

プレゼンテーション映像における話者の行動理解

† 中村 裕一, † 西谷 正志, † 大田 友一

† 筑波大学 電子情報工学系

〒 305 つくば市 天王台 1-1-1 (yuichi@is.tsukuba.ac.jp)

あらまし: ノンバーバルな情報伝達手段では、強く制限されない限り明確な法則や規約がないため、その解釈において、人間のコミュニケーションにおける曖昧性、状況依存性といったものを直接扱う必要が出てくる。本研究ではこの点を重視し、コミュニケーションという視点から人間の動作理解の問題をとらえ、身振りの持つ基本的な情報と文脈との関係を考察した。また、文脈の中で行動と最も密接な相互関係を持つ発話との相互関係を解析し、これを用いてプレゼンテーションにおける人間の身振り動作の役割を認識し、そこから有用な情報を抽出するための簡単な枠組を提案する。

キーワード: 人物行動理解, 発話, 動作識別, 文脈情報の利用, プレゼンテーション

Human Behavior Understanding in Oral Presentation

† Yuichi NAKAMURA, † Masashi Nishitani, † Yuichi OHTA

† Institute of Information Sciences and Electronics, University of Tsukuba
1-1-1 Tennodai, Tsukuba, 305 (yuichi@is.tsukuba.ac.jp)

Abstract: In reading gestures or postures, there is a variety of ambiguity which depends on situations around the behaviors, since few strict laws or rules to be definitely applicable throughout a wide variety of situations. In this paper, we consider the relationships between behaviors and context, among which the synchronized spoken dialog is most essential. By using these relationships, we propose a simple framework to understand the speaker's behaviors in oral presentation, and to extract useful information from the movements or the posture.

Keywords: Human Behavior Understanding, Dialog, Behavior Classification, Context Utilization, Presentation

1 はじめに

近年、ヒューマンインタフェースや人物行動の検出/監視などを目的として、人物の手足の動きを認識する研究や、人物の移動を検出する研究が盛んに行われている。これらの研究が主に目的とすることは、(1) 手足の位置、体の位置を計測することにより、ポインティングデバイスとして使用する。(2) 手足の形態などを認識することにより、記号として読みとる。それによって、手話の解釈を行ったり、計算機に特別な命令を与える。(3) 人物の歩行、作業などを追跡し、動きなどの特徴量を計測する、等である。

このように、従来の研究では人間の動作認識の問題が人体及びその周辺の物理量の計測の問題として扱われているが、人間の動作をコミュニケーション手段として認識/活用するという観点でみると、質的に異なった新しい扱いが必要となる。というのは、ノンバーバルな情報伝達手段には、物理量や自然言語¹のような明確な法則や規約があるわけではなく、その解釈において、人間のコミュニケーションの持つ曖昧性、状況依存性といったものを直接扱う必要が出てくるからである。

そのために、本研究ではプレゼンテーション映像を題材にし、話者が明確な意図を持って発話、行動している状況を、コミュニケーションという立場から解析する試みを行っており、本稿ではその基本的な考え方について述べる。以下では、コミュニケーションのための人物行動理解の基本的な考え方、発話内容(音声言語)の利用について、簡単な実験を交えて説明する。

2 人間の動作理解とコミュニケーション

2.1 コミュニケーションとモダリティ

人間は、絵図や身振りなどの画像的情報、発話やテキスト等の言語的情報、また、その他の情報を状況に応じて併用したり使い分けたりしている。ここで提示される情報は、お互いに参照されたり、説明されたりしながら、総合的に送り手の伝えたい意図を表現する。また同時に、明示的な情報ばかりでなく、非明示的な情報も与えられる(暗黙の了解)。そのうちのあるものは、他のモダリティ(あるいはメディア)によって与えられたり、複数モダリティ間の対応関係から与えられる²。

例えば、図1のように、話者が

「このリンゴは見ためほど甘くないんですよ。」

とリンゴを指さしながら言う時、発話内容、リンゴ(実空間での物体)、身振りといった情報が話者と聞き手の共有空間に導入され、りんご、りんごの外観、甘さといった情報や概念が関係づけられる。さらに、その前提となること、例えば、目の前の空間にりんごが実在すること、指さした

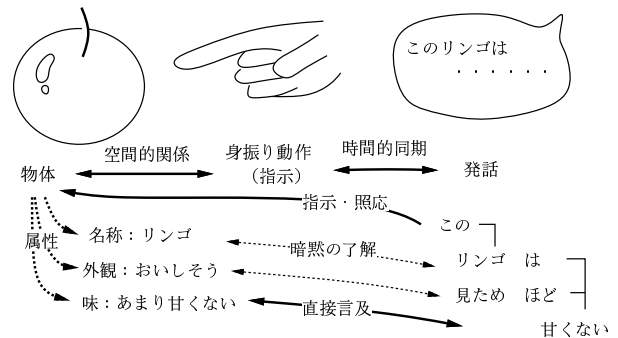


図1: コミュニケーションとモダリティ

物体がりんごであること、そのりんごが甘く見える外観を持っていること、その他多くの事柄が暗黙的に示される。

このような送り手側の意図を人間が簡単に理解できるのは、豊富な経験や知識を用いていること以上に、その場で利用可能な他の情報(文脈)をうまく利用していることがあげられる。つまり、送り手から与えられた情報をお互いに関連づけ、それによって、さらに必要な情報を取り出すことが、マルチモーダルなコミュニケーションの必須要素となっている。

本研究では、以上のようなコミュニケーションの場において、人間の身振り動作の伝える情報を抽出することを目的としている。そのために必要となることとして、まず、人間の身振りの持つ意味を考え、それを有効に利用することについて考える。

2.2 身振り動作の心理学的分類

人間の動作の心理学/記号学的分類としては、Ekmanらの分類が有名であり、多くの研究でこれをもとにしている[EF69, 黒川94]。

標識(emblem): 言語翻訳可能な動作。emblemはgestureと言い換えることも可能な動作であり、言語と等価な性質を持つ動作と言える。代表的な例としては、手話やサインなどがあげられる。

例示子(illustrator): 発話内容の補足強調。発話に付随する空間的な表現や指示動作など。以下の6種類。

指示動作(deictic movements): 対象を指し示す

空間動作(spatial movements): 空間的にアナログ量を提示する

象形動作(pictographs): 対象物を空間に描く

活動動作(kinetographs): 特定の動きを描写する

思考動作(ideographs): 思考過程を描く

バトン(baton): 語句の強調や、発話のリズムをとる

この内、最初の二つは会話を聞かなければ全く意味を持たない。残りは会話の内容を補う性質を持ち、会話を聞かなくても送り手の何らかの意図を解釈できることもある。

情感表示(affect display): 感情の意識的、無意識的の表出。表情による怒り、喜びなどの情緒の表れが最も顕著。

¹自然言語の意味が常に明確であるという意味ではない。

²その他の知識(常識)を用いることによって得られる情報も存在するが、ここでは考えない

調整子 (regulator): 会話の流れの調整. 主に視線の動きによる発話の促進や話者の交替. 大抵の場合は送り手が意図的に行う動作ではなく, 無意識の習慣になっている.

適応子 (adaptor): 身体的要求を満たしたり, 情緒を管理したりといった, 状況に適応するための動作. 発話内容と無関係に表れる.

2.3 身振り動作の持つ情報

Ekman の分類を参考に, 身振り動作から取り出すことのできる特徴量をまとめてみると, 以下のようになる.

記号: 標識に見られるような, 自然言語と一対一に置き換えられる記号, 概念.

指示, 照応: 対象の指示, 実在することの表示, 仮想的な存在 (の導入)

物理量: 量 (大きさ, 長さ, 速さ, 重さ, 移動量), 形態 (静止形, 運動軌跡), 位置 (位置, 方向) 等

心理的態度: 何らかの部分の強調, 感情の表現, リズム等
これらのうち, 記号については身振りと記号の対応辞書を作成し, それと対応づけることによって, 身振りの表現する内容を取り出すことができる. また, 物理量に関しては, 手, 足, 頭等の主要な部分の位置や動きを計測することで抽出することができる. ただし, 指示や照応, 心理的な態度については状況や発話との関連を考えなければ意味を持たない (これについては, 3 章で議論する).

以上のように, 身振り動作の持つ物理量またはパターンを抽出することによって多くの有用な情報を取り出すことができるが, 実際のコミュニケーションの場面を考えた場合には, 身振り動作の曖昧性, 状況依存性を無視することができない.

2.4 身振りの状況依存性と文脈

個々の身振りの役割やそれが本質的に持つ情報は, 上で述べた数種類に大別されるが, ある特定の身振りから実際に話し手の意図を推定し, そこから必要な情報を取り出す問題は簡単ではない. 例えば, 図 2 に示す例を考えてみよう. この身振りを文脈情報なしに解釈する場合には, その候補として様々なものが考えられる. 例えば, 相手に何かの大きさを伝えるために両手の間隔を提示している場合 (例示子の空間動作), 何かを保持している状態を表している場合 (例示子の活動動作), 又は相手に何かを要求している場合 (標識) も考えられる. 同様に, 片手を上げるような動作でも, 上の方向をさす指示動作にもなりうるし, 高いところの物をとる動作ともなりうる. また, 場合によっては, タクシーを拾う表現の一部分かもしれない.

このような対象を認識し, 送り手の意図を理解するためには, 常に文脈を考慮した処理を行うことが必要となる. そのために考えるべき文脈として, 最も直接的に関係するものは, 身振りに同期した発話と実空間での環境 (物体や



図 2: 身振動作の曖昧性と状況依存性

その他の 3 次元世界) であろう. 本研究ではこの点に着目し, 身振り動作に伴う発話によって, 動作の持つ曖昧性, 状況依存性を解消し, そこから有用な情報を取り出すことを提案する.

3 複数モダリティ間の関係とその利用

前章で述べた文脈を利用するためには, まず, 他のモダリティとの相互関係を明らかにし, それを有効に用いることが必要となる. ここでの基本的な考え方は, 複数のモダリティの対応部分を見つけ, それを用いてお互いの表す情報の曖昧性を解消すること, さらに, それをもとに個々のモダリティが固有に持っている情報, 暗黙に対応している部分を見つけることである.

3.1 複数モダリティの相互関係

まず, 身振り, 発話 (言語) が伝える情報の間の相互関係を大まかに分類しておく.

共通部分: 複数のモダリティで等価な情報を伝えている部分. 例えば, “これ” という発話が指示を表すということと, それと同期した指をさす身振りが指示動作であるということは, 等価であると考えて良い.

固有部分: 複数のモダリティで相補的に一つのことを説明する場合に, 各モダリティ独自の情報となる部分. 上記の例では, 指がさす方向は動作だけが持つ固有な情報となる.

他モダリティの説明部分: 他のモダリティの構成, 意味を別のモダリティを使って説明している部分. 主に, 言語情報を用いて他のモダリティが表す意味を述べる場合が多い. 例えば, 手の形の持つ意味を発話で説明している場合等.

他モダリティの強調部分: 他のモダリティを強調する部分. 相手の注意を特定の要素に向ける. 例えば, 手を大きく振って発話を強調する等.

ここで, 共通部分, 他モダリティの説明部分が, モダリティ間の対応関係を見つけるための大きな手がかりとなる.

3.2 動作と発話の相互関係の表出

本研究では, 動作と発話において上記で示したような共通部分, 説明部分がどのように表出されるかについて, 実際

表 1: 品詞と動作の関係

品詞	対応動作種別
指示詞 (名詞形態)	標識, 指示動作
指示詞 (名詞修飾形態)	(係受け先の名詞による)
指示詞 (述語修飾形態)	(係受け先の述語による)
名詞 (空間的属性, 量を表す)	量:空間動作, 形:象形動作
名詞 (空間語) ³	指示動作
形容詞 (空間的属性, 量を表す)	量:空間動作, 形:象形動作
名詞 (具体物を表す)	標識, 指示動作, 象形動作
動詞/さ変名詞 (空間的動作を表す)	標識, 活動動作

のプレゼンテーションに対して調査を行った。その結果、以下のような形の相互関係が多いことがわかった。

時間的な同期による対応: 最もよく現れる関係であり、同時刻に表出する複数のモダリティが示す対象が一致する。ただし、完全な同時性は必ずしも必要でなく、どちらかが先行することが多い。

意味的な対応: 発話中の単語、文章と意味的に一致する動作、形状の表出を行う。特に顕著なのは指示を伴う対応であるが、指示以外にも多く見られる。

例 1: “このリンゴ”(指さして),

例 2: “このくらいの大きさ”(両手で一定の間隔を示す),

暗黙の対応: 仮想物体の提示などのように、直接的、意味的な対応関係がない場合。時間的な対応、前後する指示や意味

的な対応と照らし合わせることで対応関係がわかる。

例 3: “カメラを回転させ”(カメラに見立てた右手を回す)

直接指示による説明: あるモダリティを用いて直接他のモダリティの要素を指示、説明する。ほとんどの場合、指示する側は自然言語。

例 4: “この形が”(手で形を作りながら),

上記の関係のうち、時間的対応関係は発話、身振り動作が起こった時刻を計測することで簡単に見つけることができる。また、意味的な対応、直接指示による説明については、発話中の単語や文節と動作種別の間に緩い関係があり、これを利用することができる。これを表 1 に示す。指示詞は対応関係を発見するための非常に良い手がかりになるが、名詞修飾形態、述語修飾形態の場合には係受け先の単語によって分類する必要がある。

また、これらの関係はお互いに排他的ではなく、実際の身振りの表出場面では、複数の関係が成立することが多い。

3.3 対応関係を用いた曖昧性の解消

上記の関係を用いることによって、身振り動作の曖昧性を減らすことが可能である。その基本的な考え方は、(1) 身振りの表し得る種別、(2) 発話の内容から予測される身振り動作、(3) 上記の対応関係の存在、の 3 つの条件をチェックすることによって、(1) や (2) で起こる大きな曖昧性を減らすことである。その簡単な方法を以下の例で示す。

例 5: “このリンゴ”+ 指さす (指示動作) + リンゴ (実体)
指示動作の抽出は比較的簡単であるが、他の動作の可能

性が残る場合もある。しかし、時間的な対応によって関係付けられた“この”という指示詞(名詞修飾形態)と空間的に実体を持ち得る名詞“リンゴ”との対応によって、指示動作であることが特定できる⁴。

例 6: “こんな形のリンゴ”+ 長方形を空間に描き出す (象形動作) + 実体なし

“こんな”(名詞修飾形態)は、係先の単語によって、指示動作、象形動作、空間動作のいずれとも対応し得るが、ここでは“形”(象形動作の属性を表す名詞)に係っていることから、象形動作である可能性が大きい。ただし、この場合にも空間的実体がある可能性を完全に否定することはできない。

ただし、複雑な例になると、このような大まかな対応関係では判断できない場合が出てくるため、さらに理論的な考察が必要である。

3.4 身振りと発話の固有情報の抽出と利用

前節までで、身振りと発話の対応づけができたと仮定すると、対応関係をもとに、個々のモダリティが独立に持っている情報を比較、統合することができる。本研究では統合の一般的な方法について議論しないが、2.3 節で示した特徴量を取り出して、発話と対応付けることを行う。まず、前節のような対応関係がわかったとしよう。これによって得られるのは、主に身振りの動作種別である。これをもとに、2.3 節で述べたような身振りの持つ特徴量をそのまま新しい情報として抽出することができる。例えば、上記の例に対して:

- 例 5 では、“この”という指示詞と指示動作が共通の意味を持った部分となる。これ以外の部分、つまり、“リンゴ”、指示された空間的な実体(リンゴ)等が固有な情報となる。これらに関係付けることによって、以後、空間的実体(リンゴ)の位置(方向)、属性、その他の説明事項などが参照可能になる。

- 例 6 では、“こんな+形”という“指示詞+属性を表す名詞”と象形動作(であるということ)が共通な情報となり、“リンゴ”、手で示す大きさ、形等が固有な情報となっている。これらに関係付けることにより、リンゴの形が四角で大きさが 10cm ぐらいであるといったことがわかる。ただし、“形”の代わりに“四角い”という単語が用いられた場合には、身振りから抽出された形状と“四角い”という表現の同一性を確認することが必要となる。

このように、対応関係をもとに身振りの特徴量を抽出することによって、多くのことが新しい情報として得られ、また、関係付けられる。ただし、後者の例で述べたように、

⁴環境中にリンゴが実在すれば判断は簡単であるが、一般的には話者が仮想的に物体や環境を提示する場合との区別が必要である。これを判断する最も確実な方法は画像認識等によって確かめることであるが、それが可能でない場合には、身振り動作の種類や発話内容などから大まかに推定することが必要となる。この点については、まだ十分にわかっておらず、さらに調査が必要であろう

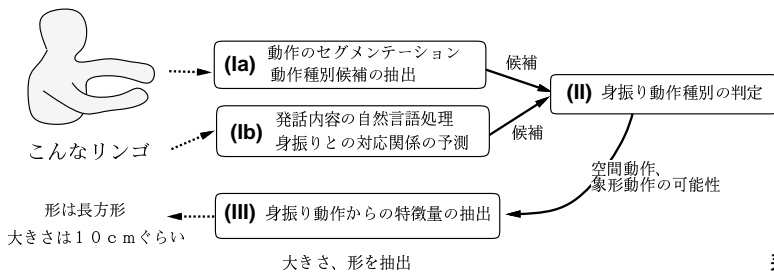


図 3: 動作理解システムの枠組

単に新しい特徴量を抽出するだけでなく、既に与えられている情報と比較することが必要である。

4 プレゼンテーション動作の理解システム

問題設定: 本研究では、以上の考えをもとに、プレゼンテーション映像における話者の動作理解を自動的に行う枠組を提案する。ただし、現在の段階では最も基本的な2つのことに目標を絞っている。

1. 話者の動作種別の認識 (動作が 2.2 節のどの種類であるかを判別する⁵⁾)
2. 動作からの新しい情報の抽出 (動作から 2.3 節で述べた情報を抽出する)

その大まかな枠組を図 3 に示す。まず、(Ia) 動作のセグメンテーション (動作の時系列からの切り出し) を行い、各種類の動作に含まれる候補を探す。2.4 節で述べたように、ここでは大きな曖昧性が残る。同時に、(Ib) 発話内容に対して、形態素解析、係受け解析を行い、各動作と対応する可能性のある単語、言い回し (文節等) を探す。(II) で (Ia) の結果と (Ib) の結果を比べて、最も妥当な身振りの種別を選択する。(III) で (II) で得られた結果をもとに、身振りから重要な物理量、指示等の情報を取り出す。ここでは対象としていないが、得られた結果を発話の解析にフィードバックすること、身振り、発話、環境を統合した意味解析等を行うことなどが、この研究の延長線上に考えられる。

データの取得: 動作情報の取得にはプレゼンテーションを行う発話者に複数の磁気センサを装着し、装着部位での位置・方位角の情報を用いた。ただし、TV カメラによる撮影を補助的に併用している。言語情報に関しては映像に付随する音声言語の自動認識が现阶段では難しいため、発話内容を自然言語の文章として手動で与える。ただし、各文節が映像フレームのどこで発話されたのかを表す情報を同時に与え、言語と動作の時間的な同時性・連続性を利用できるようにする。

(Ia) 動作のセグメンテーションと候補抽出: 得られた時系列の動作情報から、2.2 節の動作種別に対応する区間を抽出する。そのために、身振り特徴量を導入し、身振り特徴量の時間的変化が各動作種別の典型的な動作と一致する場合にその動作種別の可能性があるとする。表 2 に本

⁵ただし、話者が伝達を意図しない無意識 (あるいは implicit) な表情や行動は、本研究の対象からは省いている。その結果、上記の分類のうち、標識と例示子を主に扱うことになる。

表 2: 身振り特徴量

手の位置	(a) 右手/左手の高さ, (b) 右手/左手と胴体の距離, (c) 視線方向への右腕/左腕の伸び, (d) 右手/左手と注視軸線の距離
手の形態	(e) 右手と左手の間隔, (f) 右手/左手の方向
手の運動	(g) 右手/左手の運動速度

表 3: 取り出す動作種別と関係する身振り特徴量

身振り動作の種別	関係する特徴量
標識 (あらかじめ登録してある動作のみ)	a, b, c, d, e, f, g
指示	c, d, f
空間動作, 象形動作	e, f
活動動作	f, g

研究で使用している身振り特徴量をあげる。これらの特徴量は、頭部、両腕、及び胴体の位置と向きから計算できる。また、現在用いている身振り特徴量と動作種別との関係を表 3 に示す。

このような方法をとると、同一の区間に対して複数の候補が存在したり、複数の動作区間 (セグメント) が重なりあってお互いに矛盾した結果が生じるが、これらは上記 (II) の処理で絞りこまれることになる。

(Ib) 発話内容の自然言語解析: 発話文を形態素解析、構文解析することで [Nag92, 黒橋 92], 発話文中の形態素と係受け構造を得る。次に、表 1 にあげた関係から、対応する動作種別の候補を得る。ここでも、複数の動作種別と対応し得る単語が存在するため、一意に絞り込むことは難しい。

(II), (III) 候補の絞り込みと身振りからの特徴量抽出: (Ia) と (Ib) の間に 3.2 節で示したような対応関係がある場合に、それを最終的な候補とすることのみを行っている。そのため、複数の候補が最終的に残ることもある。これについては、より細かい動作の解析、時間的文脈を考えて絞り込む必要があるだろう。

最後に、得られた動作種別に応じて、物理量、指示対象などのパラメータを抽出する。例えば、“このぐらいの大きさ” という場合には、大きさを表す空間動作であるので、そこから両手の間隔を抽出する。ただし、どの物理量を抽出すべきかという問題に対しても若干の曖昧性があり、これを解消するためには、今後の調査、実験が必要である。

実験例: 現在はまだシステムの構築段階であり、標識動作についての動作認識は行っていない。処理可能な範囲で実験を行った例を図 4 に示す。(a) にプレゼンテーション映像を、(b) に発話内容の処理結果を、(c) に動作のセグメンテーション結果を、(d) に (II) の処理によって得られる結果を示す。(e) が身振りから得られた特徴量である。また、その他の典型的な例について実験を行った結果を、図 5 に示す。

以上の例からわかるように、プレゼンテーションから十分な精度で情報が抽出できれば、提案する枠組によってかなりの情報抽出が可能である。ただし、現在の段階では、(II) 以降の部分でかなりアドホックな処理を行っており、

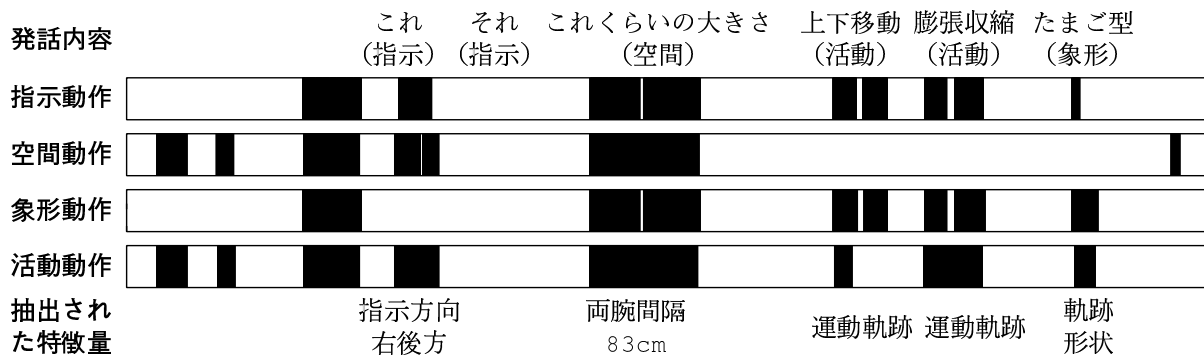
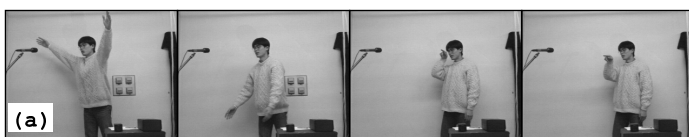
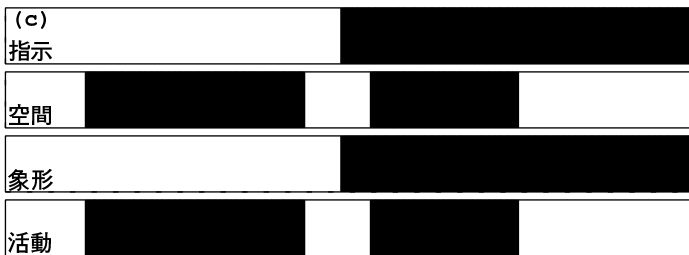
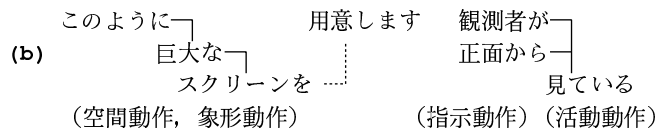


図 5: 身振り解析結果 (2)



まず、このように巨大なスクリーンを用意します。観測者が正面から見ている…



(d) 空間動作 指示動作
 (e) 両手間隔の最大 110cm 右手指示方向 話者正面方向

図 4: 身振り解析結果 (1)

一般的なプレゼンテーションに適用するためには、それらの部分の理論的再構築が必要である。

5 おわりに

本稿では、人間の動作理解の問題をコミュニケーションという視点からとらえ、身振りの持つ基本的な情報、発話との相互関係を考察した。また、これを用いて、人間の身振り動作の役割を認識し、そこから有用な情報を抽出するための簡単な枠組を提案した。ただし、ここで提案しているシステムは、まだ開発途上であり、これからの種々の改良、一般的な実験が必要である。

また、将来的に、身振りを画像から、発話を音声から抽出することを考えると、入力是非常に誤りの多い不完全なものになる。そのため、複数のモダリティ(あるいはメディア)から情報を抽出するための文脈情報としてお互いの情報を援用することの意義はより大きくなるだろう。そのため、上記のような一方向的な処理ではなく、音声、画像各々を用いて、お互いの処理の目的、視点、焦点(注目要素)を動的に決定していくメカニズムを明らかにしてい

なければならない。

最後に、画像、音声、言語といったものを有機的に統合してどんなことが可能になるかという点については、まだ未知数の部分が多いが、このような考え方は種々の分野で必要とされるだろう。その一例としては次のようなものがあげられる。

- マンマシンコミュニケーションのためのフィルタ。例えば、指示を表す表現を他の身振りから分離する等。
- 将来の、より柔軟なマルチモーダル・マンマシンコミュニケーション
- ロボットや計算機視覚における人物行動理解のための要素技術
- 通信メディアにおいて人間のコミュニケーションを支援するための視覚。例えば、相手に見せたいと意図する部分を自動的に撮影する技術等。

参考文献

- [Coh94] P. Cohen. "Natural Language Techniques for Multimodal Interaction". 信学論, Vol. J77-D-II, No. 8, pp. 1403-1416, 1994.
- [EF69] P. Ekman and W. Friesen. "The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding". *Semiotica*, Vol. 1, pp. 49-98, 1969.
- [Kur92] T. Kurokawa. "Gesture Coding and a Gesture Dictionary for a Nonverbal Interface". *IEICE Trans. Fundamentals*, Vol. E75-A, pp. 112-121, 1992.
- [McN87] D. McNeil. "*Psycholinguistics*". Harper & Row, 1987.
- [Nag92] Nagao Laboratory, Dept. of Electrical Eng. Kyoto University. "*JUMAN Manual (In Japanese)*", 1992.
- [益岡 92] 益岡隆志, 田窪行則. 基礎日本語文法. くろしお出版, 1992.
- [黒橋 92] 黒橋禎夫, 長尾真. "並列構造の検出に基づく長い日本語文の構文解析". 情処研報 NL, Vol. 88-1, , 1992.
- [黒川 94] 黒川隆夫. ノンバーバルインタフェース. オーム社, 1994.
- [中村 95] 中村裕一, 古川亮. "概念図の理解を目的としたパターン情報と自然言語情報の統合". 情報処理学会論文誌, Vol. 36, No. 1, pp. 196-205, 1995.
- [末長 92] 末長康仁ほか. "Human Reader: 人物像と音声による知的インタフェース". 信学論, Vol. J75-DII, No. 2, pp. 190-201, 1992.