複数の音声認識器とマーカを用いたマルチモーダルインターフェース

武藤　亮介 †,　嶋田　和孝 ‡,　遠藤　勉 ‡

九州工業大学大学院　情報工学研究科 †
九州工業大学　大学院情報工学研究院　知能情報工学研究系 ‡

1　はじめに

近年，コンピュータを直感的に扱えるユーザインターフェース (UI) が重要視されており，コンピュータを直感的に扱う手段の一つとしてジェスチャを用いるUIの研究が盛んに行われている [1]．しかし，これらは赤外線カメラやデータグローブなどの特殊な機器が必要で，ユーザへの負担が大きい．また，これらはジェスチャのみでしか操作が行えないという点に問題がある．人間はコミュニケーションをする際，ジェスチャ，音声，視線など複数の手段を同時に使用することで意思疎通を図っている．PCを操作する場合にも同様に単一の手段だけではなく，ジェスチャと音声など複数の手段を同時に使用することができれば，より直観的な操作が可能となる．そこで本研究では，一般家庭などでも直感的に作業が行える次世代のUIの構築を目的とし，身近な装置を用いてジェスチャ及び音声理解を行うマルチモーダルUIシステムを提案する．

2　提案手法

本節では提案手法のジュスチャ認識及び音声理解手法について説明する．

なお，本論文では提案するシステムを，写真の選択や拡大などの操作をジェスチャや音声で行う写真管理アプリケーションに適用する場合について考える．

2.1　ジェスチャ認識手法

ジェスチャ認識にはUSBカメラを使用し，色のついたマーカを指先に装着することにより認識を行う．認識の手順には大きく分けてマーカの取得と取得したマーカ位置，軌跡からのジェスチャ認識の2段階ある．

図1　キャリブレーション

まず初めに，キャリブレーションを行い，マーカを正確に取得する．キャリブレーションを行うために認識開始時にユーザに指定の場所に手を合わせてもらい「セット」と発声してもらう (図1)．その時点で，システムはユーザの肌色の取得と背景や服に含まれずマーカのみに存在する色 (マーカ代表色) の抽出を行う．マーカ代表色抽出法を以下に示す．

1. 画像をHSVに変換後，画像全体の次元数を削減

$$binnedh = h * \frac{40}{360}, binneds = s * \frac{20}{255}$$
$$binnedv = v * \frac{10}{255}$$

2. 画像全体の色 $color_i$ に以下の式を用いて重み付け．

$$weight_i = \sum inCircle$$

この時 $color_i$ が指定された場所 (図1) にある円の中にある場合には，$inCircle = 2$ となり，それ以外の場所にある場合は $inCircle = -1$ となる．

3. 以下の式より閾値を算出し，閾値以上出現した色をマーカ代表色とする．

$$threshold = 2 * \frac{sumWeight}{numColor}$$

$numColor$: $weight_i$ が正になった数
$sumWeight$: $weight_i$ が正になった値の総和

画像全体から，抽出したマーカ代表色と肌色と接している所のみ検索し，ユーザの正確な手指情報を取得する．

次に，取得したマーカ位置，軌跡からジェスチャの認識を行う．軌跡の認識はマーカの移動パターンを既定パターンとの連続DPマッチングを用いて認識する．今回ジェスチャで行える操作として以下のものを定義した (図2)．

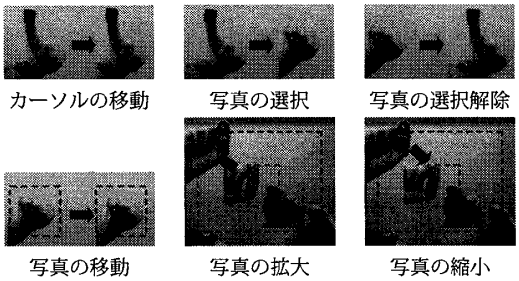

図2　操作とハンドジェスチャの対応

2.2 音声理解手法

音声に関しては，我々の研究室で開発されている複数の音声認識器に基づく，高精度な音声理解手法 [2] を適用する．これは大語彙音声認識器 (LVCSR) とタスク依存音声認識器 (DSSR) から得られる 2 つの出力結果の音素の編集距離を用いて，入力発話がコマンドかそうでないかの分別を行う手法である．この手法の概略を図 3 に示す．今回，DSSR で認識する音声として”拡大”，”ぼかし”や”セット” といった 10 個の音声を定義した．

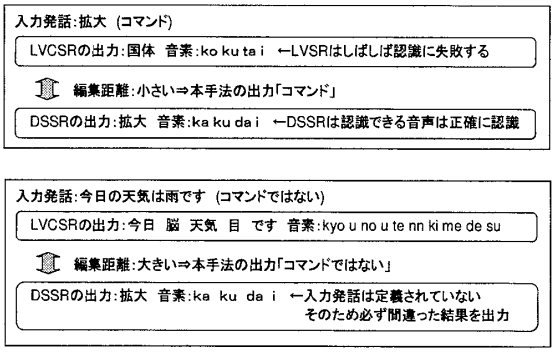

図 3 音声理解手法の概略

2.3 ジェスチャと音声の統合

以上により，ジュスチャと音声の認識が可能となった．しかし，これらを独立で使用するだけでは直観的な操作は行えない．そこで本システムでは，ジェスチャで主に操作を行い，音声はジェスチャで選択されたオブジェクトへのコマンドとして扱うことにより，ジェスチャと音声の統合を行う．

3 実験

これまでに示したように本システムはジュスチャ認識部及び音声理解部により構成される．そこで本節では，それぞれの認識精度を算出し，本システムの評価を行う．

3.1 ジェスチャの認識精度

ジェスチャの認識精度の算出するため，”選択”，”拡大”，”縮小”の各動作を一回で認識した回数を計測した．試行回数は，各動作 20 回づつ．被験者は学生 3 名で行った．その結果を表 1 に示す．

その結果，本手法はシンプルだが高い認識精度を得ることができた．また認識に失敗する理由として，主に 2 つあることがわかった．その一つ目が，図 4 に示すような時のマーカの隠れである．この解決策として，マーカの情報を時系列で見ることが挙げられる．これにより一時的なマーカの隠れによる失敗は回避できる．もう 1 つの理由は，ジェスチャの最中に手をカメラの取得領域外に持っていってしまうということであった (図 5)．これは，ユーザにカメラの取得領域を適切にフィードバックすることにより解決できると考えられる．

表 1
ジェスチャの認識精度

ジェスチャ	精度
選択	97%
拡大	85%
縮小	93%

表 2
音声理解手法の分別精度

ジェスチャ	分別精度
命令発話	87%
非命令発話	100%

図 4 マーカの隠れ

図 5 手がカメラの取得領域外

3.2 音声理解の認識精度

2.2 で定義した命令発話 10 個と非命令発話 15 個を用いて音声理解の認識精度を算出する．非命令発話には”今日の天気は雨です”や”しまった，拡大しすぎた”といった発話を定義した．以上の各発話を 5 回づつ，学生 2 名に発声してもらい，命令発話と非命令発話の分別精度を調べた．その結果を表 2 に示す．また，命令発話の認識精度は 100% となった．

この結果から，非命令発話は確実に非命令発話と分別ができており，雑談中に勝手にシステムが誤動作を起こすといったことは，起こらないと言える．また，命令発話に関しても分別精度 87%，認識精度 100% という結果から，ちゃんと命令発話と分別できれば”拡大”を”縮小”と誤認識しないことがわかった．以上の結果から，本システムの音声理解手法は UI として適切であると言える．

4 おわりに

今回，身近な装置を用いてジェスチャ及び音声理解を行うマルチモーダル UI システムを提案した．ジェスチャの認識には，指先にカラーマーカを装着し，正確なマーカ取得を行うことにより，シンプルな手法ながら高い認識精度を得ることができた．音声に関しては，複数の音声認識器に基づく音声理解手法を適用した．実験より，この音声理解手法は UI として適切であることがわかった．以上の結果より，本システムは UI として有用であると思われる．

今後の課題として，ジェスチャや音声の追加及び認識率の向上が挙げられる．

参考文献

[1] 木村他，広視野電子作業空間に関する考察とシステム試作～マイノリティ・リポート型 I/F とその発展形，インタラクション 2005，PP. 143. 150

[2] Kazutaka Shimada, Satomi Horiguchi and Tsutomu Endo, An Effective Speech Understanding Method with a Multiple Speech Recognizer based on Output Selection using Edit Distance, Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation (PACLIC22), 2008.