解釈の違いを吸収し，複数のドメイン下でもインデキシングを行うことができるようなシステムを提案する．

まず，最下層である RawData 層では，センサによって記録された断続的な生のデータを格納する．これらのセンサによるデータは，時刻と観測値の組という形式で記録される．第 2 階層である 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連続区間に分割することで，動作主体が注視と発話を行っていた区間を推定する．第

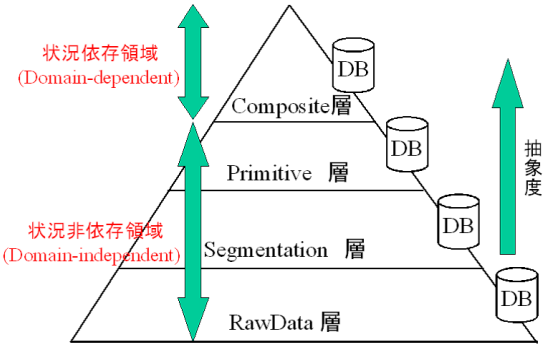

図 2: インタラクションの階層的モデル

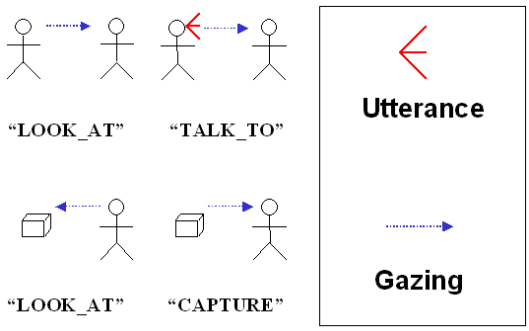

図 3: Primitive の例

3 階層である Primitive 層では，図 3 のような人のインタラクションの基本単位となる情報 (Primitive) を記録する．例えば人が対象物を視界の中に捕らえれば，人が注視を行ったとして「LOOK_AT」といった Primitive 情報を抽出できるし，さらに人に対して語りかけていれば「TALK_TO」といった情報が得られる．また，物や環境に設置されている ID センサに人が捕らえられると，人がその場所に存在することがわかるため，「CAPTURE」といった Primitive 情報が得られる．この階層までが，ドメインに依存しない領域 (domain-independent 領域) となり，インタラクション・コーパスではあらかじめこういった情報を用意しておく．続いて，最上位層である Composite 層では，「討論」「質疑応答」といった，各ドメインに依存する複合的なインタラクション情報 (Composite) を記録する．ここでは各アプリケーションがその用途に応じてインタラクションの解釈を行うことができる領域 (Domain-dependent 領域) となる．以上のように，下位の階層におけるインタラクションの解釈は，より上位の階層における解釈の一部となるため，HAS-A 関係による階層関係が成立する．

従って上位の階層ほど解釈の抽象度が高くなるが，抽象度の高い解釈を行うためには時間的・空間的にも幅の広いデータが蓄積される必要があるため，上位の階層ほどインデックスの付与に必要な時間が大きくなる．このため，よりリアルタイム性が求められるシステムに対してはより下位層のインデックス情報を利用し，より抽象度の高い情報を必要とするシステムに対してはより上位層のインデックス情報を利用することで，幅広い応用システムを構築することが可能となる．

4. 複数のドメインにおけるインタラクションの記録

今回，以下のドメインにおいてコーパスの構築を行うシステムについて試作し，実験を行った (図 4).

- 展示会ドメイン
  2003 年 11 月 6 日，7 日に開催された ATR の研究発表会におけるポスター展示会場を舞台とし，展示者と見学者のインタラクションの記録を行った．展示者全員と，見学者のうち希望者に対してはウェアラブルヘッドセットを装着してもらい，展示物には赤外線 ID タグを設置した．さらに設置型のセンサセットを天井と壁に設置し，各展示ブースには正面と背面から見下ろすような角度で人や展示物を捉えた．

- 会議ドメイン
  同研究所で日常的に開催されているミーティングの場において，参加者同士のインタラクションを記録した．参加者にはラウンドテーブルに座ってもらい，それぞれウェアラブルヘッドセットを装着してもらった．

- 講義ドメイン
  同研究所で定期的に開催されている研究会において，プロジェクタや白板を利用したプレゼン発表会における公演者と観客のインタラクションを記録した．公演者と観客にはそれぞれウェアラブルヘッドセットを装着してもらい，プレゼン用のスクリーンや説明用の白板には赤外線 ID タグを設置した．また，会場の後ろと前から公演者と観客を見下ろすような角度で設置型のセンサセットを設置した．

Composite 層ではアプリケーションの用途に応じて複合的なインタラクションの解釈を行うことを前節で述べたが，今回，その実証実験として，人の行動履歴からハイライトとなるシーンを抽出してそれらを一本の短いビデオに要約したサマリビデオを各ドメインにおいて自動生成するシステムを構築するために，ハイライトシーンとなり得るインタラクション情報を Composite 層において抽出することを試みた．まず，その抽出規則を決定するために各ドメインに対して以下の項目についての検討を行い，ドメイン間の特徴の違いを洗い出した．

- 人のグルーピング (Gr)
  人がインタラクションを行うグループが動的に変化するか否かによって比較を行う．

- 場所の有意味性 (Loc)
  インタラクションが起こる場所に意味があるか否かによって比較を行う．

- 会話場の数 (Conv)
  全体でいくつの会話場が生じているのかによって比較を行う．

- 人の役割交代 (Ro)
  インタラクションの場における人の役割が動的に変化するか否かによって比較を行う．

各ドメインに対して，以上の項目について考察を行う．

展示会ドメイン

会議ドメイン

講義ドメイン

図 4: 複数ドメインにおけるコーパスの構築

- 展示会ドメイン
  自由に相手を選んで会話を行ったり自由に展示物を閲覧することができるため，インタラクションのグループが動的に変化するし，どこでそれを行っているかによって人の興味や話題を推察することができるため，場所の有意味性は高いと考えられる．また，会場のあちこちで会話が成立するため会話場の数が多数存在することも明白であるし，人に対しても展示物を閲覧することを目的に来訪する見学者と，展示物を説明する展示者といったようにあらかじめその役割について明確に区別することができる．

- 会議ドメイン
  参加者は固定された座席に座って討論を行うため人のグルーピングが静的で場所の有意味性は低いと考えられる．また，ミーティングの参加者が同一の議題について話し合うため会話場の数は 1 としてもよいであろうし，参加者はその時々によって発話者と聞き手といったように動的にその役割を交代すると考えられる．

- 講義ドメイン
  観客は固定された座席に座って公演者の話を聞くため人のグルーピングは静的であるが，公演者が白板の前にいるのか，それともスクリーンの前にいるのかといった情報から講義の状況を推察することができるため，場所の有意味性は高い．また，会話場の数は 1 としてもよいであろう．

-	Gr	Loc	Conv	Ro
展示会	動	高	多	静
会議	静	低	1	動
講義	静	高	1	静

表 1: 各ドメインにおける特徴の比較

以上で考察した各ドメインの特徴について整理すると表 1 のようになる．これらの特徴を利用して，サマリビデオのハイライトとなり得る複合的なインタラクションの解釈を行う．ここではまず，規則を用いて時間的・空間的に共有性を有する primitive 情報を連結し，抽象度の高い解釈を加える必要があるが，各ドメインに対して表 1 の中からそれぞれ特徴的な要素に注目し，Primitive 情報の連結規則を決定する．

ここでは，具体的にポスター展示会と会議において，それらの特徴の違いを利用した Primitive 情報の連結規則について考察を行う．ポスター展示会では，人のインタラクションのグループが動的に変化し，さらに場所の有意味性が高い点に注目し，人の注視状況を手がかりとしてそのグループや場所を特定し，それに基づいてインタラクションの解釈を行う．展示会における Composite の例を図 5 に示す．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解釈して，”GROUP_DISCUSSION”というインタラクションの解釈を行う．また，物や環境に設置された ID センサに複数人が捕らえられた場合，その人達は同じ場所に滞在していることが推測されるため，”TOGETHER_WITH”といった解釈を行うことができるし，さらにその人達が同じ展示物を見ながら発話している場合には，お互いにその話題について会話を行っているとして”TALK_ABOUT”といったインタラクションの解釈を行うことができる．一方，会議ドメインでは，どの時間帯においても会話場が一つであるという特徴に注目し，人の発話状況を手がかりとして発話権を握っている人 (speaker) を特定し，それに基づいてインタラクションの解釈を行う．会議における Composite の例を図 6 に示す．例えば，発話者が一方的に話し続けている場合，発話者は演説を行っているとして”LECTURE”という解釈を行い，発話者の他にも議論に参加している人がいれば，発話者の話題について皆で討論を行っているとして”DISCUSSION”という解釈を行う．また，発話者が参加者の大半に注目されている場合，発話者は重要な発言を行っていると判断して”HOT_ATTENTION”といった解釈を行い，逆にほとんど注目されていない場合は”COLD_ATTENTION”といった解釈を行う．こういった情報はサマリビデオを生成する際に有用なインデックスとなる．

このように，Composite 層では各ドメインの特徴とインデックス情報の利用用途に応じて柔軟にインタラクションの解釈を行うことで，より有用性の高いコーパスを構築することができる．

5. おわりに

本稿では，複数の環境下で有用性の高いインタラクション・コーパスを構築するために，解釈の抽象度に基づいた階層的なモデルを利用して機械可読性の高いインデックス情報を抽出する手法について提案した．今後は，実装依存的な本モデルの評価を行うために，理論的なインタラクションのモデルを構築することを考えている．

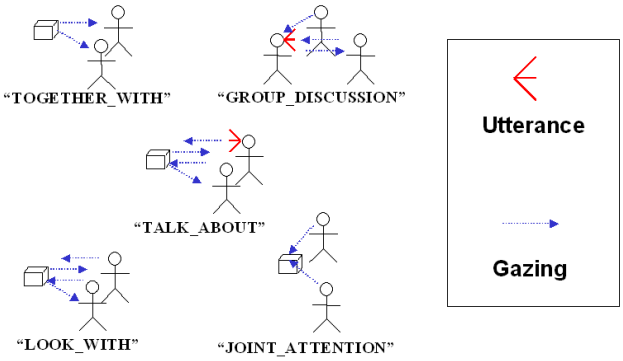

図 5: 展示会における Composite の例

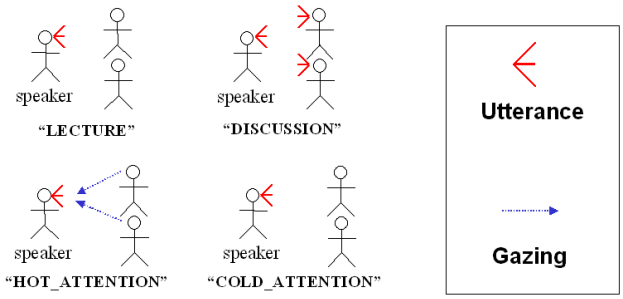

図 6: 会議における Composite の例

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，鈴木紀子氏，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，ならびに熊谷賢氏をはじめとする京都大学大学院情報学研究科知能情報学専攻の西田研究室の皆様に感謝する．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する．なお，本研究は情報通信研究機構の委託研究「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．

参考文献

[1] 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.

[2] 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.

[3] 村上正行，角所考，美濃導彦．マルチメディア一斉講義における内容に基づく受講生の注視行動の分析．人工知能学会論文誌，Vol. 17, No. 4, pp. 473-480, 2002.