TR-IT-0060

マルチモーダル・ヒューマン・コンピュータ・インタラクション関連文献調査

Survey of Multimodal Human-Computer Interaction

水梨 豪

Suguru MIZUNASHI

1994.7

概要

マルチモーダル・ヒューマン・コンピュータ・インタラクション(以下MMHCI)とは、人間と計算機の間のさまざまな情報伝達方法（モダリティ）を有機的に統合しながら行なわれる、人間と計算機の対話方式を指す。本稿では、近年さかんになってきたMMHCI関連の研究を、入力された異種の情報がどのようにして統合され解釈されるか、あるいは出力されるべき情報はどのように決定され各モダリティに分配されるかされるかという視点で概観し、整理した。さらに、音声入力と手によるジェスチャを用いたマルチモーダルシステムに関して考察を行った。

ATR 音声翻訳通信研究所

ATR Interpreting Telecommunications Research Laboratories

目次

マルチモーダル・ヒューマン・コンピュータ・インタラクション関連文献調査

1994.7.6　　水梨 豪

1.マルチモーダル・ヒューマン・コンピュータ・インタラクション(MMHCI)

CPUの処理効率、CRTの解像度の向上などのハードウエアの進歩にともない、人間と計算機の対話様式における多様性が増大している。人間が計算機に情報を入力する仕方においては、今までのキーボードやマウスに加えて、ペンによる手書き文字の入力、タッチスクリーンを使った指による指示入力、マイクロフォンによる音声入力、ビデオカメラによる映像（ユーザの動作、視線、表情など）入力、データグローブによる動作の入力など、さまざまな入力方法が利用可能になってきている。逆に、計算機から人間への情報伝達の方法も、ディスプレイ上の文字やグラフィクスによって視覚に訴えるものにはじまり、スピーカからの合成音声などを聞かせるもの、さらには触覚や力覚をデータグローブなどによってフィードバックするものなど、人間の知覚方法を適切に利用するかたちで多岐にわたるようになってきた。

マルチモーダル・ヒューマン・コンピュータ・インタラクション(以下MMHCI)とは、上記のような人間と計算機の間のさまざまな情報伝達方法（モダリティ）を有機的に統合しながら行なわれる、人間と計算機の対話方式を指す。人間同士のコミュニケーションが人間にとってもっとも「自然で効率的」なコミュニケーションの形態だと仮定すれば、MMHCIによって、人間と計算機の対話も人間にとって「自然で効率的」なコミュニケーションに近づくことができると考えられる。

現在行なわれているMMHCIの研究における問題点は、各モダリティの要素技術における問題点に加えて、各モダリティをいかに有

機的に統合させるかという問題をも含んでいるといえる。そこで、本稿では、近年さかんになってきたMMHCI関連の研究を、入力された異種の情報がどのようにして統合され解釈されるか、あるいは出力されるべき情報はどのように決定され各モダリティに分配されるかされるかという視点で概観し、整理することにする。

2.MMHCI関連研究の分類

マルチモーダル・ヒューマン・コンピュータ・インタラクションの関連研究は、マルチモーダルの技術を入力に適用しているものと出力に適用しているものに大別できそうである。前者は主に情報の認識と解釈、後者は情報の生成やレイアウトを問題としている。このように両者は技術的には異なる様相を示しているので、別個に整理するのが妥当と考える。したがって、本項では、各研究を、入力に関する研究、出力に関する研究に大別して整理していくことにする。

2.1総論

マルチモーダルHCIに関する研究の動向を解説したものには、[長尾2][Hanne et al.][Mariani][Stock1]などがある。

2.2マルチモーダル入力

マルチモーダル入力に関しては、音声入力と直接指示の二種類のモダリティを同時に用いるものが大半を占めている。ここでいう直接指示とは、ディスプレイ上のオブジェクトなどを指示する動作全般を指す。それには、マウス、ペンによるポインティング、タッチスクリーンを用いた指によるポインティング、さらにデータグロー

ブを用いた指示動作の入力まで含まれることとする。

[Bolt]では、マルチモーダル入力のさきがけとして、Put-That-Thereというシステムが提案されている。これは、手の動作を検出する装置を身につけたユーザが、「それをそこに移動しろ」というように音声と指さしの動作によって大スクリーン上のグラフィック・オブジェクトを操作するシステムである。入力の解析方法は不明である。

[Stock2]のALFRESCOは、フレスコ画に関する情報提供システムで、ビデオディスクとタッチスクリーンをそなえており、キーボードからの自然言語入力とタッチスクリーンによる直接指示を受け入れる。具体的には、たとえばキーボードからの"Who is this person?"のような参照表現を含む自然言語入力と、ディスプレイ上のあるオブジェクトを指さす指示動作とを統合して解釈（それぞれの入力情報が持つ曖昧さを解消する）して"this"が何を指しているかを同定した後、そのオブジェクトに関する情報をデータベースから検索して応答する。

[Koons]では、マイクによる音声、手の動作を検知する装置により検出される指示動作、アイトラッカーにより検出される目の動きという3種類の情報を入力として受け付けるシステムが紹介されている。それぞれの情報にタイムスタンプを押し、別々にパーズして共通の表現に変換し、その後に時間情報をもとに三者を統合する方法をとっている。扱うタスクは、二次元の地図上のオブジェクトの操作と、三次元の仮想空間でのブロックの操作の二つである。

[Vo]では、CMUが開発した、大語彙連続音声認識システム、ワード・スポッティング技術、唇の動きを検出するシステム、アイ・トラッカー、ジェスチャ認識技術、手書き文字認識技術がまず個別の技術として紹介されている。その後、それらの技術のうち、音声認識システムと唇の動きを検出するシステムを統合したシステム、音声認識システムとジェスチャ（ペン入力）認識システムを統合したシステムの有効性が実験によって示されている。唇の動きやジェス

チャが、音声認識の曖昧さ解消に役立てられている。

[Wahlster1][Allgayer]では、音声、テキスト、マウスによる直接指示を用いてインタラクティブにフォームを埋めていくシステムXTRAが試作されている。指示動作が直接チャートパーザに組み込まれていく形で、音声入力と統合・解釈され、参照表現の同定などを行っている。

[Neal]では、マルチメディア・インターフェースCUBRICONが紹介されている。入力は音声、キーボード、マウスによって行なう。インプット・コーディネータによって3つの入力信号は単一のストリームにまとめられた後、ATNであるマルチメディア・パーザ／インタープリタによって解釈される。アプリケーション例として、地図上のオブジェクトに関する問い合わせシステムが挙げられている。出力の項も参照のこと。

[Weimer]では、キーワード認識による音声入力とデータグローブによる手の動作の入力を用いた、共同作業をサポートする3次元CADシステムが紹介されている。手の動作は、物体の形状を表現したり、物体を掴んだり回転させたりするために用いられる。音声と動作の統合に関しては明らかでない。

[伯田1]では、言語情報と画像情報との統合による理解の基礎的な研究を行なっている。言葉で述べられて指示される画像を発見するための言語・画像情報間のリンクについての考察を、「グループ化の法則」を用いて行なっている。タスクは複合電話器の操作ガイダンスを想定している（システムは実際には作られていない）。ユーザは電話器の画像を見て指で指示しながら、「このボタンの意味は何ですか」のように指示語を使ってガイドを得るしくみになっている。

[伯田2]では、自然言語（定型文の音声入力）と指さしなどの指示動作（タッチパネルとデータグローブ）の組み合わせによって、画像中の対象物を同定する機能を持った地図案内システムが試作されている。自然言語入力と指示動作入力のそれぞれから対象物などの

候補を別々に抽出し、その後にユーザが意図する対象物を特定する。

[望月]では、3次元仮想空間での対象物の操作・配置を実現するシステムが示されている。入力は音声とデータグローブによる手の動きである。まず音声入力だけをパーズし、文を得、その後、手の動作から対象物や指示場所を推定し、文中にある指示語などの同定を行う。

[Matsu'ura][新田]では、ユーザからの入力として、自由発話音声とタッチパネルによるオブジェクトの直接指示が用いられている。タスクは地図を用いた情報案内である。音声入力に対しては、キーワード・スポッティングが応用されている。直接指示は曖昧さを含んでおらず、発話にも参照表現は現れないので、二つの入力の間の情報のやりとりはほとんどない。出力の項も参照のこと。

[吉岡]では、マウスによる項目選択・確認と、音声による住所・名前の入力を用いた電話番号案内システムが提案されている。ここでも、ある項目をマウスで選択したあとに音声によって名前などを入力するしくみになっており、二つの入力情報が相互に補完しあうわけではない。

2.3マルチモーダル出力

[Matsu'ura][新田]では、グラフィクスと応答文合成音声（ハンドセットを通して）を統合して出力する地理案内システムが紹介されている。タスクに依存した対話モデルを参照しながら、出力のプランを作成し、それにそって指示・案内・画面展開を行なう。[金沢][瀬戸][竹林][舘森]では、音声自由対話システムTOSBURGⅡの試作と評価が行なわれているが、これらも[新田]に準ずる方法を用いている。入力の項も参照のこと。

[Neal]のCUBRICONでは、アウトプット・プランナーが、出力する情報の質、談話の文脈、ユーザのタスクにとってのその情報の重要度を基準に、その情報にとって適切なモダリティを選択し、音声、

テキスト、グラフィクスを出力する。入力の項も参照のこと。

[長尾1][竹内]では、音声対話システムと表情合成システムを組み合わせた質問応答システムが試作されている。ユーザが音声で質問すると、ディスプレイ上に表示された顔が、さまざまな表情とともに応答する。内部の処理は、まず発話意図の意味表現をもとにプラン認識モジュールによってユーザの意図が認識され、それに従って協調的な応答を生成するモジュールが起動される。その後このモジュールが、知識から得られた情報を発話パターンに埋め込んで発話を生成するとともに、対話において重要な典型的な対話状況を認識し、その状況において生成すべき表情に関するメッセージを表情合成システムに伝達する。

[阿部]のVirtual Cocktail Partyでは、画像と音場が出力に用いられている。ディスプレイ上に散在しているパーティ参加者のアイコンに近付いたり遠ざかったりする（マウスで自分のアイコンをドラッグする）ことによって、ヘッドフォンから遠近感のある会話が聞こえてくる。それによって、ユーザはそこで交わされている会話を耳にしたり、会話に参加したりできる。これはマルチモーダル出力技術が臨場感技術に用いられている例と考えられる。他にも、[前田]の場の雰囲気を重視したTV会議システム、[広明]のコミュニケーションウォールなどでも臨場感を実現する手段としてマルチモーダル出力を用いている。

[McKeown]では、メディアコーディネータが意味表現から決定した、出力すべきテキストとグラフィクスを最適なレイアウトで出力する方法が述べられている。同様なテキストとグラフィックを用いたプレゼンテーション・プランニング、レイアウト管理に関しては、[Arens][Wahlster2][Wahlster3]でも考察されている。

2.4その他

[Zancanaro]では、主に、マルチモーダル環境（ここでは質問応答システム）における談話の結束性の問題（特に照応の解析）が論じられている。

[ローケン・キム]では、音声入力とマウスによるポインティングや描画、キーボードによるテキストの入力が組み合わされたマルチモーダル対話環境（マルチモーダル対話環境における言語データを収集するためのシミュレータなので、モダリティ間の情報の統合は行なわれいない）が試作され、道案内などのタスクにおける言語データが収集されている。対話相手の顔もディスプレイ上のビデオ画像としてリアルタイムに見ることができる。

3.まとめ

3.1マルチモーダル入出力におけるモダリティ

入力に関しては、複数のモダリティのうちの一つとして、音声入力を用いているものがほとんどである。その理由は、音声入力は他のモダリティに比べて保持する情報量が格段に大きく、対話様式としても最も自然で（人間対人間のような）効率がよいコミュニケーションを可能にするモダリティであるからと考えられる。

音声入力とともに用いられる第二のモダリティ（第三もある場合があるが）としては、手によるジェスチャが大半を占め、目の動きや唇の動きなども、例は少ないが考えられている。そしてそれら第二のモダリティの多くは、音声対話システムにおいて音声入力の解釈を助ける補助的な役割を担っている。

マルチモーダル出力に関しては、音声が重要な役割を担うような音声対話システムへの応用は少ない。むしろ臨場感技術やプレゼン

テーション・プランニングの分野において、音声、グラフィクス、映像、テキストなどがほぼ対等に用いられている。

以降は、マルチモーダル技術応用の主流を占める音声対話システムにおける第二のモダリティとそのシステムが扱うタスクについて述べることにする。

3.2手によるジェスチャの入力

音声対話システムにおいては、音声につぐ第二の入力モダリティとして手によるジェスチャを用いたものが最も多い。対話の場面において、手によるジェスチャは音声についで有効なコミュニケーションの手段であろう。さらに、計算機側からみれば、表情などと比較すると現状ではかなり認識しやすい入力でもある。したがって、音声とジェスチャを用いたマルチモーダル入力システムは、MMHCI研究の基本的な課題となりうると考えられる。ここではそれについて現状と課題を述べる。

ジェスチャの制限

手によるジェスチャを用いたシステムでは、主に、ジェスチャが指示するオブジェクトなどをもとに音声入力中の参照語の同定などが行なわれている。実際の入力方法としては、タッチパネルを備えたディスプレイに直接指で触れて、表示されているオブジェクトやディスプレイ上のある領域を指示するものが大半である。

しかしながら、そこで用いることができる直接指示の仕方はかなり定型的・限定的である。今後、より自然な直接指示の仕方が許される方向へ改善の余地があるといえる。

手によるジェスチャを入力する方法として、データグローブも多く使われるようになってきた。データグローブを手にはめて、指さしの方向や手の動作（ものをつかんだり回転させたりする）によって、オブジェクトや領域、あるいはその移動方向などを指示する方法である。この場合、タッチスクリーンの平面性に比べて、手の動

作の自由度は比較的増すが、まだ動作の制限は依然として残っているといえる。

音声とジェスチャの同期

音声入力とジェスチャの同期に関しても課題が残っている。第一に、時間的な同期の問題がある。すなわち、音声入力とジェスチャが時間的にずれた場合、ジェスチャによる情報（オブジェクトや領域の指示、動作の方法など）が音声入力に対して適切に反映されないおそれがあるということである。

第二の問題は、音声入力とジェスチャの空間的な同期である。すなわち、音声入力中の参照表現とジェスチャの指示対象の間の対応関係をどのように同定するかということである。

第一、第二の問題とも、今後のより自由な発話とジェスチャが許される音声対話環境においては、非常に重要な課題となると考えられる。

3.3音声対話システムにおけるタスク

マルチモーダル音声対話システムにおいて扱われているタスクとしては、地図などのグラフィクスを用いた特定の事柄に関する質問応答や、二次元あるいは三次元の仮想世界でのオブジェクトの操作が中心である。今後は、計算機側の知識の拡充や仮想現実・臨場感技術の進展にともなって、より高度な対話システムや協調作業支援システムが開発されると考えられる。

参考文献

[阿部] 阿部、大木、寺本、岡田、松下 "VCP：仮想音場空間を利用したコミュニケーションツール" 情報処理学会オーディオビジュアル複合情報処理研究会4-13(94.3.18)

[金沢] 金沢、瀬戸、新地、竹林 "音声自由対話システムＴＯＳＢＵＲＧ｢におけるデータ収集と評価環境" 電子情報通信学会技報SP93-114,NLC93-54(1993-12)

[広明] 広明、國枝、宮井 "臨場感技術とオフィスコミュニケーション" 情報処理学会オーディオビジュアル複合情報処理研究会4-3(94.3.18)

[瀬戸] 瀬戸、永田、新地他 "音声自由対話システムＴＯＳＢＵＲＧ｢の試作" 電子情報通信学会技報SP92-109,NLC92-50(1992-12)

[竹内] 竹内、長尾 "新たなコミュニケーションモダリティとしての表情" 情報処理学会情報メディア研究会9-4(93.1.14)

[竹林] 竹林、永田、瀬戸、新地、橋本 "音声自由対話システムＴＯＳＢＵＲＧ｢-マルチモーダル応答と音声応答キャンセルの利用-" 情報処理学会情報メディア研究会9-3(93.1.14)

[舘森] 舘森、瀬戸、金沢、竹林 "音声自由対話システムＴＯＳＢＵＲＧ｢におけるデータ収集と評価" 情報処理学会情報メディア研究会15-5(94.3.11)

[長尾1] 長尾、竹内 "コンピュータとの自然な対話のための新しいモダリティ-表情つき音声対話システムの試作と実験-" 人工知能学会研究会SIG-SLUD-9204-2(2/5)(1992)

[長尾2] 長尾 "マルチモーダル・ヒューマンコンピュータインタラクション" 人工知能学会研究会SIG-SLUD-9303-4(2/3)

[新田] 新田、正井、岩崎、田中他 "自由発話音声入力と直指（直接指示）を利用したマルチモーダル対話システムの検討" 信学技報SP92-120(1993-01)

[伯田1] 伯田、高橋、小林 "言語・画像情報統合理解の研究" ATRテクニカルレポートTR-C-0007

[伯田2] 伯田、高橋 "地図案内システムＩＭＡＧＥ" ATRテクニカルレポートTR-C-0037（非公開）(1989)

[前田] 前田、Jeong、市川、岡田、松下 "ＭＡＪＩＣ：場の雰囲気を重視したＴＶ会議" 情報処理学会グループウエア研究会5-8(94.1.28)

[望月] 望月、岸野 "自然言語と手指示を統合した３次元仮想空間内での対象物操作と配置" ATRテクニカルレポートTR-C-0082（非公開）(1993)

[吉岡] 吉岡、南、鹿野 "電話番号案内を対象としたマルチモーダル対話システムの作成と音声入力の評価" 電子情報通信学会技報SP93-128(1994-1)

[ローケン・キム] ローケン・キム、谷戸、森元 "マルチモーダル音声翻訳通信のためのシミュレータ" 情報処理学会オーディオビジュアル複合情報処理研究会4-16(94.3.18)

[Allgayer] Allgayer,J.,Jansen-Winkeln,R.,et al. "Bidirectional use of knowledge in the multi-modal NL access system XTRA" Proc. IJCAI-89(1989)

[Arens] Arens,Y.,Hovy,E. and Mulken,s. "Structure and Rules in Automated Multimedia Presentation Planning" Proc. IJCAI-93(1993)

[Bolt] Bolt,R.A. ""Put-That-There": Voice and Gesture at the Graphics Interface" Computer Graphics 14(3) 262-270 (1980)

[Hanne] Hanne,K.-H. and Bullinger, H.-J. "Multimodal Communication:Integrating Text and Gestures" Multimedia Interface Design (ed.)Blattner,ACM Press(1992)

[Koons] Koons,D.B.,Sparrell,C.J. and Thorisson,K.R. "Integrating Simultaneous Input from Speech, Gaze, and Hand Gestures" INTELLIGENT MULTI MEDIA INTERFACE, (ed.) Maybury,AAAI Press/MIT Press(1993)

[Mariani] Mariani,J.J. "Speech in the context of Human-Machine Communication" Proc. ISSD-93(1993)

[Matsu'ura] Matsu'ura,H.,Masai,Y.,Iwasaki,J.,et al. "Applying a spontaneous speech recognizer, a touch-screen,a rule-based speech synthesizer,and photoelectric sensors to a multimodal dialogue system" Proc. ISSD-93(1993)

[McKeown] McKeown,K. and Feiner,S. "Interactive Multimedia Explanation for Equipmennt Maintenance and Repair" Proc. Speech and Natural Language Workshop,DARPA(June,1990)

[Neal] Neal,J.G. and Shapiro,S.C. "INTELLIGENT MULTI-MEDIA INTERFACE TECHNOLOGY" Intelligent User Interfaces, (ed.) Sullivan,J.W. and Tyler,S.W.,ACM Press(1991)

[Stock1] Stock,O. "A Third Modality of Natural Language?" Proc. ECAI 92(1992)

[Stock2] Stock,O. and the ALFRESCO Project Team "ALFRESCO:Enjoying the Combination of Natural Language Processing and Hypermedia for Information Exploration" INTELLIGENT MULTI MEDIA INTERFACE, (ed.) Maybury,AAAI Press/MIT Press(1993)

[Vo] Vo,M.T. and Waibel,A. "Multimodal Human-computer Interaction" Proc. ISSD-93(1993)

[Wahlster1] Wahlster,W. "USER AND DISCOURSE MODELS FOR MULTIMODAL COMMUNICATION" Intelligent User Interfaces, (ed.) Sullivan,J.W. and Tyler,S.W.,ACM Press(1991)

[Wahlster2] Wahlster,W.,Andre,E.,Finkler,W.,et al. "Plan-based integration of natural language and graphics generation" Artificial Intelligence 63(1993)

[Wahlster3] Wahlster,W. "Multimodal Dialog Systems:The Coordination of Vision, Graphics and Speech" Proc.ISSD-93

[Weimer] Weimer,D. and Ganapathy,S.K. "Interaction Techniques Using Hand Tracking and Speech Recognition" Multimedia Interface Design (ed.)Blattner,ACM Press(1992)

[Zancanaro] Zancanaro,M.,Stock,O. and Strapparava,C. "Dialogue Cohesion Sharing and Ajusting in an Enhanced Multimodal Environment" Proc. IJCAI-93(1993)