ヒューマンインタフェース 75－9
（1997. 11. 14）

音声とポインティングジェスチャを利用した指示物同定

山田寛康, 福本文代, 今宮淳美
山梨大学工学部電子情報工学科

概要

本稿では, 音声とマウスを入力手段とするマルチモーダルインタフェースで, ディスプレイ上の図形を指示したとき, その指示物を同定する手法を提案する. 本手法は時間同期性を用いた指示物候補の生成（generation of candidates) と, 過去の対話情報を利用した候補中からの指示物同定（identification) の 2 つの手続きにより実現される. Generation of candidates では, すべての発話を segment と呼ぶ単位に分割し, segment と同期した時間内にマウスによって指示した図形を指示物候補として生成する. Identification では, 過去に指示物同定で得られた結果を利用することで候補中から指示物を同定する.

Referent Identification by Using a Speech and Deictic Gesture

Hiroyasu Yamada, Fumiyo Fukumoto, Atsumi Imamiya
Dept. of Electrical Engineering and Computer Science
Yamanashi University

Abstract

In this paper, we propose a method to deal with the reference of deictic expression to visual objects on a terminal screen. Users can point to visual objects on a terminal screen by using both an indirect pointing device, a mouse and natural language. The procedure for determining the referent of the deictic expression, i.e. determining the visual objects being pointed to, consists of two procedures: Generation-of-Candidates and Identification. In Generation-of-Candidates every utterance is divided into a segment using linguistic constraints called verb phrase rank. Then, candidates of visual objects are generated from trace by a mouse, together with a segment. This segment includes Japanese deictic expressions such as 'this (これ)' or 'that (あれ)'. In the second procedure, Identification, a deictic expression is identified with visual objects using pairs of deictic expressions and the visual object candidates, which have already been obtained. The results of the preliminarily experiment show the effectiveness of the method.

1 はじめに

人間同士に見られる，音声，ジェスチャなどを用いた対話を人とコンピュータとの間で実現するためマルチモーダルインタフェースの研究が行なわれている [Winograd, 1973], [Grosz, 1986], [Neal, 1988], [Moore, 1990], [Wahlster, 1991], [Takebayashi, 1992], [Mizunashi, 1997], [Oviatt, 1997]. 人間同士の対話では指示語を頻繁に使用する．指示語のなかでも実際に存在する物や場所を指す指示を現場指示という．人の指示の多くは，指示語を発話すると同時に指示物を指す．このような場合，人は発話された指示語と指示動作の異なる情報を相補的に利用することで指示物同定を実現している．従ってマルチモーダルインタフェースの研究において指示物同定は重要な問題の一つとなっている．

菊池らは，ペンによる指示動作と音声に対し，時間同期性を用いることでこれら2つの情報を統合する手法を提案した [Kikuchi, 1995]．さらにこの統合手法をファイルの移動，削除，複写の3つのタスクが実行可能なマルチモーダルウィンドウシステムに適用することでその有効性を示した．彼らの手法は発話された音声入力の単位を文節とし，これと同期する時間内にペンで指示したオブジェクトを指示物としている．例えば「このファイルを複写する」といった場合，「この」という指示語に対して，「この」が発話された時間内にペンで指したオブジェクトを指示物と同定する．しかし人の指示は曖昧であり，指示語を発話すると同時に指示物を指すとは限らない．従って彼らの提案した文節を単位とする時間同期性による統合手法では，指示語の発話と指示動作が非同期の場合，正しく指示物を同定できない．

Koons らは，音声，ハンドジェスチャ，凝視（目の動き）を使用して，スクリーン上の図形を操作するシステムを作成した [Koons, 1993]．また Kobsa らは，自然言語とタッチパネルによるポインティングジェスチャを使用して学会の会費を自動的に支払うシステムを構築した [Kobsa, 1986]．彼らの手法は，音声あるいは自然言語文に対し深い構文意味解析を行なった結果，人手で作成された概念知識ベースとの照合により指示物候補の中から指示物を一意に決定している．しかし様々な分野に対応できる言語情報を概念知識ベースとしてあらかじめ網羅的に記述しておくことは難しいため，結果的に汎用性に欠けることが問題となっている．

本稿では，音声とマウスを用いてディスプレイ上の図形を指示したとき，その指示物を同定する手法を提案する．本手法は，2つの手続き，すなわち時間同期性を用いた指示物候補の生成（generation of candidates）と，過去の対話情報を利用した候補中からの指示物同定（identification）の手続きからなる．指示語と指示動作の統合手法は菊池らと同様に時間同期性を用いる．しかし菊池らが発話された音声入力の単位を文節として同期をとるのに対し，本手法では，segment とよばれる単位を用いる．これは動作の継続を示す接続助詞を用いて決定される．Generation of candidates では segment と同期している時間内にマウスによって指示した図形を指示物の候補とする．これにより指示語の発話と指示物を指す動作とが非同期である場合にも指示物の候補を生成することが可能である．さらに Koons らや Kobsa らが，人手で作成した知識ベースとの照合により指示物候補の絞り込みを行なっているのに対して，我々は identification において過去の対話情報を利用することで候補の中から指示物を同定する．

以下2章では，時間同期性の単位となる segment について述べる．3章では指示物同定のための2つの手続きである generation of candidates と identification について述べ，4章で本手法の有効性を検証するために行なった実験について報告する．5章でまとめと今後の課題について述べる．

2 時間同期性の単位

指示語とマウスの指示動作は時間同期性を用いて統合することができる．本手法では，以下に示す4つの制約により発話された文を segment と呼ぶ単位に分割する．

1. 発話者の変更による分割

   発話文は発話者が変わった時を境にして分割する．

2. 一文単位での分割

   発話文を一文単位で分割する．ここで文とは少なくとも一つ以上の動詞か，一つの名詞句を含んでいる単語列とする．

3. 感動詞による分割

   発話文中に感動詞が存在する場合，その感動詞の前後で文を分割する．

4. 接続助詞の種類による分割

   表 1 に動詞句の分類を示す．主節と従属節の結び付きの強さは，従属節の動詞に付与される接続助詞の種類により決定される [Minami, 1986]．例えば，表 1 の **phrase-A** に属する「比較しながら」は，主節との結び付きが強い．一方 **phrase-B** に属する「比較すると」及び **phrase-C** に属する「注目

表 1: 動詞句の分類

phrase-A	継起条件句 (接続助詞 'たり','ながら', 'まま', 'つつ', etc.)
phrase-B	仮定的条件句 (接続助詞 'なら', 'ば', 'ても' 'と', etc.)
phrase-C	逆接条件句 (接続助詞 'ので', 'ため', 'から') 'けれど', etc.)

したけれども」は, 主節との結び付きが弱い. 我々は動詞句が **phrase-B** 及び **phrase-C** に属する場合, 発話文を主節と従属節との境で分割した.

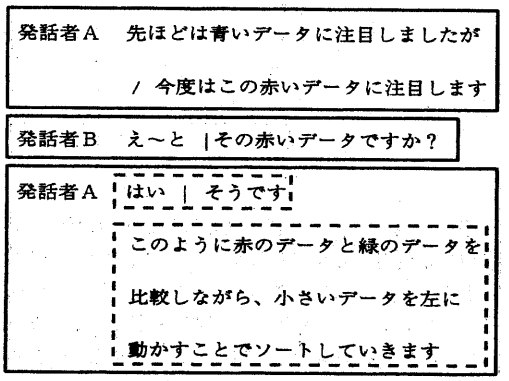


図 1: 4つの制約による発話文の分割

図 1 に上記 4 つの制約により分割された対話例を示す. 図1において, □, ⬚, |, /はそれぞれ制約 1, 2, 3, 4 により分割された結果を示す. 分割された各々を segment と呼ぶ.

3 指示物同定

3.1 候補生成 (generation of candidates)

指示物の候補は segment に指示語が含まれている場合, segment の時間と同期しているマウスの位置情報を利用することで生成される. 図 2 に segment を単位とした指示語とマウスの指示動作との時間同期の例を示す. これにより, 指示語の発話と指示動作が非同期な場合にも指示物の候補を生成することができる.

図 3 に指示物候補生成の例を示す.

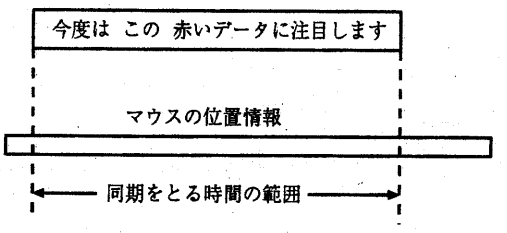
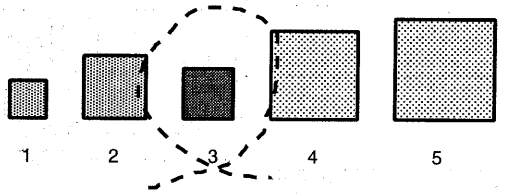


図 2: segment を単位とする時間同期性



"今度はこの赤いデータに注目します"

図 3: 指示物候補の生成

図 3 において四角形はオブジェクトであり, 点線は segment と同期したマウスの軌跡を表す. マウスカーソルがオブジェクトを通過するかその軌跡がオブジェクトを囲んだ場合, それらのオブジェクトを指示物候補とする. 図 3 ではポインティングデバイスが通過したのはオブジェクト 2, 4 であり, 軌跡によって囲まれているのはオブジェクト 3 となるため, 候補として生成されるのは, オブジェクト 2, 3, 4 となる.

3.2 指示物同定 (identification)

本手法では, 生成した候補の中から正しい指示物を決定するために, 過去の指示物同定情報を利用する [Grosz, 1986], [Kobsa, 1986]. 過去の指示物同定情報とは, それまで指示物同定で得られた指示物の情報であり, 指示語とその指示物であるオブジェクトの組で表す. 指示物同定の過程の例を図 4 に示す. 図 4 において (a)〜(c) は対話例を示す. (a')〜(c') は対話で得られた指示物同定情報を表す.

図 4 (a) では, 「この赤いデータに注目します」と発話しマウスを点線のように動かしたことを示している. この時の指示物候補はオブジェクト 2 〜 4 である. この発話が対話の始まりなので, 過去の指示物同定情報は空である. よって「この赤いデータ」に対する指示物はオ

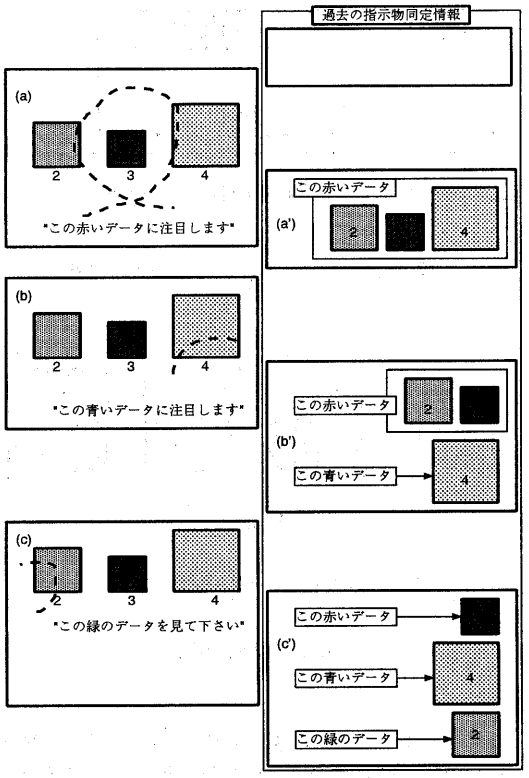


図 4: 指示物同定

ブジェクト 2～4 であり, 図 4 (a') に示すような, 指示語と指示物候補の組が過去の指示物同定情報として記録される.

次に図 4(b) に示すような発話とマウスによる指示を行なったとき,「この青いデータ」という指示語に対して候補はオブジェクト 4 となる. この指示語の名詞句の構造と, 過去の指示物同定情報の中にある「この赤いデータ」という構造は異なっている. よって「この赤いデータ」という指示語に対する指示物候補オブジェクト 4 を候補中から取り除く. その結果図 4 (b') に示すように過去の指示物同定情報が更新される.

図 4(c) で,「この緑のデータ」という指示語に対して候補はオブジェクト 2 の 1 つとなる. 現在の指示語の名詞句の構造と過去の指示物同定情報である「この赤いデータ」の構造は異なる. よってオブジェクト 2 を「この赤いデータ」という指示語の候補から取り除く.

以上により,「この赤いデータ」という指示表現に対してオブジェクト 3 を指示物として同定することができる (図 4 (c')).

4 実験

4.1 データ

本手法による指示物同定の有効性を検証するため実験を行なった. 実験では, 2 人のユーザのうち, 一人が説明者となって, もう一人にソーティングアルゴリズムを説明する. 説明ではディスプレイ上に表示されたアルゴリズムの過程を示した図を使用し, 指示にはマウスを使用する.

表 2: 指示語の種類と出現個数

指示語	個数
これ	47
それ	9
この	50
その	1
合計	107

表 2 は実験で扱った指示語の種類と出現個数を示す. 説明者が使用した指示表現のうち,「このような」「このように」「ここで」「こっち」「ここに」という指示語はディスプレイ上に表示された図形を指示していないため指示物同定の対象から削除した. その結果, 表 2 で示された 4 種類の指示語, 合計１０７個について指示物同定手法を適用した.

4.2 結果

実験結果を表 3 に示す. 表 3 において '完全正解' とは, ユーザが指した図形と本手法により出力した結果が完全に一致した (指示物同定) 場合を示す. '部分正解は', 本手法の出力結果にユーザの指示した図形だけでなく余分な候補が含まれていた場合を示す. '不正解' は出力された中にユーザが指示した図形が含まれてない場合を示す.

表 3: 実験結果

		候補生成 (%)	⟶	指示物同定 (%)
正解	完全	38(35.7)	⟶	65(60.9)
	部分	46(42.9)	⟶	20(18.6)
不正解		23(21.4)	⟶	22(20.5)

4.3 考察

表3より, 候補生成 (generation of candidates) では完全正解, 部分正解がそれぞれ38, 46個であったが, 指示物同定 (identification) を適用することで最終的に65個の完全正解を得ることができた. しかし正しく指示物を同定することができない場合が42個あった. 原因として以下の3点が考えられる.

1. 位置に関する情報

   本手法では, 語の表層的な情報を用いて指示物の同定を行なっている. 従って空間的な位置関係を示す情報が必要な同定には対処できていない. 例えば, 図5で示されるように, 「この2番目の図になります.」と発話して図の点線のようにマウスを動かしたとき, 人は2番目という位置の情報から5つの図形を指示物として同定している. しかし本手法では位置に関する情報を使用していないため, マウスの軌跡からは最後の5番目の図を候補として生成することができない.

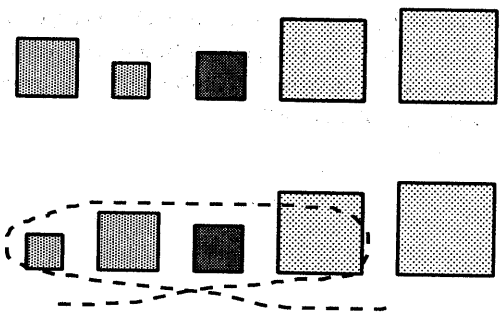


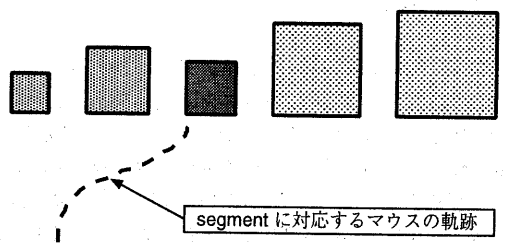
図 5: 空間的な位置関係を表わす指示

2. 時間同期性の単位

   本手法では, segment に同期しているマウスの位置情報を使用して指示物の候補を生成している. しかし図6に示すように segment に対応するマウスの軌跡が指示物となるオブジェクトを通過してない場合が生じる. 従って今後より詳細な分割方法を導入する必要がある.

3. segment 間の関係

   本手法では, 発話された指示語の名詞句の構造と, 過去の指示物同定情報の指示語の名詞句構造を比較することで指示物同定を行なっている. しかし図7に示すように, 発話された指示語が「これ」「それ」だけであると, 指示語の名詞句の構造を比較する情報がないため, 候補の中から指示物を決定できない. 今後このような問題に対処するために, segment 間の関係を利用する必要がある. 例えば 図7において, 次の segment の中の「すなわち」という言い換えの接続詞の情報を利用すれば, 指示語「これ」に対して「赤い図形」,「緑の図形」といった名詞句の比較に必要な情報を得ることができる.



"緑の図形と比較すると、

この赤い図形は一番右にいきます"

segment

図 6: segmetn に対応するマウスの軌跡がオブジェクトを通過していない場合

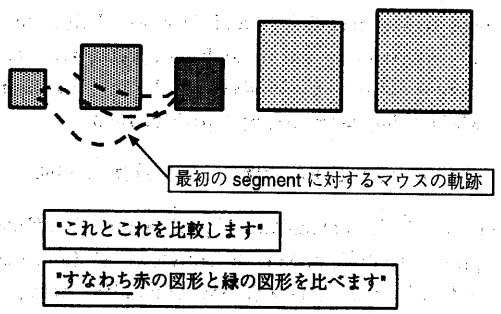


図 7: 名詞句の情報が不足している場合

5 まとめ

本稿では, 音声とマウスを用いてディスプレイ上の図形を指示したとき, その指示物を同定する手法を提案した. 本手法では, 発話を segment と呼ぶ単位に分割し, segment の時間と同期したマウスの位置情報を用いることで候補を生成した. また過去の指示物同定情報を利用することで生成した候補の中から指示物を一意に決定した. 今後は考察で述べた問題点に対処すると同時に, より複雑な図形や対話を扱った例題に適用することで, 本手法の有効性を検証する予定である.

謝辞

本研究の一部は通信・放送機構の援助を受けています. ここにそれを記し謝意に代えさせていただきます.

References

Grosz B.J., and Sidner C.L., "ATTENTION, INTENTIONS, AND THE STRUCTURE OF DISCOURSE", in: *Computational Linguistics*, Vol.12, No.3, 1986

菊池 英明, 安藤 ハル, 畑岡 信夫, "音声とペンを入力手段とするマルチモーダルインタフェースの構築" *SIGSLP, Vol.95, No.7*, pp. 113-117(1995)

Kobsa A. et al. "Combining Deictic Gestures and Natural Language for Referent Identification", in: *Proc. of International Conference on Computational Linguistics*, pp. 356-361, 1986

Koons D.B., Sparrell C.J., and Thorisson K.R., "Integrating simultaneous input from speech, gaze, and hand gestures", in: *Maybury, M.T.(Ed), Intelligent Multimedia Interfaces*, pp. 267-276, 1993

南 不二男, "現代日本語の構造", 大修館書店, 1986

Mizunashi G., Loken-Kim K., Morimoto T., "Integrated Analysis of Speech and Gesture", in: *Interaction '97 (in Japanese)*, Information Processing society of Japan, pp. 41-42, 1997

Moore J. D., and Swartout W.R., "Pointing: A Way Toward Explanation Dialogue", in: *Proc. of the 8th National Conference on Artificial Intelligence (AAAI-90)*, pp. 457-464, 1990

Neal J.G., Dobes Z., Bettinger K.E., and Byoun J.S., "Multi-Modal References in Human Computer Dialogue", in: *Proc. of the AAAI-94 workshop on Integration of Natural Language and vision Processing*, pp. 7-13, 1988

Oviatt S., and DeAngeli A., and Kuhn K., "Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction", *Proc. of CHI97*, pp. 415-422, 1997.

Takebayashi Y., "Integration of Understanding and Synthesis Functions for multimedia Interfaces", in: *Multimedia Interface Design*, Blattner (Ed), ACM Press Book, 1992

Wahlster W., "User and Discourse Models for Multimodal Communication", in: *Sullivan and Tyler*, pp. 45-67, 1991

Winograd T., "A Procedural Model of Language Understanding", in: *Computer Models of thought and Language*, W.H.Freeman and Company: San Francisco, pp. 152-186, 1973