

人間の肩の上で動作する計算機の
マルチモーダルインターフェース
に関する研究

2003 年 3 月

園 田 智 也

人間の肩の上で動作する計算機の マルチモーダルインターフェース に関する研究

2003 年 3 月

園 田 智 也

Tomonari Sonoda

目次

第 1 章	序論	11
1.1	本研究の目的	11
1.2	本研究の背景と関連研究	15
1.3	本研究の意義	19
1.4	本論文の構成	21
第 2 章	コビット計算機	23
2.1	まえがき	23
2.2	コビット計算機の定義	24
2.3	本研究におけるコビット計算機の概要	25
2.4	コビット計算機と人間のマルチモーダルインターフェース	25
2.5	交信インターフェースを実現する上での主要な課題	27
2.6	課題の解決手法	28
2.7	むすび	28
第 3 章	空書認識の手法	31
3.1	まえがき	31
3.2	空書認識システム	33
3.3	タッチパネル型インターフェースによる利用者の手の領域の抽出法	37
3.4	輝度差分画像の重心位置による手の移動位置の決定法	42
3.5	連続 DP マッチングを利用した入力文字の特定法	42
3.6	むすび	48
第 4 章	歌声認識の手法	51

4.1	はじめに	51
4.2	歌声による曲検索	52
4.3	検索サーバシステムの定式化	57
4.4	曲検索サーバの旋律情報の照合処理	63
4.5	動的閾値決定法	65
4.6	利用者側のクライアントシステム	71
4.7	むすび	74
第 5 章	コビット計算機システムのモデル	75
5.1	はじめに	75
5.2	コビット計算機のシステムモデル	76
5.3	視聴覚信号処理	78
5.4	信号イベント生成	79
5.5	動作仕様書	80
5.6	信号イベント処理	83
5.7	むすび	85
第 6 章	マルチモーダルインターフェースの実装手法	87
6.1	はじめに	87
6.2	空書認識の実装	88
6.3	歌声認識の実装	89
6.4	マルチモーダルインターフェースの拡張に関する考察	92
6.5	むすび	92
第 7 章	認識手法の評価	95
7.1	まえがき	95
7.2	空書認識の実験	95
7.3	歌声認識の実験	100
7.4	コビット計算機の実時間動作の実験	105
7.5	むすび	107

目次

第 8 章	研究の今後の方向性	109
8.1	まえがき	109
8.2	空書認識に関する研究の今後の方向性	109
8.3	歌声認識に関する研究の今後の方向性	110
8.4	コビット計算機に関する研究の今後の方向性	110
8.5	むすび	111
第 9 章	結論	113

目次

2.1	コビト計算機の概要	26
3.1	空書を行う環境	34
3.2	画像フレーム内に文字を書く様子	35
3.3	タッチパネル型インターフェースの例	39
3.4	手の領域の輝度値の取得用ボタン	41
3.5	手の領域の輝度値でフィルタリングした 2 値化画像	41
3.6	'B' を書いた際の重心 $G(gx, gy)$ の軌跡	43
3.7	連続 DP マッチング	44
3.8	'A','B','C' に対し、'A' を入力した際の距離 D	46
3.9	空中での手書き文字のための英数字の表記法	48
4.1	コビト計算機での曲検索	52
4.2	並列化された計算機上での歌声による曲検索	55
4.3	サーバシステムの拡張性	56
4.4	DP マッチング行列 $m(i, j)$	64
4.5	相対音高差の分布と閾値	66
4.6	相対音長比の分布と閾値	66
4.7	粗いカテゴリを用いたマッチングに見られる問題	69
4.8	Coarse-to-Fine マッチングの例	70
4.9	各音符の音高と音長の推定	72
5.1	コビト計算機システムのモデル	77
5.2	信号イベントのモニタリング	80

5.3	計算機の動作仕様書の例	81
5.4	信号イベント処理	84
6.1	空書認識に対する動作仕様書の実装モデル	89
6.2	歌声認識に対する動作仕様書の実装モデル	91
7.1	赤外線ライト照明によって得られた画像	99
7.2	システム全体のスループットと並列化する曲検索サーバの数の関係	102
7.3	試作したコビット計算機	105

表目次

5.1	視聴覚情報のネットワーク通信データ形式	78
6.1	コビット計算機の信号イベント (空書認識)	88
6.2	コビット計算機の信号イベント (歌声認識)	90
7.1	空中での手書き文字認識率	97
7.2	赤外線照明と蛍光灯照明との認識率の比較	99
7.3	曲データベースの拡張性の評価実験	103
7.4	並列化した計算機上での曲検索実験の結果	104
7.5	コビット計算機システムの画像信号取得のスループットと遅延時間	107

第 1 章

序論

本章では、本研究の目的を示し、関連する研究分野における本研究の位置付けを述べた後、本研究の意義を明らかにする。最後に、本論文の構成を述べる。

1.1 本研究の目的

本研究では、人間の活動を拡張・補助する計算機環境の概念として、**人間の肩の上で動作する計算機**を提案し、人間が、その計算機と自然なコミュニケーション手段で通信できることを目的として、「空書（空中で描く文字）」と「歌声」を用いた**マルチモーダルインターフェース**の実現を目的とする。

本論文では、この計算機インターフェースの手法として、(1) 利用者の空書を、動画像解析によって認識する手法、(2) 利用者が口ずさんだ旋律の曲目を、音声波形解析によって多数の曲目候補の中から特定する手法の 2 つを提案し、さらに、これらの認識手法を計算機上の統一システムで扱う手法を提案する。

人間に装着する計算機が、ビデオカメラやマイクロホンなど、人間の五感に対応する感覚センサーを有している場合、利用者の環境から得られる視聴覚信号を各センサーで取得し、解析処理することで、利用者の活動を拡張・補助することが望ましい。例えば、利用者が目の前にいる人の名前を思い出せない、あるいは街中で聞こえている音楽の曲目が分からないといった場合に、画像処理・音響信号処理した結果にしたがって、適切な情報を示唆する計算機の存在は、有用である。また、その計算機と人間との通信手段として、人間が自然に発する視聴覚信号（ジェスチャや音声など）を用いると便利である。

このように、人間の活動を拡張・補助する装着型計算機の手法として、ヘッドマウント

ディスプレイ (HMD) と装着型ビデオカメラを有したウェアラブル・ビジョン・システムが提案されており、利用者の眼前の光景に応じて生成した注視情報等を、実際の光景に重ね HMD 上に表示することで、人間の視覚能力を拡張する「拡張現実感」が実現された [Jebara et al., 1997, Kourogi et al., 1999, Kurata et al., 2000, Kurata et al., 2001, Murakami 2001]。これらのシステムでは、計算機に対する利用者の入力、主に視覚信号であるジェスチャを認識処理の対象としていたが、ジェスチャ動作に加えて、聴覚信号である音声など、人間にとって自然な交信手段を複合して計算機と対話ができることは望ましい。

一方、外界や人間から得られる様々な感覚信号を複合して認識する手法は、ヒューマノイドロボット (Humanoid-Robot) の分野で研究が進められているが [Hirai et al., 1998, Hashimoto, 2001, Yamasaki et al., 2002, Inoue et al., 2001, Ono et al., 2000]、人間と常に行動をとともにすることが、前提となる手法は知られていない。しかし、前述の「拡張現実感」のように、計算機が人間と行動をとともにすることで、人間本来の知覚能力を超えた有益な情報を提供する手法は、極めて有用であることを、本研究では主張する。

そこで、本研究では、

「人間の肩の上で動作する計算機が、利用者や外界の環境から得られる感覚信号を認識し、人間にとって自然なコミュニケーション手段で交信して、人間の活動を補助・拡張する」

という概念を提案する。また、この計算機のことを、文献 [Merton, 1965] の良く知られた思想

「巨人達の肩の上に乗った小人は、その巨人達よりも遠くを見渡せる」^{*1}

^{*1} 1 つの素晴らしい業績は、過去の多くの素晴らしい業績（巨人達の肩）の上にあることを意味した比喻。英国の学者アイザック・ニュートン (Isaac Newton : 1642/1643-1727) が先輩学者のロバート・フック (Robert Hooke:1635-1703) に送った手紙 (Feb. 5, 1675/1676) の中で、自らの「万有引力の法則」の発見を謙遜して、”If I have seen further, it is by standing on the shoulders of giants.” と表現したことが有名である。

から連想した日本語の「小人」の発音を用いて、「コビット計算機 (kobito computer)」と定義する。

肩の上で動作するコビット計算機は、利用者の頭部と極めて近い位置で各種の感覚信号を取得できる。特に視聴覚信号に関しては、利用者の耳と喉と眼の位置に近いとため、利用者が置かれている環境の状況を認識しやすく、かつ利用者の視聴覚信号（ジェスチャ・音声）を認識する上で環境ノイズの影響を除去しやすいなど、感覚信号を用いた計算機インターフェースを実現する上で、非常に多くの利点がある。また、肩の上は、人間の日常生活を大きく妨げることがなく、望ましい。

このように多くの利点があるにも関わらず、装着型計算機の視聴覚インターフェースは、移動するカメラで撮影した画像の処理問題、音響信号処理の問題の各々が大きな問題であったことが原因で、視聴覚信号を複合的に用いる手法の実現が難しかった。そこで、本論文では、コビット計算機のマルチモーダルインターフェースの実現を目的とし、「視聴覚信号処理」を研究の中心分野に据えて、以下の議論を進める。

人間が日常生活で行う視聴覚情報を交信する事象としては、身振り手振りなどのジェスチャ、顔の表情、話し言葉、歌声、泣き声、手拍子・足拍子などが挙げられるが、計算機と人間との対話という意味では、話し言葉を認識することは重要であり、音声認識の研究は広く行われてきた [Kawahara et al., 2000]。しかし、ジェスチャ認識や音声認識以外の音声信号処理も計算機インターフェースとして重要な交信手段である。

そこで、本研究では、視聴覚情報の認識技術の分野で、特に重要だと考えられる以下の 2 つの分野を対象とする。

1. 空書認識の手法

利用者が空中で文字を描く動作を撮影した画像を解析することで、その文字を特定する手法。この手法により、人間が日常で用いる基本的な情報の単位である「文字」を、コビット計算機で認識することが可能となる。

2. 歌声認識の手法

利用者が街中などで耳にした曲を口ずさむと、曲データベースの中から、その曲目を特定する手法。この手法により、従来用いられてきた話し言葉の認識技術と同様に、音声情報として日常で頻繁に使用される「歌声」をコビット計算機上で認識する

ことが可能となる。

ここで、上述の提案する 2 つの認識手法を実現する上で、以下の課題が挙げられる。

1. 装着型の計算機環境においては、毎回異なる照明条件・背景において、安定して動作検出する手法が必要であるが、利用者が空中で文字を書く際には、その書いた軌跡が残らないため、利用者の文字入力開始・終了や、文字を構成する線分の始点・終点を特定することが極めて難しかった。
2. 装着型の計算機環境で、街中などで聞いた曲目を特定するためには、予め曲データベースにその曲目を登録しておく必要がある。このため、一般には WWW などのネットワーク上に存在する、最新の曲目を含んだ大規模な曲データベースに対して検索することが前提となるが、楽曲の検索では、利用者の入力する旋律情報が曖昧であるため、効果的な旋律情報の索引を作成することが困難であった。このため、高速かつ高い精度で特定することが難しかった。さらに、多くの利用者が共有できる曲検索サービスを実現することが難しかった。

以上の問題に対して、本研究では各々、次の解決手法を提案する。

1. 輝度差分重心位置による手の動作検出法

利用者が空中で描く文字を特定するために、可視光線と赤外線照明の条件下で、画像フレーム中の利用者の手の領域を抽出し、画像フレーム間の輝度差分の重心位置を手がかりとして、空中での利用者の手の動きを追跡する手法。この手法と連続 DP マッチングを併用することで文字を特定することが可能となる。

2. 曲検索サーバの並列化手法

歌声を手掛かりとした曲検索サーバを複数の計算機上で実装し、並列処理によって検索を実現する手法。この手法により、高い検索精度を保ちながら、高速な検索が実現できるようになる。

以上の手法を用いることで、コビト計算機上で、利用者の発する視聴覚信号を認識する。

また、本研究では、以上の 2 つ認識手法に加え、計算機プログラムによる文章読み上げ技術、ビデオカメラの動作制御技術を 1 つのシステム上で扱う「コビット計算機システム」のモデルを提案することで、視聴覚情報を用いたマルチモーダルインターフェースとして統合する。

以上のように、本論文では、コビット計算機の実現を研究の最終目標に据え、人間とコビット計算機の交信手段として、視聴覚情報を複合的に用いる、マルチモーダルインターフェースの実現を研究の目的とする。

1.2 本研究の背景と関連研究

以下では、本研究の位置付けを、視聴覚情報を用いた計算機インターフェースの研究 (1.2.1)、空書認識の研究 (1.2.2)、歌声認識の研究 (1.2.3) の 3 つの観点から明らかにする。

1.2.1 視聴覚情報を用いたマルチモーダルインターフェースの研究における本研究の位置付け

人間と計算機との交信を、ジェスチャ・音声等の視聴覚情報を複合的に用いて行う手法は、マルチモーダルインターフェース (Multi-modal Interface) の研究分野や、人間が通常有する機能を実現することを目的としたヒューマノイドロボットの研究分野において、研究が進められてきた ([Mukai et al., 1998, Hasegawa 2000, Imano et al., 1995, Hirai et al., 1998, Hashimoto, 2001, Yamasaki et al., 2002, Inoue et al., 2001, Ono et al., 2000])。これらの研究では、計算機は、常に利用者の側で動作することを前提とはされておらず、人間とのコミュニケーションは、向かいあう形式であった。

これに対し、本研究が目指す計算機システムは、利用者と常に行動し、利用者が置かれている環境から得られる視聴覚情報（光景・周囲の音）をとらえて、利用者の日常生活を補助し、人間の活動を拡張する能力を有したものである。このような能力を有する計算機手法として、装着型計算機の研究分野では、利用者の頭部近辺に装着したビデオカメラで撮影した映像に、注釈情報や利用者を補助する情報を付加した上で、利用者が装着しているヘッドマウントディスプレイに提示する拡張現実

感の手法が提案された [Jebara et al., 1997, Kourogi et al., 1999, Kurata et al., 2000, Murakami 2001]。[Jebara et al., 1997] では、利用者がビリヤード台の前に立つと、ボールの配置を装着型計算機で認識し、次にどの方向に手玉を打てば良いかの情報を提示した。[Kourogi et al., 1999] では、予め登録した現実世界の対象物の映像が、装着型カメラのフレーム内に入った場合に、その対象物に関する注釈情報を、現実世界の映像に重ねて、ヘッドマウントディスプレイ上に提示した。また、[Murakami 2001] では、拡張現実感のエンターティメント・芸術への応用として、装着型計算機によって、複数の利用者の手の上に仮想的に水場を表示し、お互いの水場に対して、仮想的な魚をやり取りすることができるシステムを開発した。

これらの装着型計算機システムは、映像処理が研究の中心分野であったが、音情報を映像情報と複合的に用いた装着型計算機があれば、利用者の状況を、より多くの信号によって解釈でき、望ましい。

そこで、本研究では、日常生活において利用者のそばで常に動作することを前提とした装着型計算機上で、視聴覚情報を複合的に用いたマルチモーダルインターフェースを実現することを目的とする。

このような前提で視聴覚インターフェースを複合した手法は、これまで知られていない。本研究では、視聴覚情報を用いたインターフェースの要素技術として、「空書認識」の実現手法、および「歌声認識」を認識する手法を提案する。さらに、計算機上でこれらの認識技術を 1 つのシステムとして扱う「コビット計算機システムのモデル」を提案する。

以上のように、本研究は、視聴覚情報を用いた計算機インターフェースの研究において、「肩の上で動作する装着型計算機上で、視聴覚情報を複合したマルチモーダルインターフェースを実現する」という課題を解決することを目的としたものである。

1.2.2 空書認識の研究における本研究の位置付け

空中で描く文字を、計算機上の画像処理によって認識する技術は、ビデオカメラで撮影した画像解析によって文字入力を可能にする手法として、ペンの形状をした小型入力装置を用いる手法 [Yamamoto et al., 1999] や、先端が発光するペンを用いた電子署名の手法 [Katagiri et al., 2001] が提案されている。これらの手法は、特別な機器を手を持つことで、計算機上での認識処理を容易にしたが、本研究が対象とする装着型計算機環境におい

ては、手に何も持たずに認識できるとより便利で望ましい。

これに対して、手に特定の機器を持たずに、空中で描いた英数字を特定する技術も提案されているが [Yoon et al., 1999]、その認識を行うためのビデオカメラは利用者の正面に固定され、肌の色と手の動きの情報を解析に用いたが、本研究が対象とするような肩の上のコピト計算機では、映像は利用者側から撮影するため、利用者の手の動きの解析には、背景画像の影響を考慮する必要がある。また、撮影時の背景や照明条件は、利用する場所によって大きく変化するため、常に肌の色の情報を用いることが可能であるとは限らない。

そこで、本研究では、特定の機器を手に持たないことを前提に、装着型計算機環境で空書認識を実現することを目的とする。

ここで、空中で描かれる文字は、入力開始・終了や文字を構成する線分の起点・終点が明確ないことに加え、画像処理において、以下のことが問題となる。

1. 毎回異なる照明・背景の条件を考慮した上で、文字を書く手の領域を検出することが難しい。
2. 利用者の手は、文字入力最中に様々な方向に傾き、文字を描く軌跡を追跡することが難しい。
3. 実際には見えない文字を、手の動きから特定することは難しい。

これらの問題を解決するために、本研究では以下の手法を提案する。

1. 利用者の手の領域の画素情報を、文字入力の直前に得ることで、手の領域を大まかに抽出する手法。
2. 隣接画像フレーム間の輝度値差分の重心位置によって、手の移動量を解析する手法。
3. 手の移動量を利用した連続 DP マッチングにより、即時に入力文字を特定する手法。

さらに、'K' と 'R' のように、空中で書いてだけでは判別が難しい文字が存在するため、本研究では、空中で文字入力するための表記法についても提案する。

本研究では、これらの一連の手法によって、装着型計算機の環境において、手の動きを認識することが可能となることを提案し、特定の機器を持たずに空書を実現できることを主張する。

以上のように、本研究は空書認識の研究の観点からは、「特定の機器を持たずに、装着型計算機環境で空書を認識する」という重要な課題を解決することを目指したものである。

1.2.3 歌声認識の研究における本研究の位置付け

歌声を入力インターフェースとする技術は、音声認識技術（音声ディクテーション技術や単語認識技術等 [Kawahara et al., 2000]）とは別に、歌声やハミングによる曲検索の分野で研究が進められてきた。旋律によって曲目を特定するという思想は、山本らによって、1980年代から研究されており [Yamamoto 1988]、計算機の演算処理能力の向上に伴って、蔭山らによって、音の高さの違いを有効利用した曲検索手法が計算機上で実装された [Kageyama et al., 1993]。その後、Ghias らによって研究がなされ、音の高さのみを手掛かりにした検索では、曖昧性を考慮した検索の精度が上がらないことが指摘された [Ghias et al., 1995]。

この問題に対して、最近の研究では、適切な閾値によって旋律情報を曖昧にすることで、音の高さの情報だけでなく、音の長さの情報を用いて、精度の高い検索を実現できることが明らかとなっている [Sonoda et al., 1998j, Sonoda et al., 1998]。

しかしながら、多数の曲目の候補に対して、歌声の旋律を手掛かりとして、高速に曲目を特定する際には、以下の問題が存在した。

1. 歌声の認識では、利用者の記憶の違いや歌唱能力によって、予め計算機に登録されている旋律と完全には一致しないため、利用者の入力する歌声の最初の音さえも誤りを含む可能性があるため、検索処理用の索引を作成することが難しかった。このため、多数の曲目の中から、意図する曲目の候補を精度高く高速に絞り込むことが難しかった。例えば、音声認識では、「コンピュータ」という言葉を認識する際に、利用者が最初の「コン」を発音した時点で、認識する言葉の候補を高速に絞り込むことができるが [Kawahara et al., 2000]、歌声の場合は、曲の途中から入力される場合や、最初の発音が正しくない場合も頻繁に起こり得るため、絞り込むことが困難であった。
2. 上述の精度良く高速に曲目を絞り込めない問題に起因して、WWW 上などで、不特定多数の利用者が共有で利用できる曲検索サービスを構築することが難しかった

た。このため、街中で聞いた旋律の曲目を、利用者の歌声を使って、すぐに問い合わせるシステムを構築するためには、各利用者が予めその曲の旋律をシステムに登録しなければならなかった。

上述の問題に対して、旋律情報の索引を作成する手法も明らかとなっており、検索精度をわずかに落としても検索処理時間を飛躍的に短縮することが可能なことが明らかとなっている [Sonoda et al., 2000]。しかしながら、検索精度を高く保ちながら、高速に検索できる手法が存在すると、より実用的で望ましい。

そこで、本研究では、以上の問題を解決する手法として、従来は 1 台の計算機上で実装することが前提となっていた曲検索システムを、複数台の計算機上で実装する手法を明らかにする。また、2 種類の検索データ中継手法を提案することで、論理上、無限に拡張できる曲検索システムの構築が可能であることを明らかにし、拡張性高いシステムを利用することで、多数の利用者で、ネットワーク上の曲検索システムを共有できることを主張する。

以上のように、本研究は、歌声を入力インターフェースとする手法の研究分野においては、「精度が高く、処理時間が短い曲検索手法を、ネットワーク上の多数の利用者によって共有化する」という重大な課題を解決するとともに、従来は考慮されていなかった装着型計算機の環境で、利用者の歌声の認識を実現することを、目指したものである。

1.3 本研究の意義

本研究には、次のような意義がある。

1. 概念的意義

常時、計算機がネットワークに接続される環境・社会において、計算機の新しい役割・概念を示唆する研究である。

2. 工学的意義

人間と計算機とのインターフェースとなる各認識技術において、新しい工学的手法を提案する研究である。

3. 学際的意義

情報科学の多岐にわたる研究分野において、基礎技術・応用技術の広範囲に貢献す

る研究である。

本研究は、第一義的には、計算機の新しい役割・概念を示唆する研究である。計算機が常時ネットワークに接続可能となり、多くの一個人が複数台の計算機を用いるユビキタス・ネットワーク (ubiquitous network)、パーバシブ・ネットワーク (pervasive network) の環境においては、利用者が携帯・装着する計算機が、常に利用者の状態を把握し、利用者の環境状況やネットワーク上の情報を適切に処理して、利用者の活動補助を行うことが望ましい。また、このような計算機ネットワークが前提となる社会では、計算機と人間は、より人間にとって自然なコミュニケーション手段で交信できることが、より多くの利用者にとって障壁の少ない計算機環境を実現する意味で有意義である。コビット計算機は、このような計算機社会の命題に対する一つの答えである。

第二の工学的意義は、本研究が、人間と計算機のインターフェースとして用いる 2 つの視聴覚情報処理の手法が、新しい工学的なモデルを提示している点である。まず、「空書認識」では、非接触型インターフェースによって、利用者と同じ視点で行うという新しい手法を実現する。このため、利用者の手の動きを、肩の上のカメラから撮影して特定し、画像フレーム中の時系列の変量を計測して、筆跡の残らない空書を認識する一連の手法を提案する。また、夜間や照明条件が悪い場合にも赤外線カメラによって空書認識を実現する。「歌声認識」では、多数の曲目の中から、利用者の意図する曲を精度良く高速に求めるため、曲検索システムを任意の規模に拡張できるモデルを提案する。

第三の学際的意義は、本研究が、画像認識、音声認識、情報検索、ロボティクスなど情報科学の多岐の分野に渡り、基礎技術・応用技術に貢献する点である。まず、画像認識の分野では、ハンドジェスチャ認識や手話認識など、人間の手の動きを認識する画像処理の問題に対して、本研究の空書認識手法で提案する一連の認識手法が、従来の固定カメラで撮影した画像に適用された手法や色情報を用いた手法とは異なるアプローチで解決している点が重要である。音声認識の分野では、従来の言葉の認識とは異なり、音声入力される情報の開始位置が極めて曖昧な歌声を用いて、高い精度でパターンマッチングを行う、重要なモデルを示している。情報検索の分野においては、従来のテキスト検索が、クエリとデータベース中のデータが部分一致する検索を実行していたことに対して、本研究が対象としている空書認識と歌声認識の各手法は、必ずしもクエリとデータベース中のデータが部分一致しない「曖昧な情報」を検索するというより高次の課題を、独自の情報照合手法

で解決している。また、ロボティクスの分野では、本研究の各種認識技術を統合するマルチモーダルインターフェースが、人間と計算機との交信のみならず、ロボットとの交信インターフェースとして応用できることが意義深い。

1.4 本論文の構成

本論文は、本章を含め 9 つの章から構成される。

第 2 章では、本論文で提案する**コビット計算機**の概念を述べ、人間とコビット計算機との交信インターフェースについて議論する。

第 3 章では、視覚情報を利用した交信インターフェースとして、**空書認識の手法**を提案し、その技術的な問題点を指摘し、本研究における解決法を議論する。

第 4 章では、聴覚情報を利用した交信インターフェースとして、**歌声認識の手法**を提案し、その技術的な問題点を指摘し、本研究における解決法を議論する。

第 5 章では、**コビット計算機のシステムモデル**を提案し、**マルチモーダルインターフェース**を実現する手法について議論する。

第 6 章では、第 3 章、第 4 章で述べた各認識手法を、第 5 章で述べたコビット計算機のシステム上に具現化する手法を述べ、本システム上の各処理を述べる。

第 7 章では、第 3 章～第 5 章で述べてきた手法について、定量的な評価実験を行った結果を示し、一連の処理モデルの有効性を示す。また、各結果に関して考察を行う。

第 8 章では、第 3 章～第 5 章で提案した研究に関する課題を述べ、今後の研究の方向性を議論する。

第 9 章では、本論文で述べる研究成果のまとめを行う。

第2章

コビット計算機

本章では、コビット計算機概念を述べ、コビット計算機と人間との交信を行うためのマルチモーダルインターフェースについて議論する。また、本研究が対象とする交信手法を構築する際の主要な課題を述べる。

2.1 まえがき

動物は本来、一固体の五感（視覚・聴覚・嗅覚・味覚・触覚）によって知覚される「感覚信号」を処理し、脳内に記憶した後に、互いの経験や意思を、声や動作、臭いなどの五感の信号に置き換えて発することで、情報の伝達や共有を行っている。また、この情報の記録には、可塑性の高い脳内シナプス、あるいは遺伝子などの体内の記憶手段の他に、体外の物体に、臭いや目印を残すなどの手段が用いられ、特に人間は、複雑な情報を体外の記録媒体に記録し、能動的に情報を検索・取得する高度な手段を有している。

人間は、さらに、ビデオカメラやマイクロホンなどの感覚センサーを有した計算機を装着することで、光景や音響などの感覚信号を処理し、人間本来の五感では知覚できない情報を取得することが可能である。このような計算機の実現手法には、様々なモデルが考えられるが、ここでは、利用者の肩の上に乗せて動作する計算機を考えることにする。すなわち、環境から得られる感覚情報を、肩の上に装着している計算機が取得し、人間では知覚・想起が困難、あるいは不可能な情報を提示して、利用者の活動を補助する計算機モデルを議論の中心に据える。

このような装着型計算機では、人間と計算機との交信手法が、モデル構築の上で最も重要な課題となる。この手法には、従来のマウスやキーボードなどの計算機インターフェー

スではなく、ジェスチャや音声など、人間にとって、より自然な感覚情報を複合した**マルチモーダルインターフェース**による交信手法が求められる。しかし、装着型計算機を前提とした場合、各々の感覚情報の認識処理には技術的課題が存在したため、このような計算機インターフェースは実現が難しかった。

本章では、以上の問題を踏まえ、人間本来の知覚能力を超えた感覚情報を認識し、人間にとって自然な交信手法で、人間の活動を補助する計算機概念を**コビット計算機**と定義し、そのマルチモーダルインターフェースの実現手法を議論する。

2.2 コビット計算機の定義

本研究では、以下の要件を満たす計算機を「コビット計算機」として定義する。

- **要件 1**

利用者が日常で、肩の上に装着して利用することを前提とする。

- **要件 2**

人間の感覚器に値する感覚センサーを 1 つ以上有する。

- **要件 3**

人間にとって自然なコミュニケーション手段で交信可能である。

- **要件 4**

感覚信号を認識処理した結果に応じて、利用者にとって有益な情報を提示し、利用者の活動補助を行う。

第 1 の要件は、コビット計算機の利用場面が利用時間・利用場所の条件を問わない日常であることを示す。人間の頭部周辺で用いる計算機は、利用者の眼・耳・喉の近くにあるため、人間の感覚情報の中でも特に重要な視聴覚情報を得やすいという利点がある。また、肩の上で動作する計算機は、日常の活動の大きな妨げにならないため、望ましい。

第 2 の要件は、コビット計算機が利用者から発せられる感覚信号や環境から得られる感覚信号を取得する機能を持つことを示す。

第 3 の要件は、コビット計算機の仕様としての入力インターフェースを示し、利用者からの感覚信号を、入力信号として認識することを前提条件とする。

第 4 の要件は、コビット計算機の存在意義を示している。利用者や環境から得られる感覚信号を処理した結果、利用者にとって有益な情報を抽出し、提示することを目的とした計算機であることを示す。

以下では、以上の定義に基づいて、本論文で対象とするコビット計算機モデルの概要を述べる。

2.3 本研究におけるコビット計算機の概要

本論文では、人間の五感と同様の感覚器を持つコビット計算機を最終的な目標と見据え、その研究の第一歩として、視覚情報と聴覚情報の 2 つを処理対象としたコビット計算機を考える。この 2 つの情報を対象とする理由は、人間の眼と耳によって得られる視聴覚情報は、五感の感覚情報の中でも、言葉による概念を形成する上で極めて重要であり、コビット計算機が人間の活動の補助を行う上で、非常に大きな役割を果たす点にある。

図 2.1 は、視聴覚情報の処理を前提としたコビット計算機の利用イメージを示している。本研究では、コビット計算機が有する感覚センサーとして、ビデオカメラ（視覚センサー）とマイクロホン（聴覚センサー）の 2 つを利用者の肩の上に設置し、各感覚センサーから得られる視聴覚信号を、装着したデータ処理装置に送信する。データ処理装置では、音響信号解析と動画解析を行い、利用者と環境から得られる視聴覚信号を認識する。ここで、利用者から得られる視覚情報は、利用者の手の動きを撮影した画像を解析して得られる情報、聴覚情報は、音声を対象とする。コビット計算機は、次に、視聴覚情報の認識結果にしたがって、コビット計算機の動作信号と発話信号を生成し、データ処理装置から、肩の上のシステムの駆動部とスピーカーに対して各々送信することで、利用者に対して、視聴覚信号を使って交信する。また、利用者は、コビット計算機の視点で得られる映像を確認するための液晶ディスプレイ（LCD）を、体に装着し、補助的に用いることが可能である。

2.4 コビット計算機と人間のマルチモーダルインターフェース

前節で述べたコビット計算機の全体の処理において、視聴覚センサーで得られる視聴覚信号を、高い精度で認識することは、コビット計算機システム全体の性能を決定する上で、最も重要な課題である。特に、利用者が発する視聴覚信号の認識は、利用者とコビット計算機

の交信インターフェースが必要となるすべてのアプリケーションにおいて重要である。そこで、本研究では、提案するコビット計算機と人間との交信インターフェースの実現手法に注目して、以下の議論を進める。具体的には、次の 2 つの手法を提案する。

1. 空書認識

まず、利用者の発する視覚情報として、利用者が空中に描く文字を認識する手法を提案する。本手法は、人間が通常扱う文章を構成する最小単位の文字情報を、計算機に入力できる点で有益な手法である。

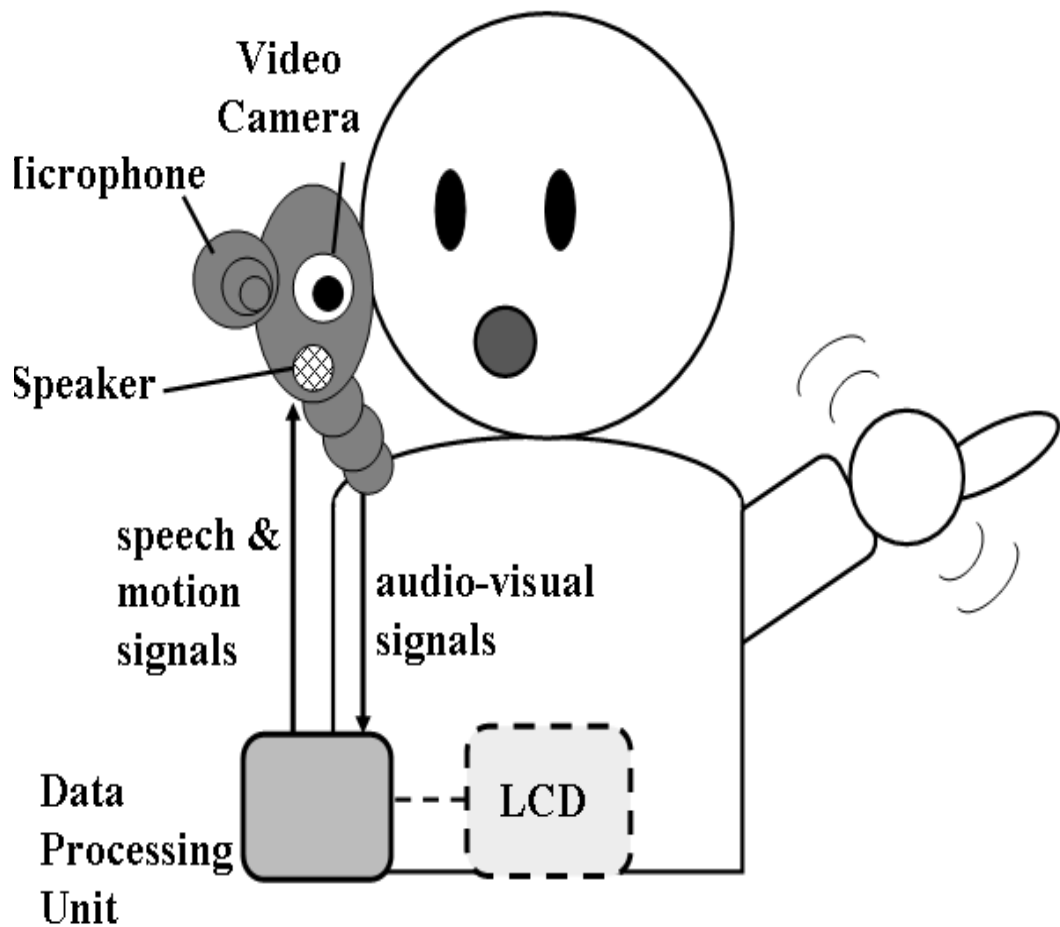


図 2.1: コビット計算機の概要

2. 歌声認識

次に、利用者の発する聴覚情報として、利用者が記憶している曲の一部を口ずさむと、その曲目を特定する手法を提案する。歌声は、話し言葉と同様に、日常で頻繁に発せられ、この認識手法を装着型計算機に応用することで、街中で聞いた曲をその場で調べることが可能となるなど、極めて有益な手法である。

本論文では、以上の 2 つの手法の具体的な実現方法を、各々第 3 章 (空書認識の手法)、第 4 章 (歌声認識の手法) で明らかにし、これらを統合することで、人間とコビット計算機とのマルチモーダルインターフェースを実現することを目的とする。

2.5 交信インターフェースを実現する上での主要な課題

前節で提案した空書認識と歌声認識を実現する上で、各々、以下のような技術的な課題が挙げられる。

1. 毎回異なる背景・照明条件での手の動作の特定問題

装着型の計算機環境においては、夜間も含め、毎回異なる照明条件での利用を考慮する必要があった。また、撮影するビデオカメラも完全には静止しないという条件も勘案する必要があった。さらに、利用者が空中で文字を書く際には、その書いた軌跡が残らないため、利用者の文字入力 of 開始・終了や、文字を構成する線分の始点・終点を特定することが極めて難しかった。

2. 多数の候補からの曲目の特定問題

装着型の計算機環境では、予め利用者個人が持つ曲データベースの他に、常に最新の曲が存在するネットワーク上の曲データベースに問い合わせる必要がある。しかし、楽曲の検索では、利用者の入力する旋律情報が曖昧であるため、効果的な旋律情報の索引を作成することが困難であった。このため、高速かつ高い精度を維持して、多数の候補を含むデータベースから、曲目を特定することが難しかった。また、多くの利用者で共有して利用できる曲データベースを有したオンライン検索サービスを実現することが難しかった。

以上の各認識技術に関する問題に対して、本研究では次節で述べる手法で解決を図る。

2.6 課題の解決手法

前節で掲げた各問題に対して、本研究では各々以下の解決アプローチを提案する。

1. 輝度差分重心位置による手の動作検出法

利用者が空中で描く文字を特定するために、(1) 可視光線および赤外線照明条件のもとで、画像フレーム中の利用者の手の領域を抽出する手法、(2) 抽出した手の領域の輝度を用いて、画像フレーム間の輝度差分の重心位置を求める手法、(3) その輝度差分の重心位置を手がかりとして、空中での利用者の手の動きを、連続 DP マッチングによって追跡する手法の 3 つを提案する。また、空中で描かれる文字は一筆書きになることを考慮した、空書に適した英数字の表記モデルを提案する。

2. 歌声による曲検索の並列処理手法

楽曲検索を複数の計算機上で並列に行う手法を提案する。旋律情報の索引では、数パーセントの検索精度の低下を回避することが難しかったことに対して [Sonoda et al., 2000]、本手法は、高い楽曲検索精度を保ちながら、高速な検索サービスを多数の利用者に提供することを可能にする。

以上の解決法に関する具体的な手法は、上述の各認識手法と同様に、各々第 3 章、第 4 章において明らかにする。

2.7 むすび

本章では、コビット計算機を定義し、人間とコビット計算機のマルチモーダルインターフェース実現の重要性を述べた。また、そのインターフェース構築のための具体的な視聴覚情報の認識手法として、空書認識と歌声認識の 2 つを提示した。以下に本章のまとめを示す。

- コビット計算機は、(1) 利用者が日常で、肩の上に装着して利用することを前提とし、(2) 人間の感覚器に値する感覚センサーを 1 つ以上有し、(3) 人間にとって自然なコミュニケーション手段で交信可能であり、(4) 感覚信号を認識処理した結果に応じて、利用者の活動補助を行う計算機として定義する。
- 視聴覚情報を用いたコビット計算機モデルの実現には、人間とコビット計算機の交信手法の実現が重要である。
- 人間とコビット計算機の交信手法として、本論文では、「空書認識」と「歌声認識」という 2 つの具体的な認識手法を明らかにし、コビット計算機のマルチモーダルインターフェースの実現を目的とする。

第 3 章

空書認識の手法

本章では、コビット計算機の入力手法として、利用者が空中で英数字を書く動作を行うと、その文字を計算機に入力できる認識システムを提案する。本手法では、装着しているビデオカメラによって、利用者の手の動作を撮影し、計算機上で白黒階調画像の解析を行うことで文字入力を実現する。また、空中での文字認識が根本的に抱える問題を解決するために、英数字の表記法を提案する。

以下では、本システムの概要・仕様を述べ、前提とする撮影条件のもとで、空書を認識する際の問題を定義する。次に、その問題に対する解決法を提案し、空書に適した英数字の表記モデルを議論する。

3.1 まえがき

装着型計算機的环境では、ヘッドマウントディスプレイ（HMD）と装着型ビデオカメラを用いることで、利用者の目の前の光景を撮影し、注釈などの付加情報を重ねた上で、HMD 上に表示することが可能であり [Jebara et al., 1997, Kourogi et al., 1999]、仮想現実空間、拡張現実空間の研究で入力インターフェースの研究が進められている。このような計算機環境では、場面に応じた様々な情報を入力できることが望ましいため、装着しているビデオカメラによって撮影した画像を解析することで、画像中の人物の顔を認識する手法や、手の動きを用いたマウスの構築手法が提案されている [Kurata et al., 2000, Kurata et al., 2000-2]。これらの手法では、目の前の光景に結び付いた入力となってきたが、画像解析によって、計算機の制御や文字入力できれば、メールや ATM の暗証番号の入力などが実現し、望ましい。一方、ビデオカメラで撮影

した画像解析によって文字入力を可能にする手法として、ペンの形状をした小型入力装置を用いる手法 [Yamamoto et al., 1999] や先端が発光するペンを用いた電子署名の手法 [Katagiri et al., 2001] が提案されているが、機器を手を持たずに入力できるとより便利である。

そこで、本研究では、利用者が空中で任意の文字を書く動作 (**空書**) を行くと、その動作を撮影した画像を解析することで、その文字を計算機に入力できるシステムを提案する。本システムは、利用者が日常生活の中で、日中の屋内外の自然光・室内灯、あるいは暗闇で装着型計算機を利用することを前提として、可視光照明と赤外線照明の 2 種類の照明条件の下で使用できるものとする。

ここで、空書の文字を認識する手法として、文献 [Yoon et al., 1999] では、固定カメラで利用者を正面から撮影し、肌の色と手の動きの情報を解析に用いたが、本研究が対象とする装着型計算機環境では、映像は利用者側から撮影するため、利用者の手の動きの解析には、背景画像の影響を考慮する必要がある。また、撮影時の背景や照明条件は、利用する場所によって変化するため、肌の彩度や輝度のモデルを導入する場合には、動的に取得する必要がある。

さらに、空中で描かれる文字は、入力開始・終了や文字を構成する線分の起点・終点が明確ないことに加え、画像処理において、以下のことが問題となる。

- (P1) 可視光・赤外線照明の照明条件、および毎回異なる背景条件を考慮した上で、文字を描く手の領域を検出することが難しい。
- (P2) 利用者の手は、文字入力の最中に様々な方向に傾き、利用者の手が描く軌跡を追跡することが難しい。
- (P3) 実際には見えない文字を、手の動きから特定することは難しい。

これらの問題を解決するために、本研究では以下の手法を問題提案する。

- (S1) 利用者の手の領域の画素情報を、文字入力の直前に得ることで、手の領域を大まかに抽出する手法。
- (S2) 隣接画像フレーム間の輝度値差分の重心位置によって、手の移動位置を解析する手法。
- (S3) 手の移動位置を利用した連続 DP マッチングにより、即時に入力文字を特定する

手法。

さらに、'K' と 'R' のように、空中で書いただけでは判別が難しい文字が存在するため、本研究では、米国 Palm 社の Graffiti(TM) を参考にし、空中で手書きするための文字表記法を設計した。

3.2 空書認識システム

以下では、本システムの概要、撮影環境、および空中での手書き文字認識で用いる画像処理の手順を述べる。また、処理上の問題点に対する本研究での解決手法を提案する。

3.2.1 本システムの仕様

本システムでは、利用者の手が描く軌跡を、図 3.1 に示すようなコビット計算機的环境中において解析することで、文字入力を可能とする。本システムでは、利用者は、装着型のビデオカメラ、ディスプレイおよび計算機を身に付けている。ビデオカメラでは、利用者が実際に見ている光景が撮影され、計算機に映像が送られる。ディスプレイには、撮影された映像に重ねて、計算機上で生成された注釈情報が表示される。図 3.2 は、本システムの利用者のディスプレイ上の表示画像を示す。本研究が対象とするコビット計算機的环境中では、第 6 章で述べるように、音声発話手法と組み合わせることで、利用者が入力する文字を音声で表現することも可能であるが、ここでは、利用者のディスプレイ上に表示することで利用者が自分の描いた文字の認識結果を確認するものとする。図 3.2 では、計算機上画像中の右上の文字は、利用者が入力できる文字の種類を'A'(英字) と'0'(数字) で示し、左上の文字は、最後に入力された文字を示す。本システムでは、まず、利用者はビデオカメラの画像フレーム内に収まるように、手に何も付けずに人差し指あるいは親指を立て、空中で 1 つの文字を書く動作を行う^{*1}。この際、文字の入力開始は、ディスプレイの中央に表示される仮想的なボタンを押す動作を行うことで認識され、文字を書き終った時点で手を画像フレームの外に出すものとする。撮影された画像フレームは、連続的に画像処理部である計算機上で解析される。システムは、ある時点までの利用者の手が描く動作パ

^{*1} この制約は、本手法が指先を認識しているものではないことから、システムの精度に影響を与えるものではないが、通常、利用者が空書を行う際には、指先で文字を書くことが多いと考えられるため、このような仕様としている。

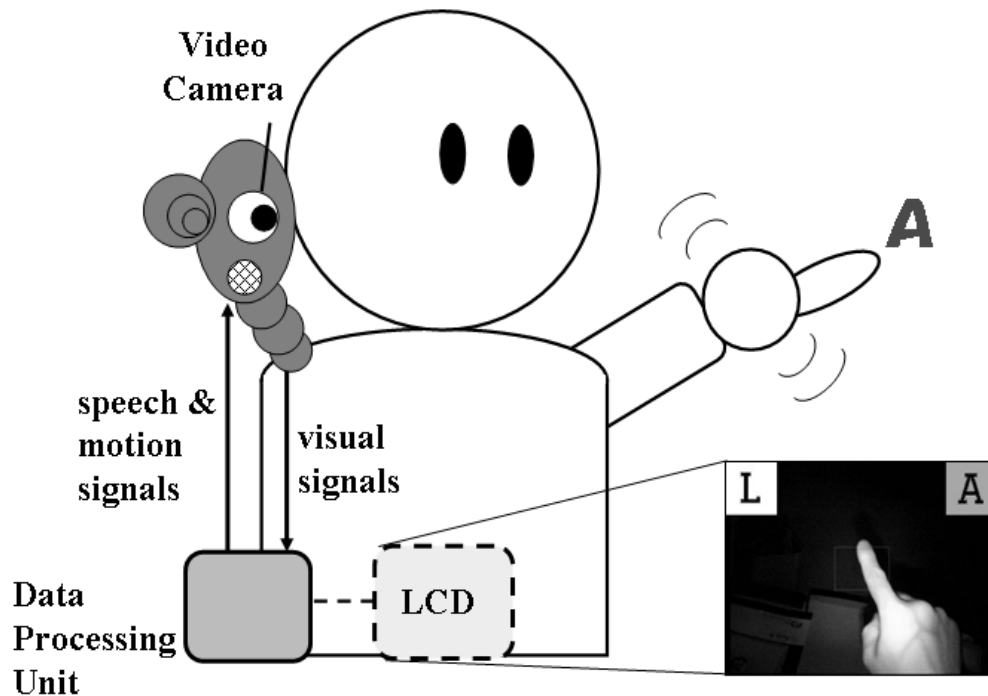


図 3.1: 空書を行う環境

ターンが、計算機の記憶領域にあらかじめ登録されている各文字の動作パターンの中の 1 つと一致したと判定した場合、その文字が入力されたと見なす。

ここで、利用者の文字の書き順などが、あらかじめ登録している文字のものとは異なる場合は、利用者は新しく自分の描くパターンを登録することができる。また、1 つの文字に対して、複数のパターンを登録することが可能であり、その際には最も類似しているパターンが優先となる。

このシステムでは、さらに星型やハート型など、様々な形を描くパターンを記憶できるため、文字入力以外の計算機制御にも応用できる。

以下では、本システムの撮影条件と画像処理の手順、および処理上の問題点と、それに対する本研究での解決手法について述べる。



図 3.2: 画像フレーム内に文字を書く様子

3.2.2 撮影条件

本システムでは、文字入力の際に、背景には移動物体が写っていないことを前提とし、カメラはほとんど揺らさないこととする。この制約は、利用者にとっては、地面や壁を向いて利用する等の動作で容易に実現できるため、大きな負担にはならず、また、システムは精度の高い認識を行うことができる。さらに、背景と手は肉眼でははっきりと区別できるほど、明るさが違うものとする。利用者の手とカメラとの距離は約 30cm～50cm 程度とする。

照明は、特に制約を設けず、自然光、屋内外の灯等で撮影された画像内で、利用者の手の位置が判別できる程度の明るさとする。また、夜間でも赤外線ライトをカメラ側から照射することで、手の動きを判別できる [Sonoda et al., Sept.2001]。この際、カメラは赤外線の光を撮影できるものとし、赤外線ライトとカメラとの相対位置は固定する。

以上の環境で撮影する画像に対して、以下に述べる処理を行うことで、手の描く軌跡パ

ターンを認識する。

3.2.3 画像処理上の問題と解決手法

本システムの計算機上での画像処理の手順と問題点、およびその解決手法を述べる。

利用者が空中で書いた文字を認識する上で考慮しなければならないことは、利用者が文字を書き始めた時点と書き終えた時点をいかにして検出し、その間の手の動きをいかにして読み取るかである。

ここで、本システムの仕様では、利用者は、文字を書き終えた時点で画像フレームの外へ手を出すので、その手の存在を検出すれば良いが、前提とする撮影条件では、背景は完全には静止しておらず、撮影場所も毎回異なるため、照明条件や背景が一定でない。そこで、これらを考慮し、本システムの画像処理は、カメラからの画像フレーム入力があるたびに、以下の手順で行う。

- (a) 画像フレーム中の利用者の手の領域の検出。
- (b) 前回の画像フレームとの間での手の動きの検出。
- (c) 現在の時刻までに描かれた文字の特定。

ここで、3.1 で述べた、(P1)~(P3) の問題を考慮する必要がある。

(P1) の問題は、利用者の手の存在が大まかに分かれば良いので、隣接画像フレーム間の輝度値などの差分を検出する手法が考えられるが、カメラが完全に固定されていないため、その誤差を完全に除くことは難しい。そこで、予め肌の色や明るさ (輝度) を登録して、その位置を特定する手法が考えられるが、使用する場所や照明条件ごとに、その値は異なるため、使用時にそれらの情報を特定できることが望ましい。

(P2) を解決する手掛かりとして、隣接画像フレーム間で、輝度値の差分から手の動きを検出する手法 [Oka et al., 1999] や、光学的なフローを検出し、画像フレーム全体のフローと大きく異なる箇所を、手の動きと見なす手法が考えられるが [Kubota et al., 1998, Kouroggi et al., 1999]、検出した手の動きの領域から、文字を書く軌跡パターンを、どのように検出するかが課題となる。

(P3) の問題は、利用者の文字入力の開始と終了の間の軌跡を手がかりとし、最も類似するパターンの文字を決定すれば良い。しかしながら、'K' と 'R' のように、空中で書き

た軌跡だけでは、判別が困難な文字も存在する。

以上を考慮し、本研究では各々の問題・課題に対する解決法として、以下の各手法を提案する。

1. 利用者の見ているディスプレイ上に、仮想のタッチパネルのボタンを提示し、利用者がそのボタンを押した際に、利用者の手の領域の情報 (輝度値) を取得し抽出する手法。この手法により、取得した輝度値とは大きく異なる画素値を、処理の対象から削除することが可能となり、以下の処理において、精度の高い結果を得ることができる。
2. 隣接画像フレーム間で、同位置の輝度値差分を取り、その差分画像における、値の重心位置 $G(gx, gy)$ によって、利用者の手の移動位置を決定する手法。この手法により、計算量の多い光学的なフローの解析処理を用いずに利用者の手の軌跡を解析できる。
3. 連続 DP マッチングによる文字パターン特定法。本手法では、上記の処理で得た重心位置の変量を用いて、高速な入力文字の認識処理を実現する。

以下、**3.3** では、タッチパネル型インターフェースによる利用者の手の領域の抽出法、**3.4** では、輝度差分画像の重心位置による手の移動位置の決定法、さらに **3.5** では、連続 DP マッチングを利用した入力文字の特定法を述べ、空中で書いた軌跡だけでは特定できない文字のために、本システム独自の英数字の表記法を提案する。

3.3 タッチパネル型インターフェースによる利用者の手の領域の抽出法

以下では、手の領域を抽出する際の問題を述べ、提案するタッチパネル型インターフェースによる解決法を述べる。

3.3.1 手の領域の抽出問題

ジェスチャ認識の処理において、画像フレーム内の動きのある領域を抽出することは、システム全体の性能に大きな影響を与えるため重要である。ここで、固定されたカメラでは、隣接画像フレーム間の差分を取ることで、その領域を抽出することが可能である

[Mukai et al., 1998]。

しかし、本システムの環境では、ほとんど静止した状態で撮影していても、利用者のわずかな揺れにより、背景部分が動き、差分画像において、利用者の手の領域だけを優先的に抽出することは難しい。

そこで、本研究では、文字を入力する直前に、利用者の手の明るさ (輝度値) を取得し、それに近い値を持つ画素のみを処理の対象とすることで、手の領域を大まかに抽出する手法を提案する。この際、輝度値を確実に取得するために、ある瞬間に、画像フレーム内の絶対的な位置に手の領域が重なっていることが望ましい。そこで、本研究では、利用者が任意の瞬間に手の領域の輝度値を取得する手法として、利用者の装着しているディスプレイ上にタッチパネルを表示し、そのボタンを手の動きによって押すことができるインターフェースを提案する。利用者は文字の入力の直前にタッチパネルのボタンを軽く触れるように手をかざすことで、容易に輝度値を取得できる。

なお、現行のシステムは、赤外線画像を勘案し、白黒階調画像を用いて処理しているため、本手法を輝度値の獲得に用いているが、予備実験から、肌の色情報 (色相値) を用いることができる環境では、さらに精度の高い処理が行える。

以下では、本研究で提案するタッチパネル型のインターフェースについて述べる。

3.3.2 タッチパネル型インターフェース

図 3.3 に、本研究で提案するタッチパネル型インターフェースの例を示す。この図では、仮想的なタッチパネルボタンを 3 つ (Button1~3) 表示させている。この手法では、利用者は、ディスプレイ上に示されたボタンを押すように、指先をボタンに重なるように移動させることで入力が行える。

本インターフェースの構築の際には、利用者がボタンを押す際に変化する画像中の輝度に着目する。具体的には、隣接フレーム間での輝度値の変化量の和をボタン領域ごとに求め、ある閾値以上の変化量を持つボタン領域には、利用者の手が重なっていると見なして処理を行う。すなわち、時刻 t における画像フレーム F_t 上の各画素の輝度値を $I(t, x, y)$ とし、タッチパネル上の N 個のボタン領域を B_n ($n = 0, 1, \dots, N - 1$) としたときに、 B_n 上での輝度の変化量 S_n を次の式で求める。



図 3.3: タッチパネル型インターフェースの例

$$S_n = \sum_i \sum_j |I(t, i, j) - I(t-1, i, j)|, \\ I(t, i, j), I(t-1, i, j) \in B_n$$

ここで、 $S_m = \max\{S_n | 0 \leq n < N\}$ が、

$$S_m > T * S_n \quad (T \text{ は定数. } 0 \leq n < N-1 \text{ and } m \neq n.)$$

を満たすとき、 B_n が押されたと判定する。

ここで、本研究の予備実験において、次のような誤入力が見られた。

- (A) あるボタンを 1 回だけ押したい場合にも、連続して入力される現象。
- (B) 背景画像の影響で、意図していないボタンが自動的に押されてしまう現象。

3.3. タッチパネル型インターフェースによる利用者の手の領域の抽出法

(A) の現象は、入力操作の履歴を、時間軸によって管理することで、連続する入力を容易に制限できる。

(B) の現象は、カメラのぶれによって起こる現象である。そこで、被験者 1 名に対して、カメラを動かさないように指示を与えた上で、図 3.4 の背景で、空書の文字入力の予備実験を行ったところ、文字を書く際の実際のカメラのぶれは、縦横に最大で 19 ピクセル発生したが、上述した B の現象の発生率は、実験で取得した動画像 6,400 フレーム中 4 フレームで、0.0625% であった（ボタン押下判定に用いる T の値は 5 とした）。これは、通常の利用条件下において、利用者に許容でき得るものであり、ボタン押下判定は十分な性能が得られる。

本システムでは、利用者がボタンを一度押した後は、意図しないボタン入力が発生しないように、利用者の入力がない状態が一定時間続かない限り、再度ボタンを押すことができないようにしている。

本システムでは、このインターフェースを用いることで、任意の瞬間に利用者の手の領域の輝度値を取得できる。

3.3.3 手の領域の輝度値の取得方法

図 3.4 は、画像フレームを縦横に 5*5 の領域に分割したものであり、図 3.4 中の縦横の白線は、その分割線となっている。本システムでは、この 25 個に分割された領域のうち、中央の領域を、手の領域の輝度値を取得するためのものとして利用しており、他の 24 個の分割領域をその他のボタンと見なして、利用者がボタンを押しているかどうかの判定を行う。

ここで、本システムでは、利用者が中央のボタンを押した瞬間に、画像フレーム中の特定の領域 R に手の領域が重なっていることを前提とする。本システムでは、 R の位置は利用者が右利きの場合は、25 分割した一番右下の矩形領域としており、左利きの場合は一番左下の矩形領域としている。ボタンが押された際には、システムは R 上の画素値に関するヒストグラムを作成し、最も度数の大きな画素値 v を求める。その後、現在の画像フレームにおいて、 $v \pm \epsilon$ を満たす値を持つ画素のみを残し、残りの画素をすべて処理の対象外と見なす。図 3.5 は、図 3.4 のボタンを押した結果得られた v に対して、 $v \pm \epsilon$ ($\epsilon = 10$) を満たす画素とそうでない画素を 2 値化して表示したものである。 $v \pm \epsilon$

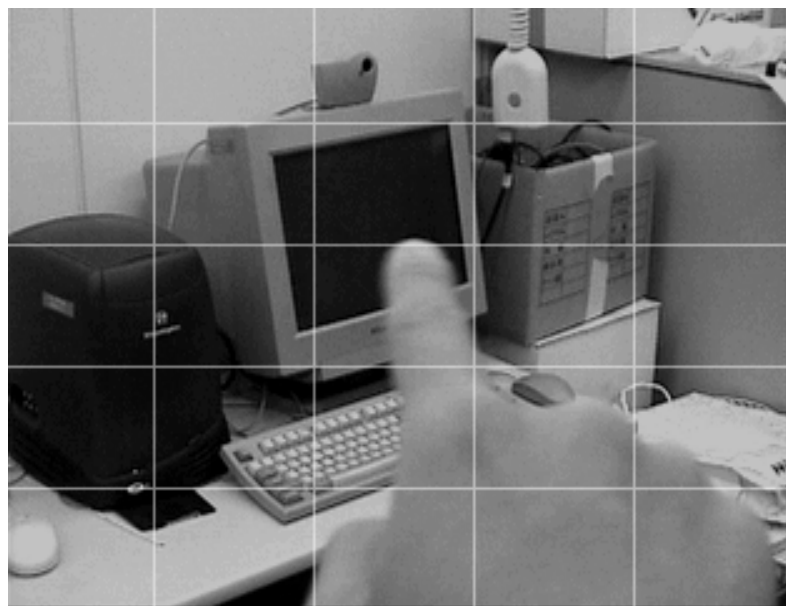


図 3.4: 手の領域の輝度値の取得用ボタン

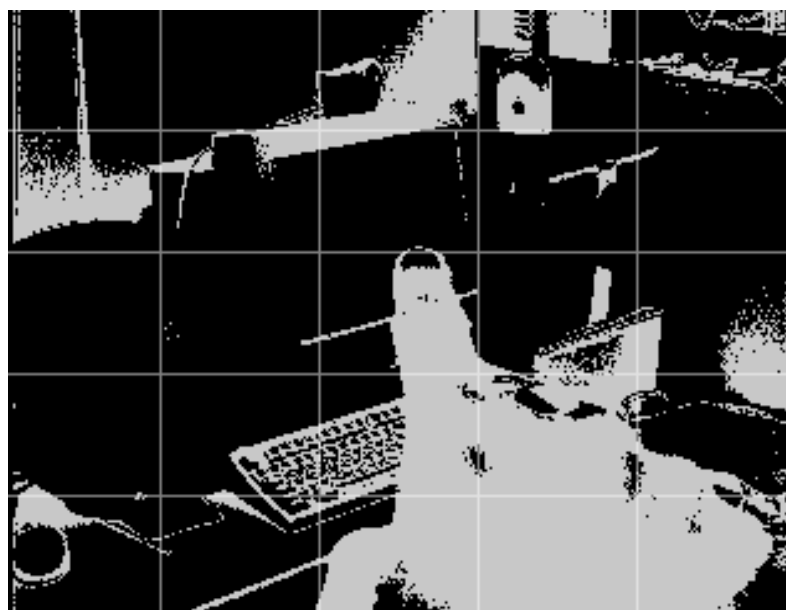


図 3.5: 手の領域の輝度値でフィルタリングした 2 値化画像

を満たすものを白く表示している。

ここで、本システムでは $v \pm \epsilon$ で残した画素について、隣接画像フレーム間で差分をとり、その総和がある閾値 $Hand$ を越える場合には、利用者の手が画像フレーム内にあり、文字を書いている最中だとみなし、そうでない場合には、文字入力を行っていないと見なす。

以下では、このようにして得られた画像を用いて、手の描く軌跡を特定する。

3.4 輝度差分画像の重心位置による手の移動位置の決定法

本手法では、利用者の手の動きを追跡する手法として、隣接画像フレーム間の輝度値差分の絶対値を重みとした、画素値の重心位置を用いる手法を提案する。すなわち、重心位置の遷移パターンによって、入力された文字を特定する。

ここで、時刻 $t, t-1$ における画像フレーム F_t, F_{t-1} 間の各画素の輝度値の差分 $|I(t, x, y) - I(t-1, x, y)|$ を重みとしたとき、その重心の位置 $G(gx, gy)$ は、次の一連の式で求められる。

$$\begin{aligned} X &= \sum_i \sum_j i * |I(t, i, j) - I(t-1, i, j)| \\ Y &= \sum_i \sum_j j * |I(t, i, j) - I(t-1, i, j)| \\ W &= \sum_i \sum_j |I(t, i, j) - I(t-1, i, j)| \end{aligned}$$

これより、

$$\begin{aligned} gx &= X/W \\ gy &= Y/W \end{aligned}$$

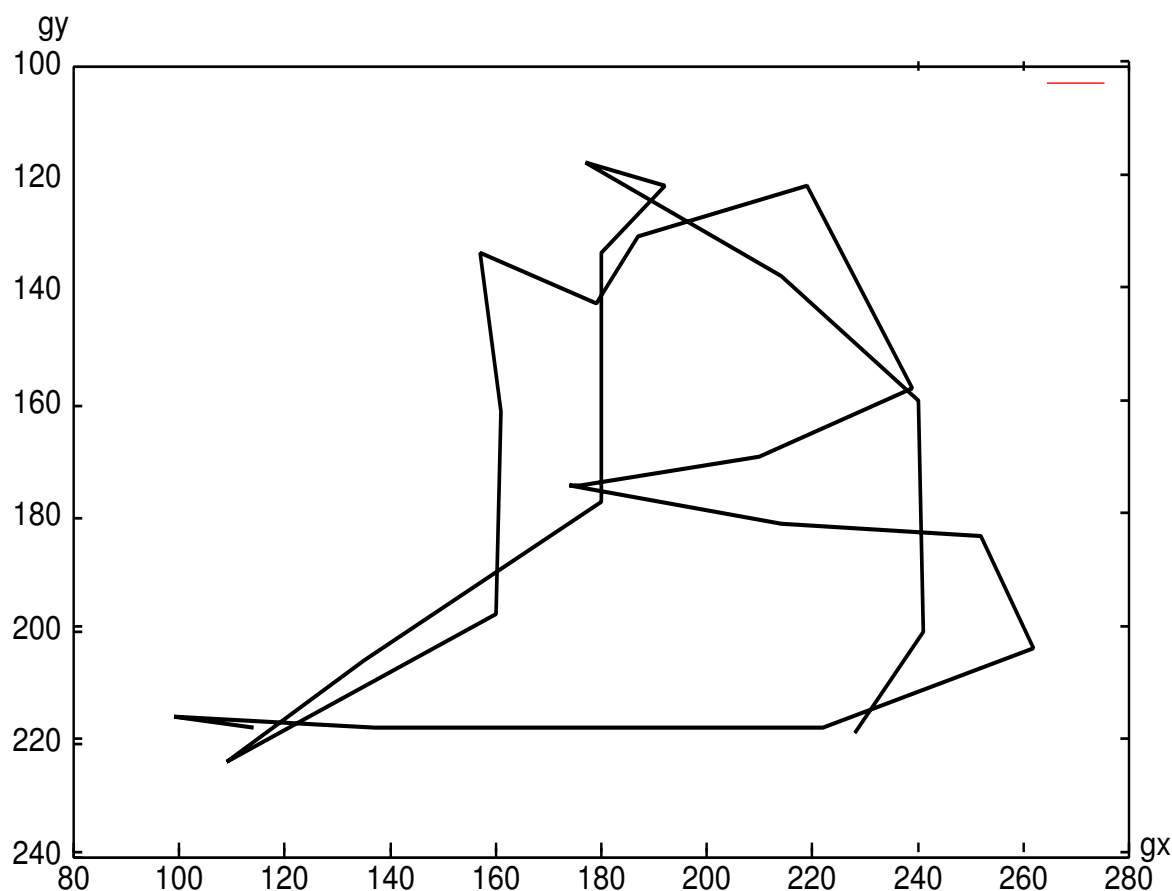
本システムでは $I(t, i, j)$ として、上述の処理で得られた、手の領域の輝度値に近いものは、その値を用いており、それ以外では値を 0 としている。

図 3.6 は、以上の式により求めた、アルファベットの 'B' を描いた際の重心位置の軌跡を示している。この図が示すように、重心位置 $G(gx, gy)$ の軌跡は、利用者の手の動きに伴っている。

以上で得られる G の遷移パターンは、次に述べる連続 DP マッチングにより、DB のパターンと照合する。

3.5 連続 DP マッチングを利用した入力文字の特定法

以下では、利用者に描かれる文字を、連続 DP マッチングにより認識する手法を述べる。

図 3.6: 'B' を書いた際の重心 $G(gx, gy)$ の軌跡

3.5.1 連続 DP マッチングによる距離の測定

本手法では、空中で描かれた文字を認識する際、画像フレーム間の輝度差分の重心位置の遷移パターンを、入力とデータベース (DB) との間で照合する。この照合には、ジェスチャ認識等でよく用いられる連続 DP マッチングを用いる [Takahashi et al., 1994, Mukai et al., 1998, Oka et al., 1999]。同様の手法として、HMM(hidden Markov model) を用いることも考えられるが [Yoon et al., 1999]、本システムでは学習させるデータが音声認識処理などと比較して少ないため、処理がより単純な DP マッチングを用いている。

ここで、利用者の入力における重心位置の系列を x, y 方向でそれぞれ Q_x, Q_y とし、同様に、長さ I の DB 系列における重心位置の系列を S_x, S_y とする。それぞれの系列の

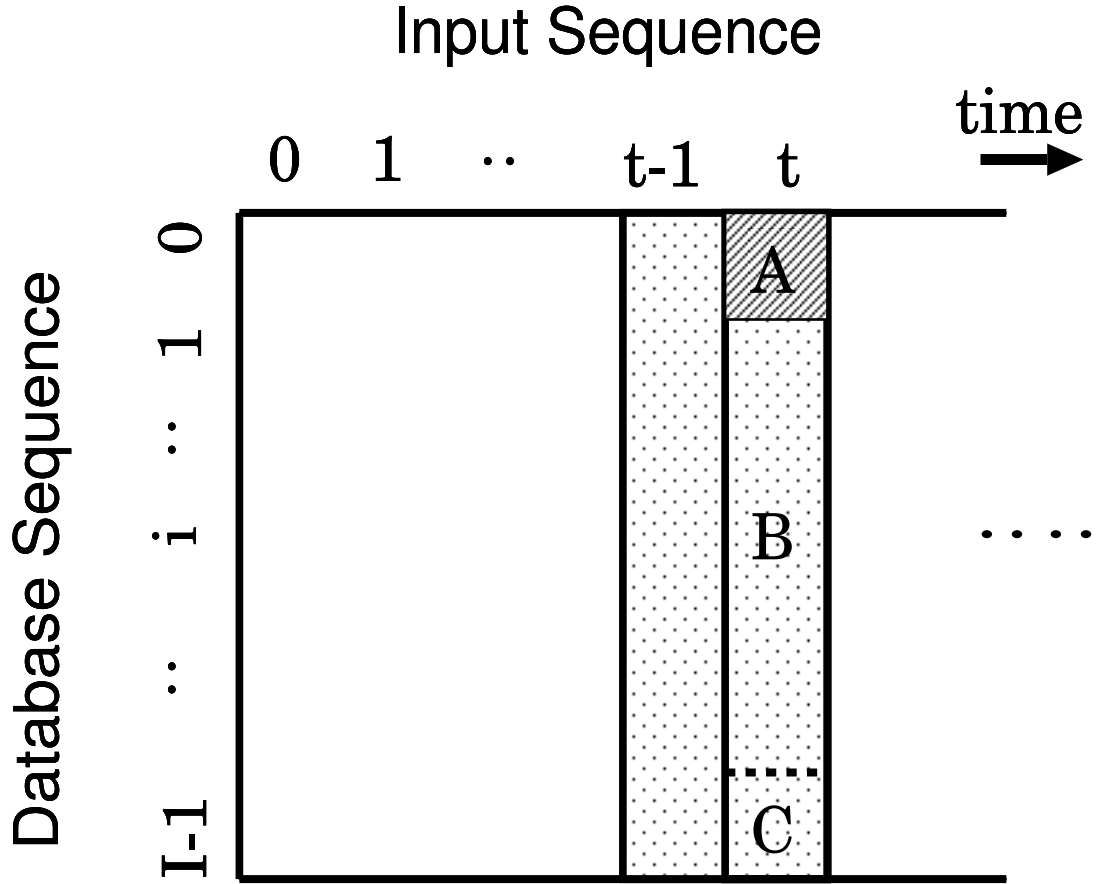


図 3.7: 連続 DP マッチング

要素が取り得る整数値の集合を C_x, C_y としたとき、入力と DB との要素間の距離 d_x, d_y を次のように定義する。

$$\begin{aligned}
 Q_n &= q_{n0}, q_{n1}, q_{n2}, \dots, q_{nt} \in C_n \\
 S_n &= s_{n0}, s_{n1}, s_{n2}, \dots, s_{I-1}, s_{ni} \in C_n \\
 d_n(q_{nt}, s_{ni}) &= |q_{nt} - s_{ni}| \\
 &\quad (\text{ただし } n = x, y)
 \end{aligned}$$

連続 DP マッチングの計算は、図 3.7 のように 2 次元のマッチング行列 $m(i, t)$ を用いて、時刻 t までの入力の重心位置の系列と DB 系列との間の距離 D を求める。はじめに $t = 0$ における距離の計算を行い、次に時刻 t における入力パターンと DB の先頭パターンの距離の計算を行う (Step A)。続いて、時刻 $t-1$ の結果を利用して、その時刻における入力パターンと DB 中のすべての位置のパターンとの距離を計算し (Step B)、最後の行で得られた結果を、時刻 t における入力系列 Q と DB 系列 S との間の距離 D と

する (Step C)。この処理を各 DB 中の系列ごとに行い、時刻 t までの入力系列とすべての DB 系列間の距離を求める。

以下に、各 Step A ～ C における要素の計算を示す。なお、 $t = 0$ おける $m(i, 0)$ の値は、すべて $m(i, 0) = \infty$ ($0 \leq i < I$) とする。

- Step A の計算

$i = 0, t > 0$ において、

$$m(0, t) = d_x(s_{x0}, q_{xt}) + d_y(s_{y0}, q_{yt})$$

- Step B の計算

$0 < i < I, t > 0$ において、

$$m(i, t) = \begin{aligned} & d_x(s_{xi}, s_{xt}) + d_y(s_{yi}, s_{yt}) + \\ & \min(m(i-1, t), \\ & \quad m(i-1, t-1), \\ & \quad m(i, t-1)) \end{aligned}$$

- Step C の計算

$$D(Q, S) = m(I-1, t)$$

以上で定義した連続 DP マッチングの距離 D は、時刻 $t-i \sim t$ の間の利用者の入力系列 Q が S の系列に類似するほど、小さな値を取り、一致する場合、0 となる。

3.5.2 入力文字の決定

DB 中の Num 個の登録データに対する、時刻 t での連続 DP マッチングの距離 $D(Q_t, S_k)$ ($k = 0, 1, \dots, Num-1$) の中から、実際に入力された文字を決定する手順を述べる。

図 3.8 は、本研究で利用した DB 中に登録している英字 'A'～'C' に対する、時刻 t (フレーム) での D の距離の遷移を示しており、時刻 $t_0 \sim t_1$ の間で、'A' の文字を入力している。この図に観察できるように、 $t_0 \sim t_1$ では 'A' に対する距離 D が他の文字の距離よりも小さくなっている。特に入力動作の終る t_1 の直前の時点では急激に小さくなり、しばらく最小値を維持しながら、再び増加していく。

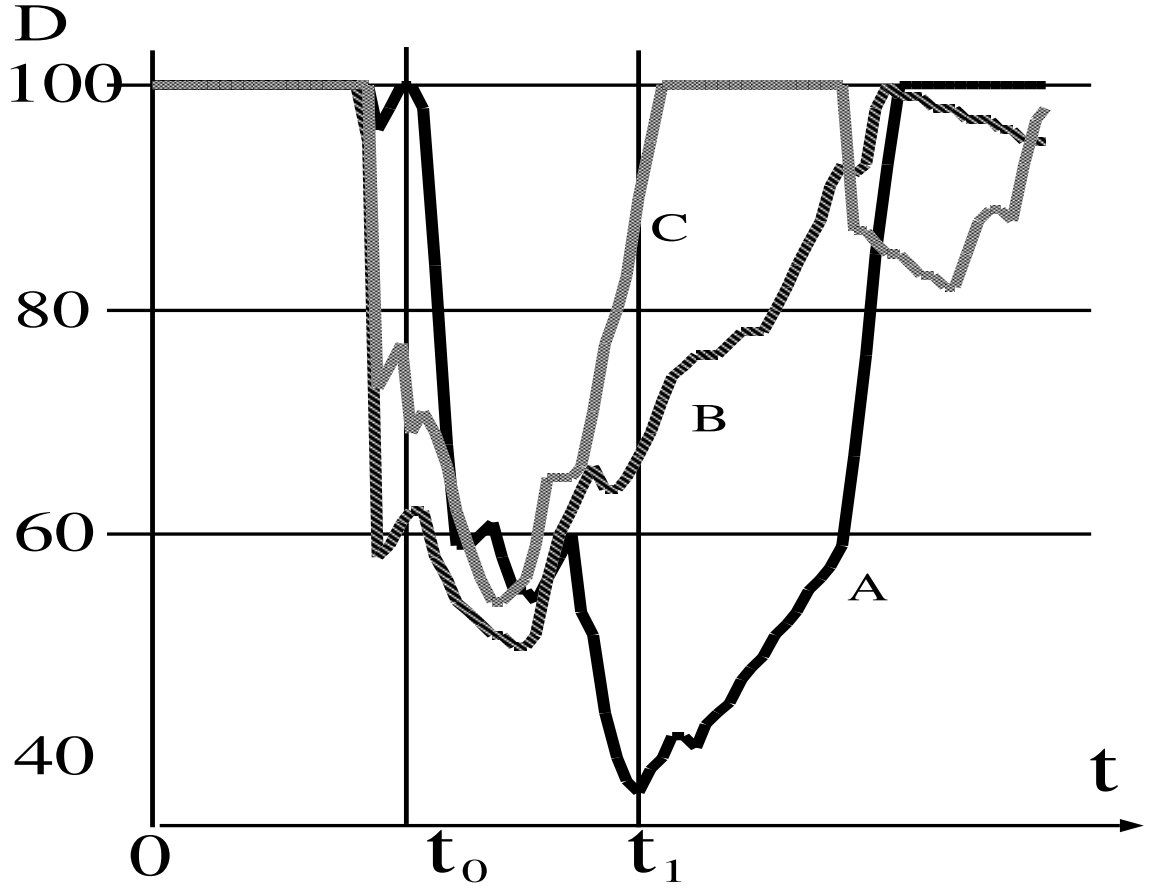


図 3.8: 'A','B','C' に対し、'A' を入力した際の距離 D

本システムでは、この性質を利用し、利用者の入力が終了している（画像フレーム外にある）という前提で、以下の条件をすべて満たす $D(Q_t, S_i)$ ($0 \leq i < Num$) が観測された時刻 t で、 S_i の系列が入力されたと判定する。

• 条件 1

$0 \leq t_x < Time$ において、 S_i が

$$D(Q_{t-t_x}, S_i) = \min\{D(Q_{t-t_x}, S_k) \mid 0 \leq k < Num\}$$

かつ

$$D(Q_{t-t_x}, S_i) \geq D(Q_{t-t_x-1}, S_i) \text{ をすべて満たす。}$$

• 条件 2

$$D(Q_t, S_i) < Distance \quad (Distance \text{ は定数})$$

本研究では各条件で用いる $Time$, $Distance$ を予備実験に基づいて、経験的に決定している ($Time = 5$, $Distance = 50$)。条件 1 は、時刻 t において、最小の D 値を持つパターンを有力候補と見なしており、利用者の文字入力が終わった際に、該当した文字パターンの D がある一定の時間 ($Time$)、最小値を保ちながら増加する傾向にあることを意味している。また、条件 2 では、最小値がある一定の閾値 ($Distance$) より小さい値を取ることを意味している。

3.5.3 空中での手書き文字表記法

上述した手法により、ブロック体の英数字を入力する予備実験を行ったところ、(B, R, K, P), (C, L), (4, 9) などの文字間で誤認識が頻繁に発生した。この中には 'R' と 'K' や '4' と '9' など実際に人間が判別が困難なものもある。

この問題は、通常書かれる文字が、紙の上などで軌跡を残すことが前提となっており、画像だけでは、その軌跡を得ることができないことが原因である。これは、学習アルゴリズムなどで、利用者の癖をより精密に判定させるだけでは、解決することが難しい。

そこで、本研究では、空中での手書き文字認識が根本的に抱えるこの問題を解決するために、空中で手書き文字入力をするための英数字の表記法を提案する。ここで、文字の設計に際し、予備実験から以下の指針を得た。

- 画像フレーム内の手の動きは、必ず一筆書きになるため、一筆書きの文字を設計した方が利用者の誤入力を抑えることができ、望ましい。
- 描かれる文字の軌跡は、輝度差分の重心によって決定されるため、利用者の手の甲部分の輝度差分に影響されやすく、重心の移動は画像フレームの縦方向よりも横方向の方の変化量が多い。このため、文字は横方向に多くの変化量を持つように設計した方が認識率が高い。
- 利用者が文字を書き終えた時点をシステムに認識させ、また、'P' と 'R' などの文字間の誤認識を抑えるために、文字の軌跡の最終部分を画像フレームの外に出るように設計した方が良い。また、その方向によって区別できることが望ましい。

以上の指針に基づいた上で、本研究では、多くの利用者に普及している米国 Palm 社の Graffiti(TM) を参考にし、空中で手書きするための文字を設計した (図 3.9)。この際、で

きるだけ Graffiti に従い、'U' と 'V', 'P' と 'R' などの誤認識が起きやすい文字に関しては、最後にフレーム外に手を出す方向を定めることで、認識率を高める工夫をした。

利用者は、中央のタッチパネルボタンを押した後、各文字の始点まで手を運び、文字を書いた後、画像フレーム外に手を出すことを想定している。なお、図 3.9 の各文字の周りの枠は、画像フレームを表している。

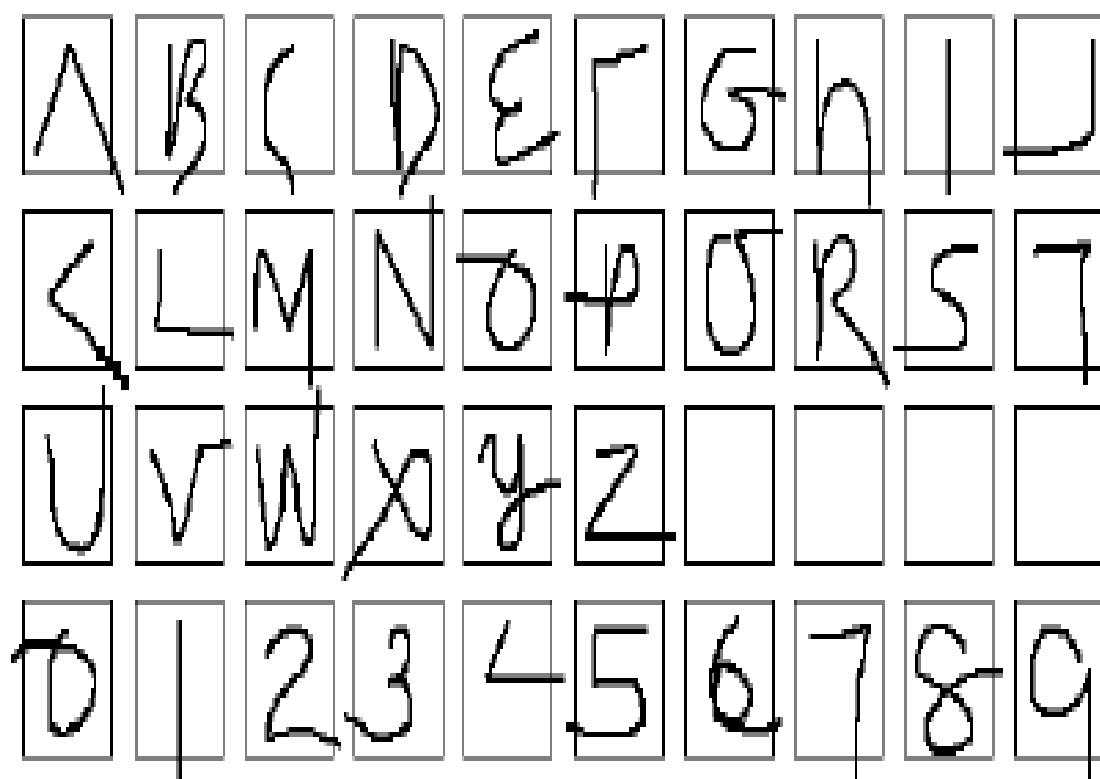


図 3.9: 空中での手書き文字のための英数字の表記法

3.6 むすび

本章では、空書認識を実現するための具体的な一連の手法を述べた。以下に本章のまとめを述べる。

- コビト計算機上で空書認識を実現する手法として (1) タッチパネル型のインターフェースにより、利用者の手の領域を特定し、(2) 輝度差分の重心位置で手の軌跡

を追跡し、(3) 連続 DP マッチングによる文字列を特定するという一連の手法を提案した。

- 空中で描かれる軌跡だけでは判別が難しい文字が存在する問題を解決するために、独自の英数字の表記法を提案した。

本章で述べた一連の手法を、具体的にコビト計算機上に実装する方法は、**第 6 章（マルチモーダルインターフェースの実装手法）**において議論する。また、**第 7 章 認識手法の評価**において、これらの手法を評価する。

第 4 章

歌声認識の手法

本章では、コビト計算機で歌声を認識する手法として、利用者が装着している計算機が無線や有線によって、WWW に接続できることを前提に、WWW 上で多数の曲目の候補から、利用者の意図する曲を検索するシステムについて述べる。

以下では、本システムの概要・仕様を述べ、本システムを実現する上での問題を定義する。次に、その問題に対する解決法を提案し、システムのモデルを定式化する。最後に、曲検索を行うサーバ・クライアントシステムモデルの具体的な処理手順を述べる。

4.1 はじめに

多くの利用者が WWW 上の曲検索サービスを利用できる環境では、利用者の歌声を手掛かりとして、その曲目が分かると便利である。特に、本研究が対象とするような装着型計算機環境においては、街中でふと耳にした曲を、口ずさんで検索できると便利である。

このように歌声を手掛かりとする曲検索システムを構築する上で、ジェスチャ認識でも用いた DP マッチングは、極めて有効であり、多くのシステムが、曲データベースの照合に採用している [Kageyama et al., 1993, Ghias et al., 1995, Sonoda et al., 1998, Smith et al., 1998, Bainbridge 1998, Nishimura et al., 2001]。

しかしながら、DP マッチングは、堅牢な照合手段である一方で、手掛かりとする入力の旋律情報の時系列の長さで照合するデータベースの旋律情報の時系列の長さの積に比例して、検索時間が増加する問題があることが指摘されている [Kageyama et al., 1993, Ghias et al., 1995, Sonoda et al., 2000]。これに対して、旋律情報に対する索引の作成手法も提案され [Sonoda et al., 2000]、旋律情報の索引を使用しない場合と比較して、数

パーセントの検索精度は低下するものの、検索時間を 1,000 分の 1 以下に短縮することが実現された。

このような旋律情報の索引の利用は、高速な検索を必要とする場合には、極めて有用であるが、一方で、検索精度を高く保つことも重要であり、実現されることが望ましい。

そこで、本研究では検索精度の高い手法 [Sonoda et al., 1998] を実装した計算機を、LAN 上で複数台並列化することで、精度高く、高速に検索する、という 2 つの条件を同時に満たす手法を提案する。また、本手法を利用することで、曲検索システムを任意の規模で構築でき、データベースの曲数や利用者の数に応じた、拡張性高いシステムを実現できることを主張する。

4.2 歌声による曲検索

本節では、歌声によって曲を検索するシステムの概要と仕様を述べる。

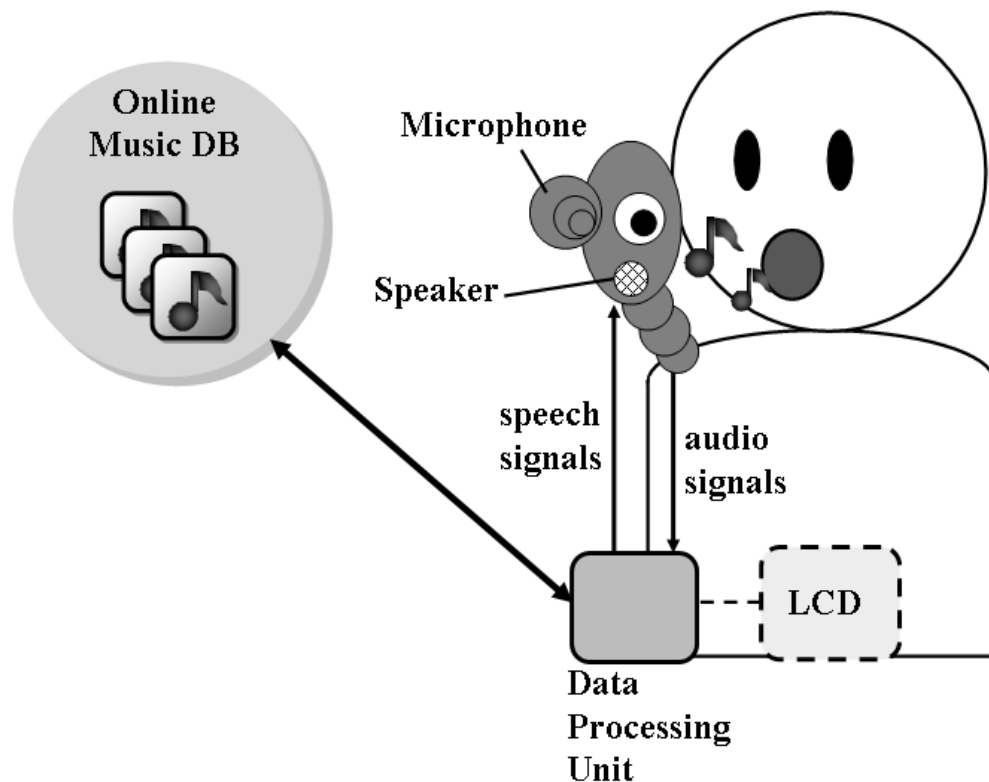


図 4.1: コビット計算機での曲検索

4.2.1 曲検索システムの概要

図 4.1 は、コビット計算機を利用した歌声認識手法の模式図である。利用者は、肩の上のマイクロホンで音声を入力し、録音された音声データを検索クエリデータとして、ネットワーク上の曲検索サービスの結果を得る。ここで、本研究が対象とするような装着型計算機環境においては、街中でふと耳にした曲を、口ずさんで検索することが利用場面として考えられるため、ネットワーク上の曲検索サービスを利用することが適している。

曲検索サーバでは、利用者から送信されたクエリデータと曲 DB の曲データ情報との間で、照合処理を行う。ここで、本研究では、曲データとして、広く普及している SMF(Standard MIDI File) を採用している。照合処理の結果、曲検索サーバは、利用者が入力した検索クエリデータに類似する、データベース中の候補の曲目を順番に並べた曲リストを生成し、利用者側に返送する。

一方、利用者側の計算機クライアントでは、利用者が歌った曲を録音し、それを音声波形解析して、旋律情報を抽出する。このとき、本システムの仕様では、利用者は自由なテンポ、自由なキーの高さで歌ってよく、曲データベースに登録されている曲のどの部分を歌っても検索できるものとする。

次に、この旋律情報に加えて、利用者が音声認識などによって入力した文字列情報を合わせたものを、検索クエリとして検索サーバに送信したする。サーバから返信される曲リストは、計算機のディスプレイに表示されるか、あるいは音声読み上げシステムによって、利用者に伝えられる。

4.2.2 多数の曲からの曲目の特定問題

上述のネットワーク上の曲検索サービスを構築する際に、最も重要となることは、高速かつ高精度の検索処理を、いかにして、多数の曲を保有するデータベースに対して、実現するかということである。また、多数の利用者でサービス共有することも勘案する必要がある。

曲検索に用いる利用者の歌声は、記憶違いや歌唱能力の差異によって、曲データベース中の意図する曲と必ずしも旋律が一致しない。このため、利用者が歌う旋律の最初の音でさえも正しいことは保障されず、通常のテキスト検索のような索引を、旋律情報に対して

作成することは、極めて困難であった。その結果、高速な曲検索を、多数の曲に対して実現することは難しかった。

一方、旋律情報の索引を作成する手法として、従来手法 [Kosugi et al., 2000, Nishimura et al., 2001] では、索引作成にテンポやキーの高さを手掛かりにしたが、利用者がメトロノームに合わせて歌わなければならない制約や、相対的なキーの高さを正しく歌わなければならない制約があった。

これに対し、本システムのように、利用者に自由なテンポや自由なキーの高さを許容する場合は、[Sonoda et al., 2000] のように、旋律情報の短い区間 (3~5 音符) を用いて索引を効果的に作成する手法が利用できる。この手法では、検索時間は 1,000 分の 1 以下に短縮されたが、索引を用いない場合と比較して、検索精度が低下することが指摘されている。しかし、本論文が対象とするような不特定多数の利用者に高速な検索サービスを提供する際には、高い精度を保ったまま、曲検索を行うことが望ましい。

そこで、本論文では、[Sonoda et al., 1998j] で提案されたような極めて高い精度を達成する検索サーバの性質を保ったまま、大規模な曲データベースに対して検索を実現する手法として、曲検索システムを PC によって並列化する手法を提案する。このように並列化された曲検索システムを構築するためには、検索サーバに登録する曲数が徐々に増加していくことを考慮しなければならない、また、利用者の増加によって、システムを柔軟に拡張できることが望ましいが、これまで、このような拡張性を勘案した、曲検索システムの並列化手法は知られていない。

このため、本研究では、これらの拡張性を実現するために、2 種類のデータ中継サーバを定義することで、柔軟にシステムを拡張でき、データベース中の曲数や利用者の増加に対応できる手法を提案する。

4.2.3 拡張性のある曲検索システムの構築手法

図 4.2 は、本研究で提案するネットワーク上の検索システムの構成を示している。図 4.2 において、曲検索サーバシステムの最小単位は、曲検索データベース (DB) を有した 1 つの曲検索サーバ (C) である。この曲検索サーバは、(A),(B) の 2 種類のデータ中継サーバを用いることで、任意の規模に並列化が可能である。2 種類のデータ中継サーバは、利用者が送信した検索クエリデータ (歌声の旋律情報、および曲目やアーティスト名

などの文字列情報で構成される検索キー)を受け取り、そのサーバから接続可能な複数のサーバの 1 つあるいは、すべてに対して、受信したデータをそのまま送信する。

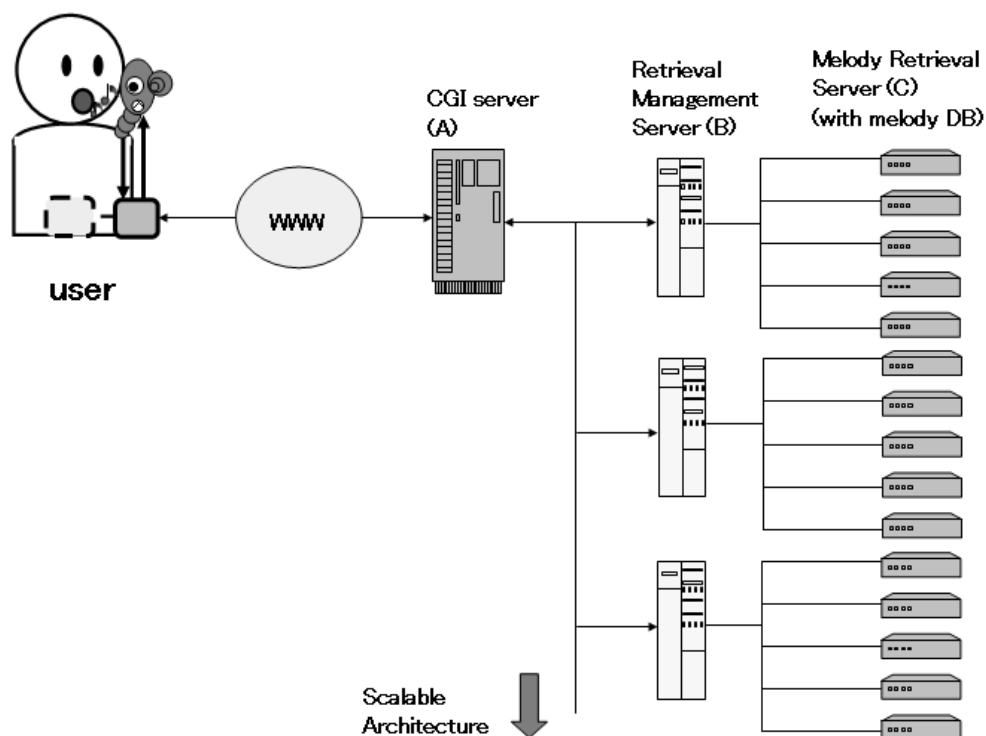


図 4.2: 並列化された計算機上での歌声による曲検索

上述のシステムにおいて、拡張性のある曲検索システムを構築する上で最も重要であるのは、以下の 2 つの種類の中継サーバである。

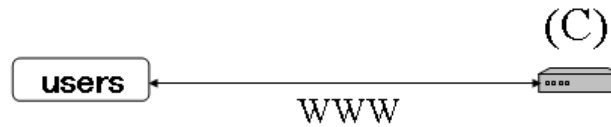
- **CGI サーバ (Common Gateway Interface Server) (A)**

利用者のクエリデータを WWW 上のクライアント端末から受け取り、次の中継サーバ、あるいは曲検索サーバに対して中継送信する。本サーバは、次に送信可能な複数のサーバの中から 1 つをランダムに選択して送信する。

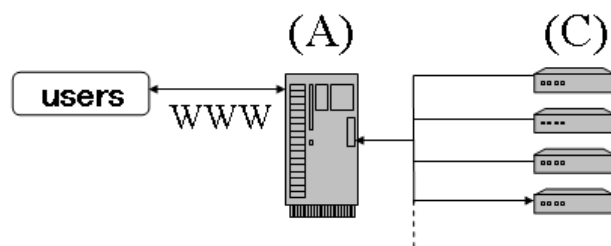
- **検索管理サーバ (Retrieval Management Server) (B)**

クエリデータを利用者のクライアント端末や前の中継サーバから受け取り、曲検索サーバに対して中継送信する機能は (A) の中継サーバと同じであるが、本サーバは、次にクエリデータを送信可能なすべてのサーバに対して、ブロードキャストの形式で送信する。

- (1) a system structure with a small database
and with few user accesses.



- (2) a system structure with a small database
and with a lot of user accesses.



- (3) a system structure with a large database
and with few user accesses.

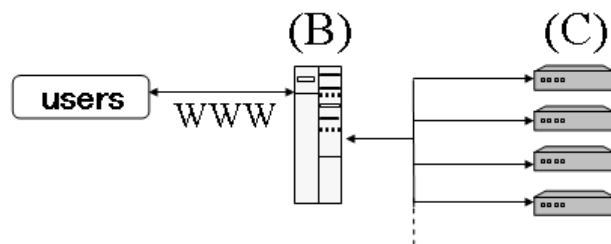


図 4.3: サーバシステムの拡張性

図 4.3 は、これら 2 種類の中継サーバを用いて、曲検索システムの規模を拡大・縮小する様子を示している。また、図 4.2 は、図 4.3 を組み合わせて、大規模な曲検索システムを構築する場合の構成である。

図 4.3-(1)~(3)、および図 4.2 は、各々以下の用途に適している。

- 図 4.3-(1)

曲検索サーバシステムの最小構成で、曲検索サーバ (C) のみから成る。この構成は、データベースの曲数が比較的少なく、かつ利用者のアクセスの頻度が少ないシ

システムに適している。^{*1}

- 図 4.3-(2)

1 つの CGI サーバ (A) と、そのサーバから送信可能な複数の曲検索サーバ (C) による構成。この構成では、各曲検索サーバ (C) に同一の曲 DB を持たせることで、利用者アクセスによる負荷を分散させることができる。この構成は、DB の曲数が比較的少なく、利用者のアクセスの頻度が多い場合に適すシステムである。

- 図 4.3-(3)

1 つの検索管理サーバ (B) と、そのサーバから送信可能な複数の曲検索サーバ (C) による構成。この構成では、各曲検索サーバ (C) に異なる曲 DB を持たせることで、検索処理にかかる負荷を分散させ、検索管理サーバ (B) でその結果を集約して、利用者側に返信することが可能である。DB 中の曲数が多く、かつ利用者のアクセスの頻度が少ない場合に対応できるシステムである。

- 図 4.2

CGI サーバ (A)、検索管理サーバ (B) と 曲検索サーバ (C) を混合した構成で、上述の図 4.3-(2),(3) の特徴を併せ持つことから、大規模 DB に対して、多くの利用者アクセスに対応するシステムである。

次節では、上述の 3 つのサーバにおける具体的な検索処理手順を定式化する。

4.3 検索サーバシステムの定式化

本節では、前節で提案した手法に関する 3 つのサーバ (CGI サーバ、検索管理サーバ、曲検索サーバ) の処理モデルを各々定式化する。

4.3.1 CGI サーバ

以下では、CGI サーバのメッセージ送受信機能を定式化する。

CGI サーバの役割は、WWW 上の利用者クライアントシステムから受信したクエリデータを、次の送信先の曲検索サーバあるいは中継サーバに送信し、その送信先から返信

^{*1} 本研究では、経験的に 1 万曲未満のデータベースを比較的の小規模だとみなしている。また、1 秒間に平均 3 回以下の利用者アクセスがある場合は、アクセスの頻度が少ないシステムだとしている。

される**曲検索結果データ**（曲リスト）を、送信元に対して、再度、中継することである。

この機能を実現するために、CGI サーバはクエリデータを中継する送信先のホストアドレス $HOST$ とポート番号 $PORT$ の対を一つのサーバ情報として並べたサーバリスト $LIST_1$ を保有している。これを以下の式で表わす。

$$LIST_1 = \{hosts_i \mid i = 0, 1, \dots, N - 1\}$$

$$where\ hosts_i = pair(HOST_i, PORT_i).$$

利用者のクエリデータ QM を受信した際には、CGI サーバは $LIST_1$ から、ランダムにホスト情報 $hosts_i$ を選び、 QM を $hosts_i$ で表わされるサーバ $SERVER_i$ に送信する。ここで、 $SERVER_i$ は、曲検索サーバか検索管理サーバのいずれか 1 つである。次に、 $SERVER_i$ からは、曲検索結果データ RM が返信されるので、それを QM の送信元に返信する。

クエリデータの中継時に、あるホスト $SERVER_i$ への接続に失敗した場合は、 $SERVER_{ii}$ ($i \neq ii, s_{ii} \in LIST_1$) に対して接続を試みる。 $LIST_1$ 中のすべてのサーバに対して接続が失敗した場合は、エラーメッセージ $ERROR$ をクエリデータの送信元に返信する。

ここで、 $SERVER_i$ の機能を F_{s_i} で定義すると、 RM と QM は以下の関係になる。

$$RM = F_{s_i}(QM)$$

また、CGI サーバの機能を F_{cgi} とすると、 QM を受け取ったときの返信メッセージは、以下の式で表わせる。

$$F_{cgi}(QM) = \begin{cases} F_{s_n}(QM), s_n \in LIST_1. \\ ERROR : when\ no\ SERVER_n \\ \quad response. \end{cases}$$

この式は、CGI サーバが単にクエリデータを $LIST_1$ 中の 1 つの $SERVER_i$ に中継送信し、 $SERVER_i$ からの曲検索結果データをそのままクエリデータの送信元に返信することを示す。このようなサーバは、WWW 上で複数の利用者が使用可能な検索サービス

を行う上で、極めて重要な役割を果たし、アクセスの負荷分散を行う以外にも、本来の曲検索サーバのホストアドレスやポート番号を、ネットワーク上からは判別できなくするため、保有する曲データベースを保護する際のセキュリティを向上する重要な意味がある。

4.3.2 検索管理サーバ

以下では、検索管理サーバのメッセージ送受信機能を定式化する。

検索管理サーバの役割は、CGI サーバと同様に、受信したクエリデータを、次の送信先である複数の曲検索サーバに送信し、それらの曲検索サーバから返信される、曲検索結果データを集計し、1 つの検索結果としてまとめて、クエリデータの送信元に返信することである。

この機能の実現のために、検索管理サーバは、曲検索サーバリスト $LIST_2$ を保有し、そのリストに記録している、 M 個の曲検索サーバのホスト情報 ($HOST$ と $PORT$ の対) をデータ中継に用いる。これは、以下のような式で示すことができる。

$$LIST_2 = \{mr_j \mid j = 0, 1, \dots, M - 1\}$$

ここで、

$$mr_j = \text{pair}(HOST_j, PORT_j).$$

検索管理サーバが、検索のクエリデータ QM を受信した場合は、 $LIST_2$ に記載されるすべての曲検索サーバに対して、 QM を中継送信する。ここで、各曲検索サーバ MR_j の機能を F_{mr_j} とし、 MR_j が返信する曲検索結果データの曲リストを、 $mLIST_j$ とした場合に、 QM を用いて以下の式で示すことができる。

$$mLIST_j = F_{mr_j}(QM).$$

ここで、各 $mLIST_j$ のリストには、 K_j 個の曲情報が記載されている。曲情報は、曲検索サーバにおいて、検索クエリデータとデータベースの曲データとの照合処理の際に得られるマッチング距離 D と、曲目やアーティスト名などの文字列情報 $dTEXT$ の 2 つが対になった情報である。これを以下の式で表現する。

$$mLIST_j = \{music(j)_k \mid k = 0, 1, \dots, K_j - 1\}$$

ここで、

$$music(j)_k = pair(D(j)_k, dTEXT(j)_k).$$

$music(j)_k$ は、データベースの各曲を示す。 $D(j)_k$ は具体的には、4.4 で述べる DP マッチング距離であり、 $music(j)_k$ と QM の間の距離を示す。すなわち、この値が小さいほど、 $music(j)_k$ は、より QM と類似した箇所を持つ曲であると判定される。

検索管理サーバがクエリデータを中継した M 個の曲検索サーバのうち、 M' 個が曲検索結果データ $mLIST_j$ ($j = 0, 1, \dots, M-1$) を返信してきた場合に、検索管理サーバは、以下のように、これらのデータの和集合を、ひとつの曲検索結果データ $mLIST_{all}$ として生成する。

$$mLIST_{all} = mLIST_0 \cup mLIST_1 \cup \dots \cup mLIST_{M'-1}$$

次に、検索管理サーバは、 $mLIST_{all}$ に含まれるすべての曲情報 $music(j)_k$ を、 $D(j)_k$ の値が小さい順番に整列し、このリスト $mLIST_{sorted}$ を QM の送信元に返信する。

$$mLIST_{sorted} = \{music_h \mid h = 0, 1, \dots, K_{sorted} - 1\}$$

ここで

$$\begin{aligned} K_{sorted} &= \sum_{j=0}^{M'-1} K_j, \\ music_h &= pair(D_h, dTEXT_h) \in mLIST_{all}, \\ D_h &\leq D_{h+1}. \end{aligned}$$

検索管理サーバに、ひとつのサーバも検索結果データを返信しない場合は、検索管理サーバは、エラーメッセージ $ERROR$ を QM の送信元に返信する。

ここで、検索管理サーバの機能を F_{rms} とすると、 QM を受け取ったときの返信メッセージは、以下の式で表わせる。

$$F_{rms}(QM) = \begin{cases} mLIST_{sorted} & \text{when } M' > 0 \\ ERROR & \text{when } M' = 0. \end{cases}$$

以上で述べたように、検索管理サーバは複数の曲検索サーバの検索結果を管理する。この機能は、高速な検索を実現するとともに、システム全体の拡張性を実現するという重要な意味がある。

4.3.3 曲検索サーバ

以下では、曲検索サーバのメッセージ送受信機能を定式化する。

曲検索サーバは、検索のクエリデータ QM を受信すると、曲検索結果データ $mLIST$ を返信する。この機能を F_{mrs} で表わすと、

$$mLIST = F_{mrs}(QM)$$

となる。ここで、曲検索サーバの保有する曲データベースを DB とすると、 DB は、各曲の曲目やアーティスト名を含む文字列情報 $dTEXT$ と、曲の旋律情報 $dMELODY$ を対にした情報で構成される。ここで、 DB に含まれる曲数が K とすると、以下の式で表現できる。

$$DB = \{data_k \mid k = 0, 1, \dots, K - 1\}$$

ここで、

$$data_k = pair(dTEXT_k, dMELODY_k).$$

ここで、 $dMELODY$ には、曲の旋律情報として、旋律を構成する各隣接音符間の相対的音高 (relative-pitch value) と相対的音長情報 (relative-span value) を、SMF から計算して格納している。これらの相対的な値が、 QM との照合処理に用いられる。

一方、 QM に含まれる文字列情報と旋律情報を、各々 $qTEXT$ 、 $qMELODY$ と表現する。 QM には、これら 2 種類の情報のうち、いずれか一方しか含まれていなくても良い。このとき、 QM は以下のように表現できる ($NULL$ は $qTEXT$ や $qMELODY$ の情報がないことを意味する)。

$$QM = \begin{cases} pair(qTEXT, qMELODY) \\ pair(NULL, qMELODY) \\ pair(qTEXT, NULL) \end{cases}$$

曲検索サーバにおける照合処理では、 $qTEXT$ が $NULL$ でない場合は、候補の曲 $CAND$ を絞り込む。このとき、 $qTEXT$ は、 $dTEXT_k$ と部分一致することを前提としている。 $CAND$ の集合は以下ようになる。

$$CAND = \{data_k \in DB\}$$

ここで

$$qTEXT(\neq NULL) \in dTEXT_k$$

$qTEXT$ が $NULL$ の場合は次のようになる。

$$CAND = DB$$

ここで、

$$qTEXT = NULL$$

以上の文字列情報の照合処理の後、曲検索サーバは旋律情報の照合処理として、DP マッチング処理 $DpMatching$ を行う。この結果、 $qMELODY$ と各 $dMELODY_k$ の系列間の DP マッチングの距離 D_k が以下のように求められる。

$$D_k = DpMatching(qMELODY, dMELODY_k).$$

ここで、

$$0 \leq k < K, \quad data_k \in CAND.$$

この結果、 $mLIST$ は、最終的には以下のように定式化される。

$$mLIST = \{music_k \mid 0 \leq k < K, \\ data_k \in CAND\}$$

ここで、

$$music_k = pair(D_k, dTEXT_k), \quad D_k \leq D_{k+1}.$$

また、 $mLIST$ に含まれる曲情報 $music_k$ は、 D_k の小さい順に整列された後に、返信される。

4.4 曲検索サーバの旋律情報の照合処理

前節で述べた一連の検索サーバシステムの中で、最も重要であるのは、曲検索サーバにおける旋律情報の照合処理である。本節では、この照合処理の具体的な処理手順として、DP マッチング処理について述べ、DP マッチングで扱う大まかな旋律情報を決定する手法として、「動的閾値決定法」と「Coarse-to-Fine マッチング」を述べる。

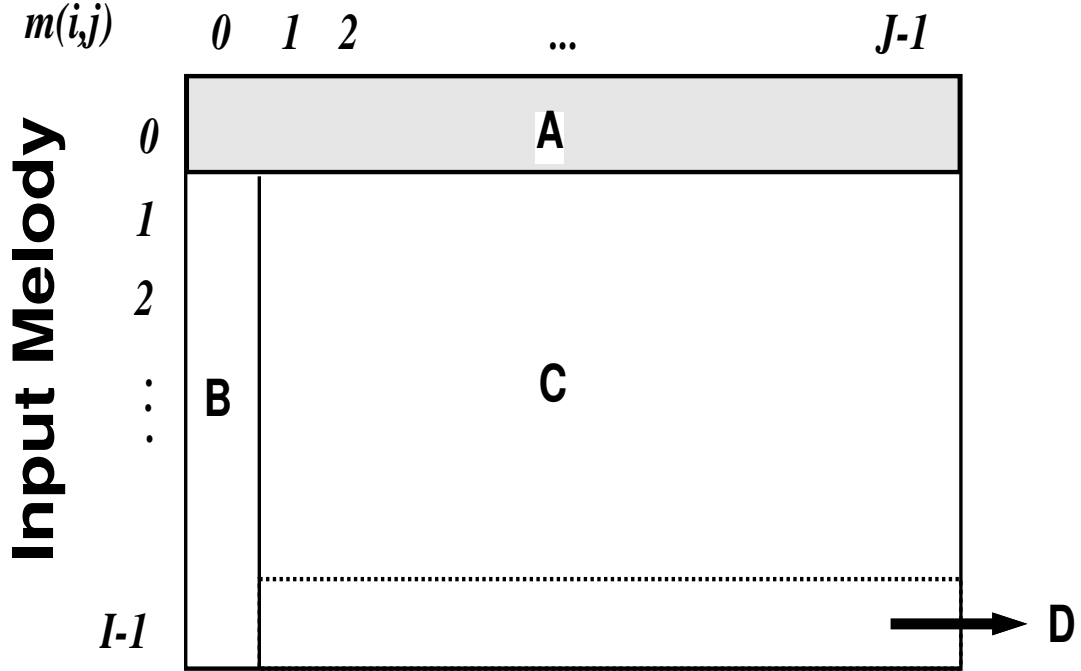
4.4.1 DP マッチング処理

DP マッチングでは、入力と DB 中の各曲の旋律の系列間の距離を求める。利用者の入力する旋律の曖昧さを許容するために、マッチングには、隣接する各音符の相対音高差・相対音長比を、それぞれ閾値によって粗いカテゴリに分割して、粗い精度に変換した相対値を用い、比較する粗い相対値同士の差を DP マッチングにおけるペナルティとし、その合計値を系列間の距離とする。具体的には次のように行なう。

入力の歌声の旋律を粗い精度に変換した、長さ I の相対音高差・相対音長比の系列を $QR_p \cdot QR_s$ とする。また、DB 中のある曲が持つ旋律を粗い精度に変換した、長さ J の相対音高差・相対音長比の系列を $SR_p \cdot SR_s$ とする。入力の系列 QR_x と DB 中の曲の持つ系列 SR_x についての距離 $D_x(QR_x, SR_x)$ を DP マッチングによって求める（ここで、 $x = p, s$ とする）。本システムでは SR_x のすべての箇所を開始点として QR_x の先頭からのマッチングを行なうことが可能となっている。

粗い精度の相対音高差・相対音長比が取り得る整数値の集合を、それぞれ $C_p \cdot C_s$ とし、相対値間のペナルティ d_x を次のように定義する。

Database Song

図 4.4: DP マッチング行列 $m(i, j)$

$$\begin{aligned}
 QR_x &= q_0, q_1, q_2, \dots, q_{I-1}, \quad q_i \in C_x \\
 SR_x &= s_0, s_1, s_2, \dots, s_{J-1}, \quad s_j \in C_x \\
 d_x(a, b) &= |a - b| \\
 &: a \text{ と } b \text{ の間のペナルティ, } a, b \in C_x \\
 &(\text{ただし } x = p, s)
 \end{aligned}$$

DP マッチングの計算は 2 次元のマッチング行列 $m(i, j)$ を用いて行なう。各要素の計算は図 4.4 のように、はじめに 1 行目の計算 (Step A)、次に 1 列目の計算を行なう (Step B)。それからすべての要素について計算を行ない (Step C)、最後に一番下の行の最小値を用いて距離 D_x を求める (Step D)。以上の処理を DB 中のすべての曲について行ない、入力系列との距離 D_x を求める。

以下に各 Step A~D の計算手順を示す。

- Step A の計算

$0 \leq j < J$ において

$$m(0, j) = d_x(q_0, s_j)$$

- Step B の計算

$0 < i < I$ において

$$m(i, 0) = \infty$$

- Step C の計算

$1 \leq i < I, 1 \leq j < J$ において

$$m(i, j) = d_x(q_i, s_j) + \min \begin{pmatrix} m(i-1, j), \\ m(i-1, j-1), \\ m(i, j-1) \end{pmatrix}$$

- Step D の計算

$$D_x(QR_x, SR_x) = \min\{m(I-1, j) \mid 1 \leq j < J\}$$

4.5 動的閾値決定法

DP マッチングの精度は、照合に用いる、粗い相対音高差・相対音長比の値によって定まる。そこで、本研究では、最適な粗い相対音高差・相対音長比の値を得るために、動的閾値決定法 [Sonoda et al., 1998j] を用いる。

ここで、まず、利用者の入力する旋律の曖昧さを、吸収するための粗い相対値のカテゴリ数を 3 つとして考える。すなわち、音符の粗い相対音高差の系列は、前音から「上がった (U)」「同じ (E)」「下がった (D)」という系列に、粗い相対音長比の系列は、前音から「長くなった (L)」「同じ (E)」「短くなった (S)」という系列に変換する。

このように、閾値を用いて、入力の旋律と DB 中の曲の旋律を、3 つの粗い相対音高差・相対音長比の値の系列に変換する場合には、DB 中の曲に出現する各カテゴリの値の出現確率は等しい方が良い。例えば、相対音高差において、DB 全体で U 、 E 、 D の 3 つのカテゴリの中で、 E が他よりも多く出現するような閾値の設定をしてしまうと、入力旋律の粗い相対音高差の系列に、 E が多く含まれる場合に、DB 中の各曲との距離に差異を付けることが難しく、正答の曲の候補の絞り込みが、まったく成功しない。これに対して、動的閾値決定法は、図 4.5、4.6 に示すような分布図に従って、最適な閾値を決定し、粗い相対値に変換する手法である。

図 4.5、4.6 はそれぞれ、歌声によって作成した DB の曲（200 曲）中出现するすべての相対音高差・相対音長比のヒストグラムである。相対音高は半音の差が 100 となるように正規化しており、相対音長は比をパーセンテージで表現している。

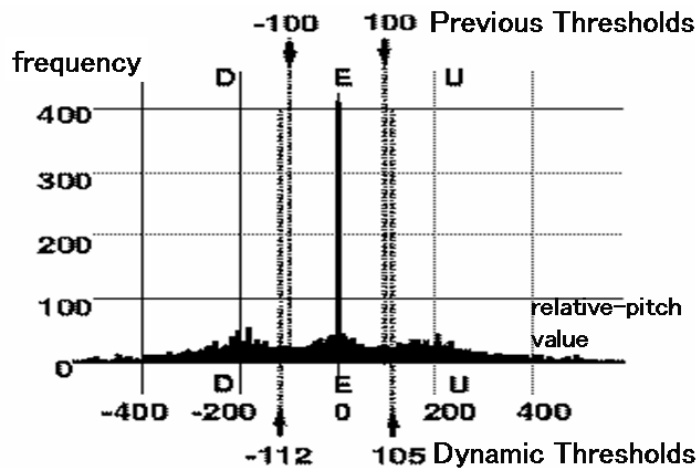


図 4.5: 相対音高差の分布と閾値

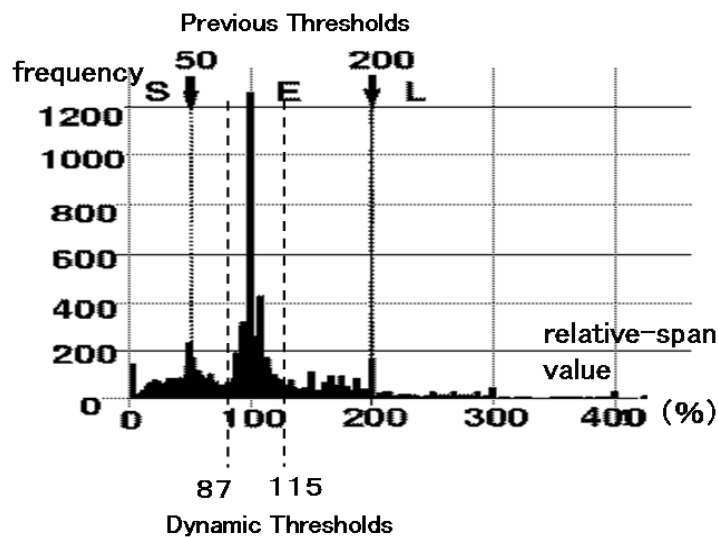


図 4.6: 相対音長比の分布と閾値

ここで、動的閾値決定法では、閾値によって分割されるカテゴリ内の合計度数が均等分されるように、その値を決める（図 4.5、4.6 の下部に示す）。この手法を用いることで、

各々のヒストグラム中で、3つのカテゴリの出現等確率は均等に近づき、対象とする DB 中の正答の曲を絞り込む上で、最適な閾値を決定することができる。また、DB 中の曲が変更されるたびに動的に閾値を変更することができるため、常に最適な閾値を保つことが可能である。

また、図 4.5、4.6 の各上部に示す、従来のアドホックな閾値??の位置との比較から予測できるように、本手法では、特に図 4.6 の相対音長比に対しては、アドホックな閾値との差異が大きく、検索精度が高く向上することが知られている [Sonoda et al., 1998j]。

以下に、一般的な場合の動的閾値の決定手法と、その閾値を用いた検索キーの生成手順を述べる。

動的閾値の決定法

動的閾値 DT の決定法は以下のようなになる。

1. DB 内の全曲中の音高、音長の相対値で図 4.5、図 4.6 のようなヒストグラムを作成する。
2. ヒストグラムの総度数を Sum とし、カテゴリ数を $CategoryNum$ としたとき、1つのカテゴリ内の合計度数の期待値を $M = Sum/CategoryNum$ とする。
3. 各カテゴリの合計度数が均等に M に近くなるようにヒストグラムを分割し、その各境界線を閾値 $DT(j)$ ($0 \leq j \leq CategoryNum - 2$) とする。

このようにして決定した動的閾値を利用することで、歌声の入力と DB 中の曲すべてに出現する音高・音長の相対値の粗い相対値が得られる。

動的閾値を用いた検索キーの生成

相対音高差の系列 Q_p 、相対音長比の系列 Q_s に対する動的閾値 DT_p , DT_s を利用し、それぞれの粗い相対値情報の系列 R_p , R_s を求める。

$DT_n(j)$ ($0 \leq j \leq CategoryNum - 2$) ($n = p$ または $n = s$) によって変換される $CategoryNum$ 段階の粗い相対値 c_k ($0 \leq k < CategoryNum$, $c_k < c_{k+1}$) の集合を C

とする。長さ m の入力系列 Q_n を

$$Q_n = q_0 \ q_1 \ q_2 \ \dots \ q_{m-1}$$

で定義するとき、 Q_n に出現する各音符の相対値 $q_i (0 \leq i < m)$ を $CategoryNum$ 段階の粗い相対値情報 $r_i (\in C)$ に変換する際の条件は、以下のようになる。

$$r_i = \begin{cases} c_0 & (q_i < DT_n(0)) \\ c_k & (DT_n(k-1) \leq q_i < DT_n(k)), \quad 1 \leq k < CategoryNum - 2 \\ c_{n-1} & (DT_n(CategoryNum - 2) \leq q_i) \end{cases}$$

この変換により、 Q_n は粗い相対値情報の系列

$$R_n = r_0 \ r_1 \ r_2 \ \dots \ r_{m-1} \quad (r_i \in C)$$

に変換される。マッチング時には、DB 中に出現する音高・音長に対しても同様の処理を行えば良い。

4.5.1 Coarse-to-Fine マッチング法

相対値を粗くカテゴリ分けして検索時のマッチングを行なうと、曲 DB 中の多くの曲で同様の系列パターンが出現する可能性が高くなり、正答の絞り込みは一般に難しくなる (図 4.7)。この問題を解決するために、より細かい分類のカテゴリを用いることも考えられる。しかし、最初から細かい精度の相対値を用いると、利用者の歌声の変動や誤差による影響を吸収できず、逆に正答率が落ちてしまう。そこで、このトレードオフを勘案しながら、精度の高い検索を行なう手法が必要となる。

そこで、本研究では、検索時に情報を粗い精度に変換ためのカテゴリ数を徐々に向上させながらマッチングを行なう Coarse-to-Fine マッチング法を用いる [Sonoda et al., 1998j]。本手法では、まず、大まかなカテゴリ数 (例えば 3 段階) の相対音高差・相対音長比を利用し、DP マッチングを行なう。その結果、DB 中の曲で入力との距離が近いと解釈された曲の上位に着目する。ここで、その精度で絞り込みが十分かどうかの判定を行ない、絞り込めていなかった場合にのみ、より細かい精度 (例えば 9 段階) の相対値を利用したマッチングを上位の曲に対して再度行なう。そして、カテゴリを細かくしながら絞り込みが十分にできるまでこれを繰り返す (図 4.8)。

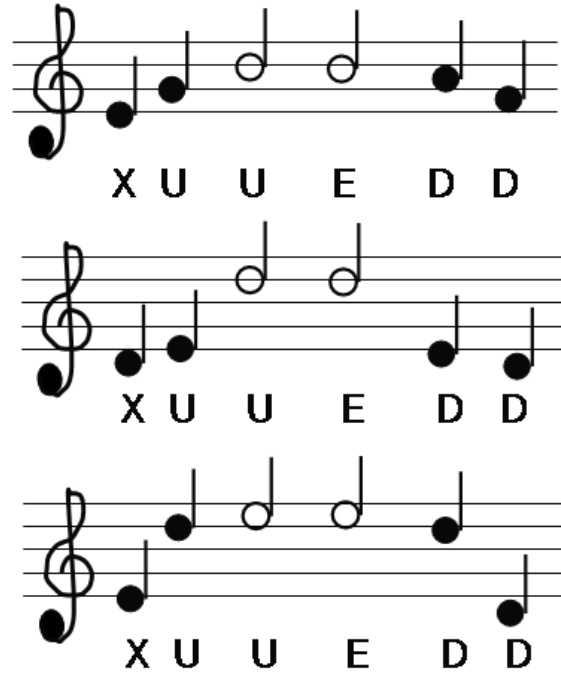


図 4.7: 粗いカテゴリを用いたマッチングに見られる問題

たとえ粗い精度のマッチングで正答を絞り込めなくても、正答の曲は上位に含まれる性質があるため、このように上位の曲を対象に繰り返しより精密なマッチングを行うことで、正答の曲が適切に絞り込める。

また、実際には、十分な分割数の閾値を用意することで、複数の閾値を持つことを回避できる。例えば、 $3^3 = 27$ 段階の閾値を予め用意することで、3 段階、9 段階、27 段階の閾値を別に用意する必要がない。

以下に、Coarse-to-Fine マッチング法の実施方法を述べる。

Coarse-to-Fine マッチング法

歌声入力 of 相対音高差・相対音長比の系列 Q_p, Q_s と DB 中の曲の相対音高差・相対音長比の系列 S_p, S_s との間で、DP マッチングを行なう際、両者のそれぞれの系列を G_i 段階の粗い精度の相対値 $QR_p(i), QR_s(i)$ と $SR_p(i), SR_s(i)$ に変換することを考える（ただし、 $i = 0, 1, \dots, m$, $G_i < G_{i+1}$ とする）。この変換には、動的閾値 $CTF_p(i), CTF_s(i)$

5. 以下、 $i + 1$ を新たな i とみなし、絞り込みが完了するか、または、 $i = m - 1$ となるまで 3. 4. の処理を繰り返す。

4.6 利用者側のクライアントシステム

本節では、利用者側のコビット計算機で実装する、曲検索クライアントシステムの処理について、処理手順に沿って述べる。

4.6.1 音声の録音

まず、利用者はマイクロホンから歌声を入力する。このとき、利用者の歌い方は「タタタ」や「チャチャチャ」など、1 つの音符に、無声破裂音の子音と母音の 'a' の対からなる 1 モーラを対応させたものとし、キーの高さやテンポには特に制約を与えない。また、DB 中の曲のどの部分を歌っても良いものとする [Sonoda et al., 1998]。なお、録音時には利用者の歌声以外は入力されないことを前提とする。本研究では、A/D 変換には、8 bits, モノラルでサンプリング周波数は 8,000 Hz を用いている (図 4.9-1.)。

4.6.2 音符を区切るための音の長さの検出

利用者の歌声に出現する音符を検出するために、まず、音の長さを決定する。本研究では、音符の区切りとして、音符の発音開始位置を、音声波形領域の振幅の強さで決定する (図 4.9-2.)。

ここで、母音の 'a' は比較的強い声帯の振動を要する特徴を利用するため、音声処理に用いるフレーム中の振幅の最大値の絶対値に注目し、それが一定値 *VOICED* を上回る場合は、母音の候補とし、それ以外の場合は、音符の区切れと見なす。本研究では $VOICED = 5$ としている。

4.6.3 各音声フレームのピッチ検出処理

次に、クライアントシステムは有声音と判定したフレームに対して、512 点のハミング窓 (hamming window) を音声波形に施した後に、FFT (Fast Fourier Transforms) を施し、各音声フレーム (16msec) におけるピッチの検出を行う (図 4.9-3.)。

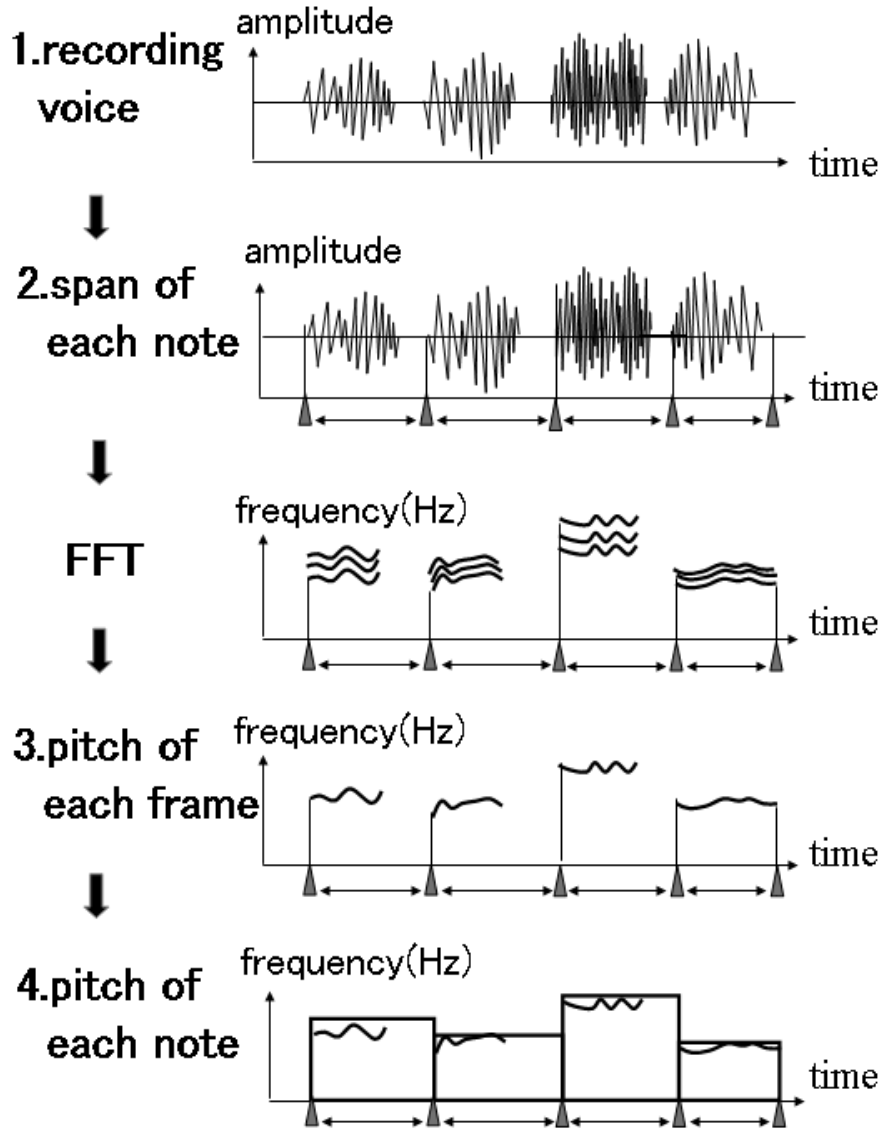


図 4.9: 各音符の音高と音長の推定

この処理において、ピッチの検出には、周波数領域に出現する倍音構造に着目する。すなわち、ある周波数領域の周波数軸上の p ($0 < p \leq 256$) の位置にピークが存在すると仮定し、また、それが $p/NPEAK$ の位置にあるピークの $NPEAK$ 倍音であると仮定したときに、 $p/NPEAK$ と p の位置の間に出現する、 $p/NPEAK$ の整数倍の位置 ($p/NPEAK, (p*2)/NPEAK, \dots, p$ の) のピーク合計値 SUM_p を求める。この処理を、すべての位置 p に対して行った結果、 SUM_p の値が最大となった p から得られる

$p/NPEAK$ を、このフレームのピッチとする。本研究では $NPEAK = 5$ としている。

4.6.4 各音符のピッチ特定処理

次に、以上の処理で求めた音符の区切りと、各有声音フレームのピッチ情報を用いて、旋律情報の最小単位となる音符の音高を求める (図 4.9-4.)。連続する有声音のフレームの区間を 1 つの音符の区間とし、その区間のフレームにおいて検出されたピッチの最大値を、その音符の音高と定める。

4.6.5 相対音符への変換処理

利用者に、自由なキーとテンポで歌うことを許容するために、クライアントシステムは、上述の処理で得られた音高・音長の時系列において、各音符の前の音からの相対的な値に変換する。このとき、音の高さは、2 音符間での半音の差を 100 という値に正規化し、音の長さは、直前の音符との長さの比をパーセンテージで表わす。

4.6.6 検索クエリデータの送信処理

以上の処理の結果、利用者の音声は、膨大な音響信号から、データ量が極めて小さな音符データに変換されるため、ネットワークにかかる負荷が小さく望ましい。クライアントシステムは、検索クエリデータとして、旋律情報に加え、文字列情報（アーティスト名や曲目など）を検索サーバに送信することができる。

4.6.7 検索結果データの受信処理

曲検索サーバから得られる検索結果は、複数の曲情報が含まれる曲リストである。ここで、1 つの曲情報は、曲目とアーティスト名を対としたものであり、利用者の入力した旋律により類似するものから順に整列されている。

本研究のコビト計算機では、検索結果の曲リストの内容を、独自に開発したプログラムに、既存の発話エンジン（米国マイクロソフト社 SAPI5.0）を利用し、利用者に音声で知らせることができる。

4.7 むすび

本章では、歌声認識を実現するための一連の手法を述べた。以下に本章のまとめを述べる。

- コビト計算機上で、歌声認識を実現するためには、ネットワーク上の曲データベースにアクセスし、検索処理によって、意図する曲目を特定するモデルが適している。
- 多数の曲を含む、曲データベースを、複数の利用者によってネットワーク上で共有し、歌声を手掛かりとして、高速かつ精度良く検索するために、曲検索サーバを並列化する手法を提案し、モデルの定式化を行った。
- 本手法では、2 種類のデータ中継サーバを新たに定義することで、任意の規模に曲検索システムを拡張でき、曲数や利用者数に応じた規模のシステムを構築できる。

本章で提案した認識手法を具体的にコビト計算機上に実装する手法は、**第 6 章（マルチモーダルインターフェースの実装手法）**で述べる。また、本手法の有効性は、**第 7 章（認識手法の評価）**で述べる。

第 5 章

コビット計算機システムのモデル

本章では、第 3 章、第 4 章で述べた各認識手法を統合された 1 つのシステムで利用するためのコビット計算機システムのモデルを提案する。

5.1 はじめに

人間同士が音声やジェスチャなどの視聴覚信号を用いてコミュニケーションを行う際には、その信号の意味は、時間や場所、コミュニケーションの相手など、様々な背景要因によって、異なってくる。

同様に、本研究でジェスチャ認識、歌声認識による視聴覚インターフェースを計算機の統一環境で、計算機インターフェースとして用いる際には、同一の視聴覚信号に対しても、計算機の利用場面や信号の組み合わせ、アプリケーションに応じて、異なる適切な動作ができることが望ましい。

このような動作を行う計算機の環境を設計するためには、計算機の動作仕様を各利用場面に対して、柔軟に変更できる仕組みが必要となる。そこで、本研究では計算機の動作制御を、静的なプログラムとして実装するのではなく、各利用場面における動作を記述した「**動作仕様書 (Action Specification Document)**」によって動的に定義できるモデルを設計する。

本モデルでは、利用者の発する視聴覚信号を解析した結果を、計算機上で「**信号イベント (Signal-Event)**」として解釈し、動作仕様書内の記述によって、各信号イベントに対する計算機の動作を決定する。動作仕様書は、利用場面ごとに定義し、各動作仕様書には、各信号イベントに対して、計算機がどのように動作すべきかの記述を定義できるもの

とする。また、利用場面の変更示す信号イベントの発生時には、次の利用場面に対応する動作仕様書を動的に呼び出す記述を定義できるものとする。

ここで、本システムのアプリケーションソフトウェアでは、計算機の複雑な動作制御をあらかじめ記述できる動作仕様書が必要になると考えられるが、本手法では、上述した 2 種類の記述よりも高度な処理はアプリケーションソフトウェアごとに、システム設計者が、実装するモデルとする。この結果、各視聴覚信号に対して、計算機の動作を人間が意図的に制御したい場合は、「利用場面」と「利用場面における信号イベント」の 2 種類の情報のみを管理すればよく、計算機の動作を非常に単純に制御できる。

以下では、利用者と計算機間の通信インターフェースとして、視聴覚信号を用いる手法を用いるコビット計算機のシステムモデルを述べる。

5.2 コビット計算機のシステムモデル

図 5.1 は、本研究で提案するコビット計算機のシステムモデルを示したものである。この図のように、コビット計算機のシステムは以下のような複数の独立した装置、データ処理プログラムから構成される。

- (1) 感覚センサー (Sensors)

計算機に接続された感覚センサーのビデオカメラとマイクロホンである。ここで取得される視聴覚信号は、データ処理装置 (Data Processing Unit) に送信される。

- (2) 視聴覚信号処理プログラム (Audio-Visual Signal Processing Programs)

計算機のデータ処理装置において、ビデオカメラとマイクロホンから得られる視聴覚信号を取得し、それらを次の「(3) 信号イベント生成プログラム」の部分に、通信ポート経由で送信するプログラム群である。

- (3) 信号イベント生成プログラム (Signal-Event Generating Programs)

視聴覚信号を解析して、その結果となる「信号イベント」を生成するプログラム群である。第 3 章、第 4 章で述べた空書認識、歌声認識の各手法は、このプログラ

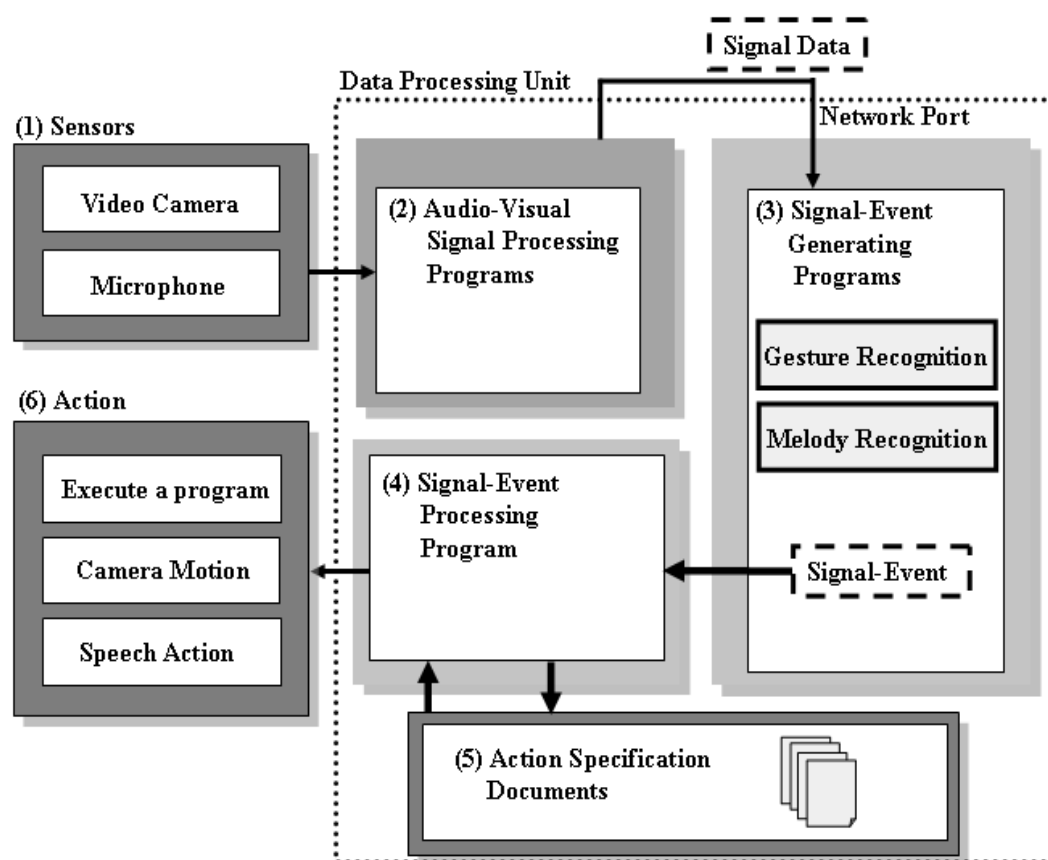


図 5.1: コビト計算機システムのモデル

ム群の 1 つとして実装できる（図中 Gesture Recognition と Melody Recognition で表記）。

- (4) 信号イベント処理プログラム (Signal-Event Processing Program)

信号イベントを解析し、現在の利用場面における計算機動作を決定するプログラムある。この動作決定は、動作仕様書に記述されている「動作命令」によって行う。

- (5) 動作仕様書 (Action Specification Documents)

各利用場面において、各信号イベントに対して、どのように計算機が動作するべきかを定義した、仕様書データである。1 つの利用場面に対して、1 つの仕様書が対応しており、各利用場面において、信号イベントが発生した場合、対応する仕様書に記述されている動作を、計算機が順次実行する。

● (6) 計算機動作 (Action)

動作命令に従って発生する、動作を示す。この動作として、任意のプログラムを実行する動作やビデオカメラを動かす動作、言葉を発話する動作などが挙げられ、このインターフェースを使って、計算機が視聴覚情報を発する。

以下に続く各節では、上述した各機能のうち、データ処理装置内の処理モデルについて、次の順に述べる。

- 1. 視聴覚信号処理 (5.3)
- 2. 信号イベント生成 (5.4)
- 3. 動作仕様書 (5.5)
- 4. 信号イベント処理 (5.6)

5.3 視聴覚信号処理

本研究では、視聴覚情報の信号を、ネットワーク通信によって集約するモデルを採用する。このモデルでは、視聴覚信号処理のプログラムにおいて取得した信号を、次の「信号イベント生成」のプログラムに対して、通信ポート経由で送信する。

「信号イベント生成」のプログラムに対して、送信される視聴覚情報の信号の単位は、動画像の 1 フレーム (24bit カラー、320*240) と音声波形 1 フレーム (サンプリング周波数 8kHz における 512 点、フレームシフト幅 16msec) としており、視覚信号と聴覚信号は、個々に、1 フレームサイズのデータが感覚センサーから取得できた時点で送信される。また、通信プロトコルには、表 5.1 で示すような極めて単純な形式を用いる。

表 5.1: 視聴覚情報のネットワーク通信データ形式

通信情報	意味
size	4byte のデータサイズを示す 0 以上の整数。
data	size で示されるバイト数分の視聴覚信号データ

ここで、表 5.1 のような通信データ形式では、視聴覚信号データの種類を区別することが不可能であるため、視聴覚情報の信号の区別は、特定の通信ポートによって区別する。

各通信ポートで受信した信号は、空書認識、歌声認識各々の解析処理を行うプログラムに渡される。

5.4 信号イベント生成

「信号イベント」とは、具体的には、動画像処理、音響信号処理の結果得られる、信号の意味解釈を示す「文字列データ」として定義する。

信号イベントを示す文字列は、利用場面に応じて、システムの設計者がプログラム内で自由に定義できる仕様とする。一方、システムの利用者は、各利用場面における信号イベントを、以下で述べる動作仕様書内に定義することで、任意に、その信号イベントを取得できる。

例えば、動画像処理のプログラムを作成したシステム設計者が、画像全体が「暗く」なったことを検出した場合に、“DARK” という文字列で表現された信号イベントを発行することを定義したと仮定する。ここで、システム利用者が、コビット計算機を利用している最中に、この“DARK” という信号イベントが発生した場合、システムの利用者が“DARK” という信号イベントを、予めその時点の動作仕様書内に定義している場合は、図 5.1 の (4) のプログラムにおいて、その動作仕様書に記述されている動作命令が順に実行される。また、動作仕様書に、対応する信号イベントが定義されていない場合は、この信号イベントは無視される。

以上の仕様に従ったプログラムを、図 5.1 の (3) において、構成する手法として、1 つの信号イベント監視プログラムと、その監視ルールに従って信号イベントを発する信号処理プログラム群によって具現化することが考えられる (図 5.2)。

本論文で提案した空書認識と歌声認識も、この信号処理プログラムとして、他の動画像処理プログラム、音響信号処理プログラム等と同一の環境で利用できる。

この図 5.2 内における空書認識、歌声認識各々の信号イベント生成手法については、第 6 章で具体的に述べる。

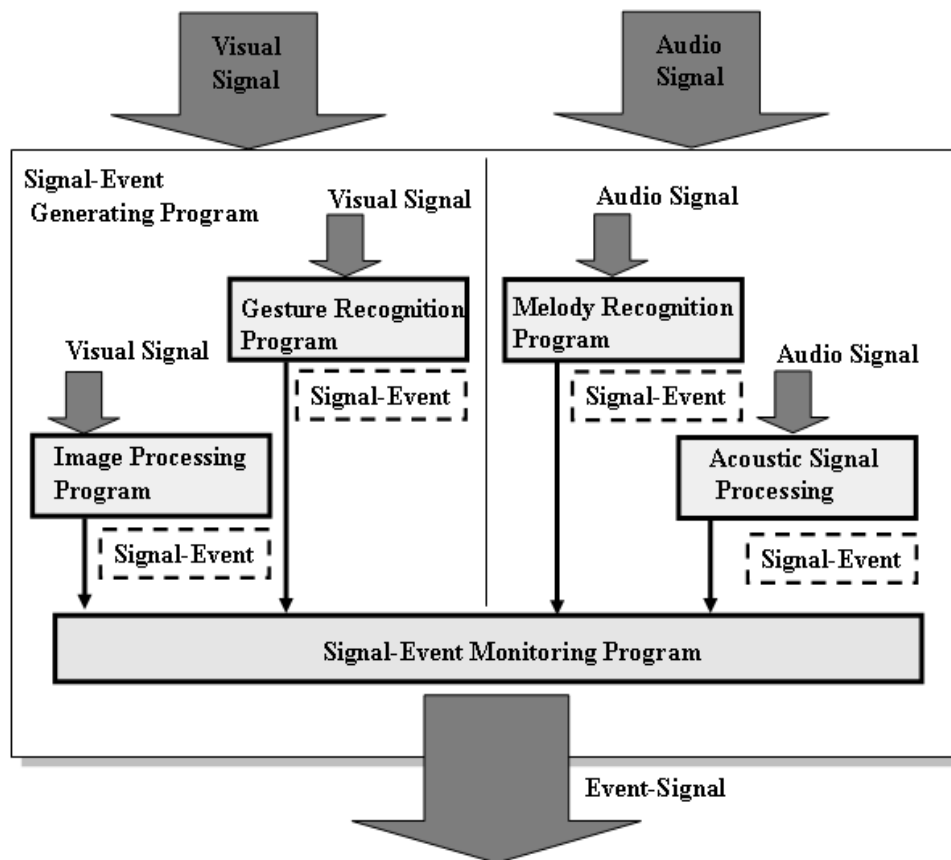


図 5.2: 信号イベントのモニタリング

5.5 動作仕様書

本節では、以上で述べた信号イベントを用いて、コビット計算機の動作制御を行うことを目的として、「動作仕様書」を用いる手法を提案する。この動作仕様書は、視聴覚情報を用いたインターフェースを統一環境で扱う上で、重要な役割を果たす。以下では、動作仕様書の例を示した後に、具体的な定義を行う。

5.5.1 動作仕様書の例

図 5.3 は、本研究で提案する動作仕様書を模式的に示したものである。ここでは、左から順に *SCENE1*、*SCENE2* の 2 つの利用場面に対応する仕様書（2 つの矩形領域）を示している。

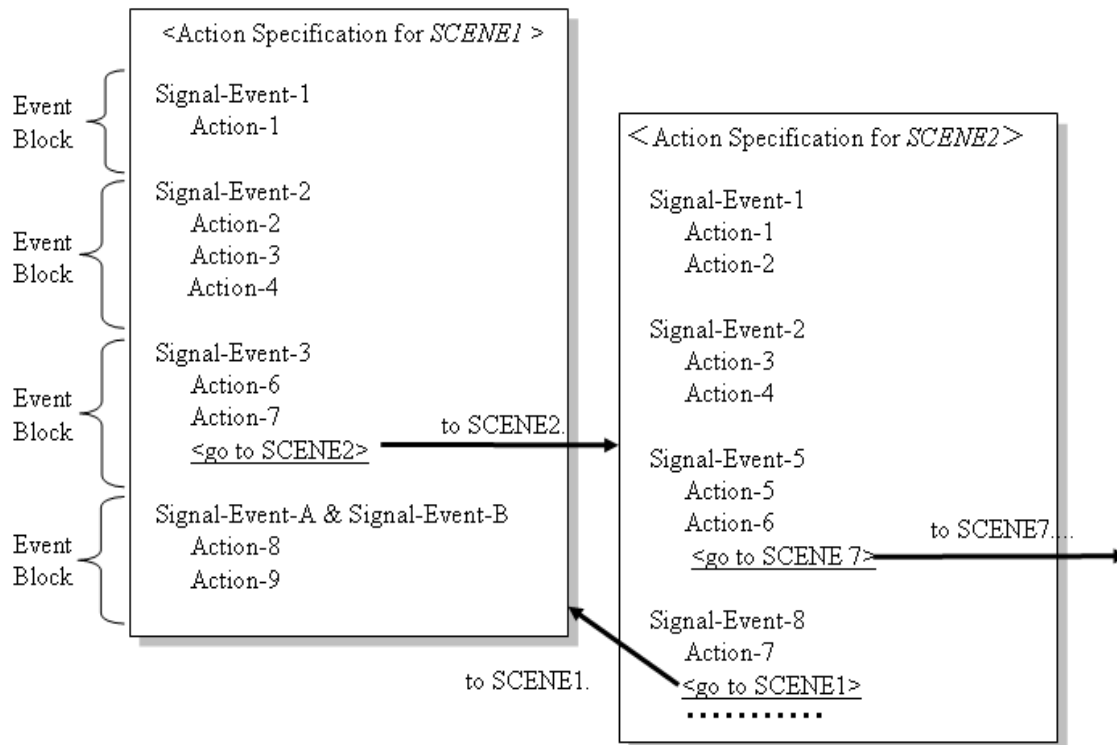


図 5.3: 計算機の動作仕様書の例

動作仕様書は、このように、1 つの利用場面に対して 1 つ定義し、1 つの動作仕様書は、1 つ以上の**イベントブロック** (event block) という単位から構成される。図 5.3 の左端の記述は、*SCENE1* には 4 つのイベントブロックがあることを例示したものである。

イベントブロックは、1 行の「信号イベント (Signal Event) の記述」と 1 行以上の「動作 (Action) の記述」によって構成される文書の単位である。コビット計算機は、図 5.1 の (4) のプログラムにおいて、「信号イベントの記述」の各行に書かれている文字列 (Action-Event-*) と一致する信号イベントのメッセージ (文字列) を受信した際に、その「記号イベントの記述」以下に続く一連の「動作の記述」に従って動作する。例えば、図 5.3 中の利用場面 *SCENE1* を例に挙げると、コビット計算機が信号イベントとして、"Signal-Event-2" という文字列を受信した場合は、Action-2、Action-3 および Action-4 で示される動作を順に実行する。

ここで、「信号イベントの記述」の行には、図 5.3 の *SCENE1* の 4 番目のイベントブ

ロックに示すように、複数の信号イベントを「&」の記号を用いて記述することができ、「信号イベント A と B が同時に起こった場合」などと、条件の複合的な利用が可能である。

一方、「動作命令」としては、以下の 2 種類を定義する。

1. 「コビット計算機に関する動作命令」
2. 「利用場面変更の動作命令」

1. 「コビット計算機に関する動作命令」は、「～を発話する」「カメラの視線方向を変更する」「～のプログラムを実行する」など、コビット計算機が主体となって作用する動作の記述である。

2. 「利用場面変更の動作命令」は、利用場面が変更されたことを知らせる信号イベントに対して、その利用場面に対する動作仕様書を読み取る動作である。図 5.3 中の *SCENE1*、*SCENE2* の *< goto... >* で示すように、ある信号イベントを引き金として、利用場面が変更する際に用いる。

5.5.2 動作仕様書の定義

例示した動作仕様書を、以下のように定義する。

まず、本システムで発生する N 個の信号イベントの文字列 $se(n)$, ($n = 0, 1, \dots, N-1$) の集合を以下のように定義する。

$$SE = \{se(n) \mid n = 0, 1, \dots, N-1\},$$

$$\text{かつ } p \neq q \rightarrow se(p) \neq se(q).$$

ここで、利用場面 $SCENE_i$ ($i = 0, 1, \dots, I-1$) に定義される j 番目の信号イベントの記述 $SignalEvent(i, j)$ ($j = 0, 1, \dots, J-1$) に、 $M(i, j)$ 個の信号イベントが含まれているとすると、その書式は

$$SignalEvent(i, j) := se(a_0)[\& se(a_1)[\& se(a_2) \dots [\& se(a_{M(i,j)-1})]]],$$

ここで、 $se(a_x) \in SE, x = 0, 1, \dots, M(i, j)$.

で表現できる。ここで、1 つの $SignalEvent(i, j)$ に対して定義される一連の動作命

令を $ActionCommand(i, j, k)$ ($k = 0, 1, \dots, K - 1$) とするとき、イベントブロック $EventBlock(i, j)$ は、以下の書式で定義できる。ここで、 $\langle return \rangle$ は改行を表わす。

$$\begin{aligned} EventBlock(i, j) := & SignalEvent(i, j) \langle return \rangle \\ & ActionCommand(i, j, 0) \langle return \rangle \\ & ActionCommand(i, j, 1) \langle return \rangle \\ & : \\ & ActionCommand(i, j, K - 1) \langle return \rangle \\ & \langle return \rangle \end{aligned}$$

このとき、 $SCENE_i$ に対する動作仕様書 $ActionSpecDoc(i)$ を以下で定義する。

$$\begin{aligned} ActionSpecDoc(i) := & EventBlock(i, 0) \langle return \rangle \\ & EventBlock(i, 1) \langle return \rangle \\ & : \\ & EventBlock(i, J - 1) \langle return \rangle \end{aligned}$$

5.6 信号イベント処理

信号イベント処理を行うプログラムは、図 5.4 に示す手順で動作する。

まず、計算機が起動した際に、プログラムは、システム利用者が任意に定義する「初期の場面」に対する動作仕様書を読み取る。

次に、動作仕様書に記述のある信号イベントを待機し、その信号を受信した場合は、動作命令を実行する。このとき、動作仕様書に記述されているものが、「利用場面変更の動作命令」の場合は、利用場面を変更するために、動作仕様書の再読み込みを行う。そうでない場合は「コビット計算機に関する動作命令」だとみなし、記述されている動作を実行する。

「コビット計算機に関する動作命令」は具体的には、以下の 3 種類の動作に大別される。

1. 任意のプログラムを実行する。
2. 可動式のカメラの視線方向を変更する。

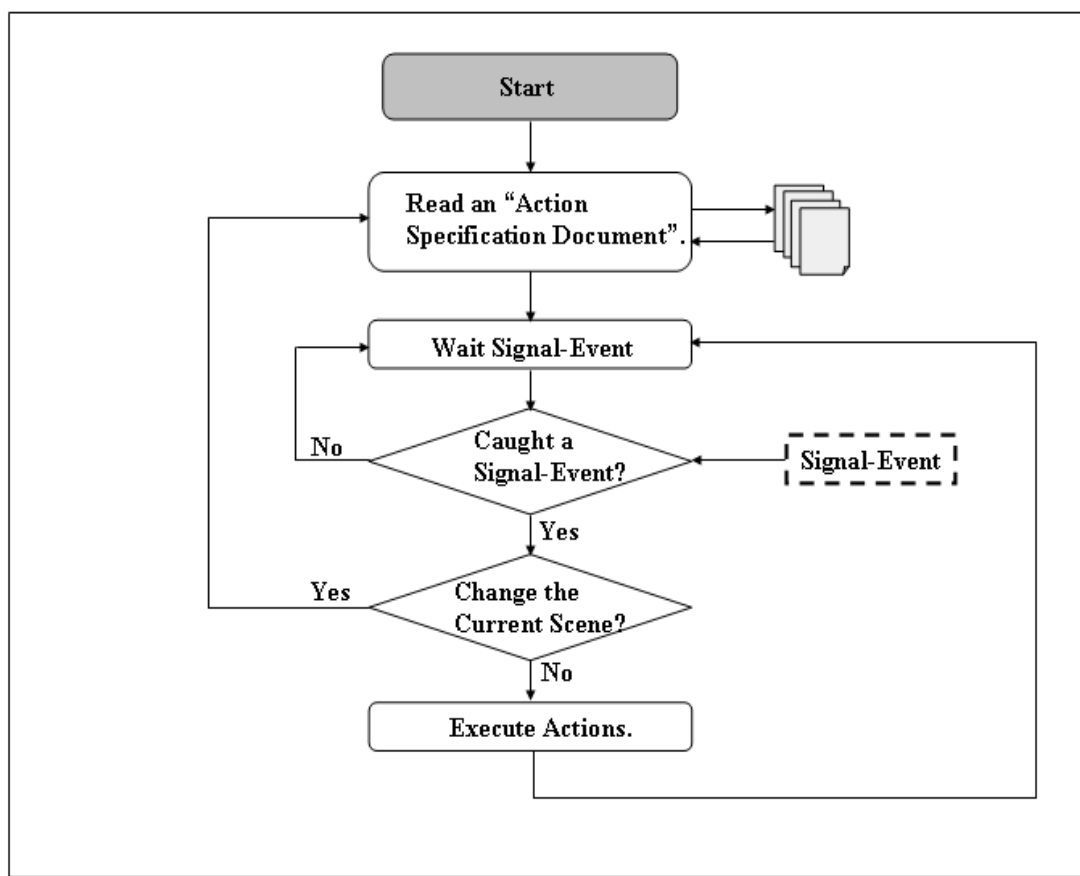


図 5.4: 信号イベント処理

3. 発話（テキスト読み上げ）プログラムによって、スピーカから発話する。

ここで、任意のプログラムの実行とは別に、カメラの視線方向の変更の機能と、発話機能を設定している理由は、本研究において、コビト計算機が人間との交信手段として、視聴覚インターフェースを用いることを前提としているためであり、計算機そのものが物理的に動く視覚インターフェースと、音を発する聴覚インターフェースを実現している*¹。

動作仕様書に記述されている一連の動作命令の実行が終了すると、プログラムは再び信号イベントの待機状態となり、この処理を繰り返すことでコビト計算機が動作する。

*¹ 現行のシステムでは、発話（テキスト読み上げ）プログラムは、マイクロソフト社がシステム開発者向けに提供する SAPI(Speech API)5.0 に含まれる発話エンジンを用いて、本研究用に開発したものである。また、カメラの動作は、KEYENCE 社が公開している、可動式カメラの仕様に従ったプログラムを独自に開発したものをを用いている。

5.7 むすび

本章では、コビット計算機システムのモデルを提示した。以下に本章のまとめを述べる。

- コビット計算機のモデルは、(1) 感覚センサー、(2) 視聴覚信号処理プログラム、(3) 信号イベント生成プログラム、(4) 信号イベント処理プログラム、(5) 動作仕様書、(6) 計算機動作の 6 つの部位から構成される。
- コビット計算機のモデルでは、計算機の動作を「動作仕様書」によって動的に制御し、視聴覚信号の認識結果を「信号イベント」という文字列情報に変換することで、利用場面に応じた、計算機インターフェースを実現できる。

次章では、本章で定義したコビット計算機のモデルに、第 3 章、第 4 章で述べた各認識手法を、マルチモーダルインターフェースとして具体的に実装する手法を述べる。

第 6 章

マルチモーダルインターフェースの実装手法

本章では、第 3 章、第 4 章で述べてきた各認識手法を、第 5 章で述べたコビット計算機のモデル上に、マルチモーダルインターフェースとして、具体的に実装する手法を述べる。

6.1 はじめに

コビット計算機のシステムでは、「信号イベントの生成」と「動作仕様書」の 2 つのモデルを有効に活用することで、利用場面によって意味解釈の異なる、マルチモーダルインターフェースが実現できることを述べた。そこで、「空書認識」と「歌声認識」の両手法の特徴を踏まえ、このモデルを適用することを考える。

まず、空書認識は、利用者の手の動きのパターンを実時間で計測しているため、空書を終えた時点で、信号イベントを発行すれば良いことが容易に推察される。この信号イベントが、空書した文字を意味すれば、動作仕様書において、信号イベントを受信した際に、ファイルに文字入力を行うなどの動作命令を記述しておけば良い。

一方、提案した歌声認識の手法は、ネットワーク上の曲 DB の検索処理を考慮しなければならない。また、ネットワーク上の情報資源のアドレスやアクセス方式の変更にも柔軟に対応できなければならない。そこで、歌声認識では、音声録音の終了を、信号イベントとして、信号イベント生成後は、録音された音を解析し、ネットワーク上の曲 DB にアクセスするプログラムを、動作仕様書内の動作命令で実行することが、実装に適していると考えられる。

以上の考察を踏まえ、以下では、本論文で述べた空書認識、歌声認識の各手法を、コビット計算機のシステム上に実装する手法について、信号イベントの生成と動作仕様書の定義方法を議論する。

6.2 空書認識の実装

本研究で提案するコビット計算機のシステムにおいて、空書認識を実装する手法を「信号イベントの生成」と「動作仕様書」の観点から述べる。

6.2.1 空書認識における信号イベント生成

空書認識の結果によって信号イベントを生成するためには、視覚信号を画像フレームとして受け取り、処理すれば良い。そこで、第 3 章で述べた一連の手法を使って、空書された文字を認識した場合に、その文字を信号イベントとして発する仕様とする。

本研究では、このように発する空書認識結果の信号イベントを表 6.1 の書式で表現する。ここで、空書認識の開始を示す信号イベントには、本システムでは、第 3 章で述べた仮想的なタッチパネルインターフェースを用いて、ボタン押下を示す信号イベントを生成して行う。

表 6.1: コビット計算機の信号イベント (空書認識)

信号イベント文字列	信号イベントの意味
CHARACTER_*	空書認識結果得られる文字。(* は英数字)

6.2.2 空書認識に対する動作仕様書

空書認識を、実際のコビット計算機のアプリケーションソフトウェアで用いる場合は、空書認識プログラムが生成する信号イベントを、適切な動作仕様書内に記述し、空書文字入力開始時に読み込めば良い。また、文字入力終了したときには、そのことを示す適当な信号イベントを定めておき、「もとの状態に戻る」、「新しい状態に移行する」などに対応する動作仕様書を、読み取ることで、計算機が次の動作に移行できる。

図 6.1 は、利用場面 SCENE1 において、信号イベント *Signal – Event1* を受け

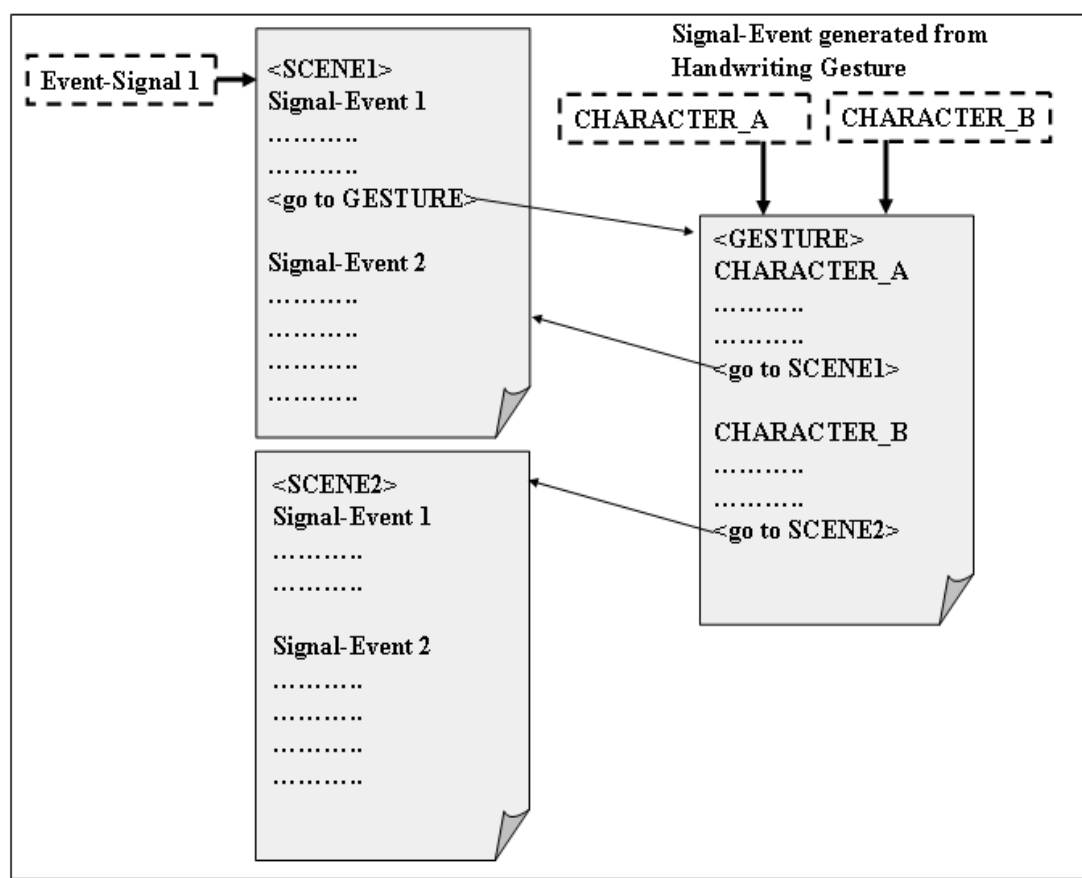


図 6.1: 空書認識に対する動作仕様書の実装モデル

取った場合、空書文字入力の利用場面 *GESTURE* に移行する動作命令の様子を示している。利用場面 *GESTURE* では、空書文字の信号イベント *CHARACTER_A* や *CHARACTER_B* を受け取った場合に、利用場面を各々 *SCENE1*、*SCENE2* に移行する。

以上のように、動作仕様書を適切に用いることで、コビット計算機システムでは、空書文字を用いた視覚インターフェースを実現できる。

6.3 歌声認識の実装

本研究で提案するコビット計算機のシステムにおいて、歌声認識を実装する手法を、空書認識と同様に、「信号イベントの生成」と「動作仕様書の定義」の観点から述べる。

6.3.1 歌声認識における信号イベント生成

第 4 章で述べた歌声認識を行って信号イベントを生成することを考える際、いくつかの具現化の手法が挙げられるが、本研究では、最も単純に処理を行う手法として、音声録音のみを、信号イベント生成処理の前に行い、他の処理は、すべて動作仕様書内の動作命令、「任意のプログラムを実行する」という機能を用いて、歌声認識用に作成したプログラムを実行することで実現する。この結果、実時間処理が必要な音声録音処理が実現でき、頻繁に変更が起こりうるネットワーク経由での曲検索処理に対しても、動作仕様書の記述変更のみで、柔軟に対応できる利点が得られる。

ここで、音声録音処理には、録音の開始・停止を検出する必要がある。音声録音の開始には、第 3 章で述べた仮想的なタッチパネルインターフェースを用いることが可能である。また、録音停止には、このインターフェースの他、沈黙時間を検出した信号イベントを発生させて行うことが可能である。

ここで、録音終了後の音声を用いて、歌声認識を実現するには、表 6.2 に示すように、録音終了を意味する信号イベントを定義し、この信号イベントの発生によって、音声解析、曲検索の実施に移行すると良い。

表 6.2: コビト計算機の信号イベント（歌声認識）

信号イベント文字列	信号イベントの意味
STOP_SINGING	歌声録音が終了した際に発行されるイベント。

6.3.2 歌声認識に対する動作仕様書

図 6.2 は、歌声認識に対する動作仕様書を示したものである。

この図 6.2 に示すように、音声録音の停止を意味する信号イベント *STOP_SINGING* が生成されたとき、動作仕様書では、以下の処理を行うプログラムを順次行うことで、歌声認識を実現する。

1. 録音された音声データから、旋律情報を抽出し、曲検索クエリを生成する。図 6.2 中では、*ANALYZE_VOICE.exe* という名称のプログラムを実行することを示

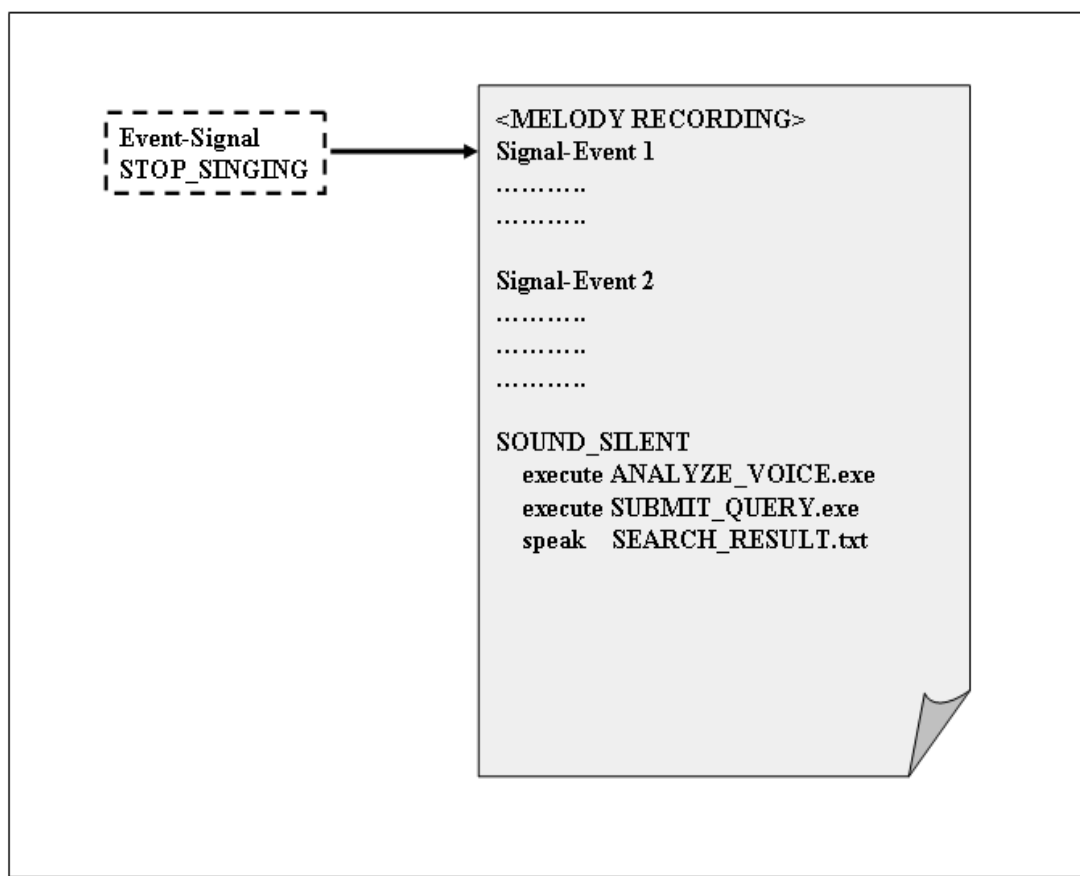


図 6.2: 歌声認識に対する動作仕様書の実装モデル

している。

2. ネットワーク上の曲検索サーバに曲検索クエリを送信する。図 6.2 中では、*SUBMIT_QUERY.exe* という名称のプログラムを実行することを示している。
3. 曲検索サーバから受け取った曲検索結果を音声によって、読み上げる。図 6.2 中では、*SEARCH_RESULT.txt* という名称のテキストファイルを読み上げる処理を示している。

本研究では、上述の処理 1,2 で用いる *ANALYZE_VOICE.exe*、*SUBMIT_QUERY.exe*

を、独自に開発したプログラムによって実現している。

6.4 マルチモーダルインターフェースの拡張に関する考察

本章で述べてきたように、コビット計算機システムでは、全く異なった信号処理を行っている空書認識、歌声認識の各手法を、「信号イベント」と「動作仕様書」のモデルによって、統一環境で計算機インターフェースとして実現できる。

ここで、特筆すべき点は、これらのモデルを利用することで、容易に他の信号を用いたインターフェースの構築に拡張できることである。

具体的には、新しく計算機インターフェースとして導入したい感覚信号がある場合には、その信号を取得する感覚センサーを計算機に接続し、プログラムによって感覚信号を取得処理する。次に、取得した信号を利用して、適切な「信号イベント」を発生させるプログラムを準備すれば、計算機に対する新たな交信手段として利用可能である。

また、動作仕様書を用いると、新しく導入した信号イベントに対しても、各々の利用場面に応じた異なる動作を、柔軟に定義できる利点がある。

以上のように、本研究で提案したコビット計算機システムのモデルは、複数の信号認識技術を、1つの計算機環境上で統合し、マルチモーダルインターフェースを実現する上で、拡張性が高く、極めて有効な手法であることが示唆される。

6.5 むすび

本章では、空書認識、歌声認識の各手法のコビット計算機のシステムモデルとして、実装する手法を述べた。本章のまとめを以下に述べる。

- コビット計算機モデルでは、認識手法の異なる空書認識、歌声認識の各手法に対して、「信号イベント」と「動作仕様書」を適用することで、計算機上の統一環境で、利用者の視聴覚信号を扱うことが可能となり、マルチモーダルインターフェースが実現できる。
- 空書認識では、利用者が空中で文字を描く動作を終えた時点で、その文字を意味する信号イベントを発することで、コビット計算機のモデルに組み込むことができる。
- 歌声認識では、利用者が歌声の録音を終了した時点で、その終了を示す信号をイベ

ントを発生させ、動作仕様書の中で、曲検索を行う一連のプログラムを実行することで、コビット計算機のモデルに組み込むことができる。

- 本モデルでは、視聴覚情報以外の信号に対しても、同様に適用でき、複数の信号を用いたマルチモーダルインターフェースを実現する上で有効である。

第 7 章

認識手法の評価

本章では、第 3 章、第 4 章で述べた空書認識、歌声認識の手法の有効性を定量的に分析する一連の実験に関して述べる。

7.1 まえがき

本論文の第 3 章、第 4 章では、各々空書認識、歌声認識を実現する各手法を提案し、その有効性を述べた。

ここで、これらの手法の有効性を、定量的に評価する必要がある。また、両手法を、コビット計算機システム上で同時に実施した場合、計算機インターフェースとして、実時間動作することを検証する必要がある。

以下では、空書認識、歌声認識の手法の有効性を評価するため、一連の実験を行った結果について述べる (7.2、7.3)。また、第 5 章で述べたコビット計算機システム上に、両認識手法を実装した際の実時間動作を評価した実験結果について述べる (7.4)。

7.2 空書認識の実験

本論文で提案した空書認識手法の有効性を実験的に確認するために、以下の一連の評価実験を行った。

- 文字の表記法による文字認識精度の差異の評価 (7.2.2)
- 赤外線照明による文字認識精度の評価 (7.2.3)

以下では、順番に実験環境、実験結果と考察を順に述べる。

7.2.1 空書認識の実験環境

本実験では、3.2.2 で述べた撮影条件 (室内、蛍光灯の照明、赤外線ライト) のもとで、デジタルビデオカメラ (SONY:PC-100) を利用者の肩に設置し撮影した。映像は、360*240 のフレームサイズで白黒 256 階調画像でデジタル画像として記録し、IEE1394 のインターフェースでラップトップ型 PC (CPU :*Intel(R) mobile pentium(R) III* 866MHz, RAM:384M, OS:WindowsXP Professional Version 2002) に転送し、処理を行った。

文字入力時、被験者 (20 代の男女 5 人 [男性 4 人:女性 1 人 (被験者 2)]、全員右利き) は、画像フレームの下側から手を伸ばし、タッチパネルボタン (押下判定の閾値 $T = 5$) を押して入力を開始した。

また、被験者が実験を行う際には、各個人の手の形状や傾きなどの特徴を活用するために、事前に各被験者の空書で、英数字 36 文字の各々に対して、ブロック体表記方法で 1 回ずつ、さらに提案した表記法で 1 回ずつ文字データベースに登録した。さらに、入力文字は、英数字のうちのいずれか 1 つを入力することを前提とし、'I' と 'l' や 'O(オー)' と '0(ゼロ)' などの類似した表記による誤りを防ぐため、入力の際に英字と数字は切替えて行った。

実験中の画像の背景は、図 3.2 のような室内の本棚 (被験者 1,2)、図 3.4 のような室内の机の前 (被験者 3,4,5) で行った。各被験者に、文字を書く際には、カメラ動かさないように指示して実験したところ、実際に空書を行った際のビデオカメラのぶれは、断続的に観察され、手の領域が重ならない画像フレームの上部におけるヒストグラムのずれによって計測したところ、最大で上下に 18 ピクセル、左右に 19 ピクセルであった。

手の領域を抽出する際の画素の値の範囲 ϵ は 5 とし、手が画像フレーム内にあるかどうかの閾値 $Hand$ を経験的に 1000 とした。また、文字判定の際の連続 DP マッチングでの距離の最大値は、100 となるように設定し、文字の決定処理の際には $Distance = 50, Time = 5$ とした。

表 7.1: 空中での手書き文字認識率

表記方法	被験者	成功数/入力数	認識率
ブロック体表記	被験者 1	44/72	61.1 %
	被験者 2	37/72	51.4 %
	被験者 3	35/72	48.6 %
	被験者 4	41/72	56.9 %
	被験者 5	36/72	50.0 %
	合計	193/360	53.6 %
提案する表記	被験者 1	63/72	87.5 %
	被験者 2	61/72	84.7 %
	被験者 3	56/72	77.8 %
	被験者 4	48/72	66.7 %
	被験者 5	43/72	59.7 %
	合計	271/360	75.3 %

7.2.2 文字の表記法による文字認識精度の差異の評価

文字の表記法が、システムの認識精度に与える違いを評価するために、通常ブロック体表記方法と **3.5.3** で提案した表記方法において、認識率を計測する実験を行った。

実験

英数字 36 文字の各々に対して、上述の 5 人の被験者が、ブロック体記法と、本論文で提案した表記法によって 2 回ずつ入力実験を行い、意図する文字が正しく入力された回数によって、認識率を計測した。被験者は事前にシステムの使用方法を習得するために、10 回程度の入力練習の後に各々の文字を登録し、実験を行った。表 7.1 に結果を示す。

考察

表 7.1 より、システムの認識率は、ブロック体の表記法の場合は被験者合計で 53.6 % であるのに対し、本研究で提案する表記法では 75.3 % であり、20% 以上の精度向上が確認できた。各々の被験者においても、本論文で提案する表記法がブロック体よりも高い精

度を示した。ブロック体の表記の数字では、特に '0' と '6' や '7' と '1' の間に、誤認識が多く見られた。

以上の結果より、文字の設計方法を変更することは、本論文で提案する空書のインターフェースの認識精度を向上する場合に、極めて有効な手法であることが考察される。例えば、文字の軌跡の最後に画像フレームの上方向に抜ける文字 (本論文で提案した表記法においては 'N', 'U' などが該当する) は、手の甲の部分が最後まで画像フレームに残るため、輝度差分の重心位置に影響が大きいため、誤認識につながる可能性が高いが、文字の設計方法を変更することで、大きな精度向上が期待できる。

7.2.3 赤外線照明による文字認識精度の評価

提案した一連の手法によって、夜間でも赤外線照明によって入力可能なインターフェースを構築できることを実験的に確かめるため、赤外線照明を用いた空書の実験を行った。

実験

外光がまったく入らない室内において、赤外線照射を利用して、**7.2.2** と同様の実験を被験者 1 で行った。このとき、図 7.1 に示すような赤外線画像が得られ、本論文で提案した 3 つの手法で解析した。

赤外線は照射方向によっては、赤外線画像フレームの中央が非常に明るく表示され、フレーム周囲の部分は逆に非常に暗くなるため、システムが指先の位置を追跡できずに、認識が成功しないことがある。このため、文字の設計を見直す必要が生じていた [?]

そこで、本実験環境のビデオカメラが赤外線ライトを設置可能な以下の 2 箇所についてそれぞれ、同じように入力を行った場合の認識率の違いについて検証を行った。

1. レンズの上部 (レンズ中央から赤外線ライトが垂直上方向 75 mm に位置)
2. レンズの下部 (レンズ中央から赤外線ライトが垂直下方向 35 mm に位置)

結果を表 7.2 に示す。

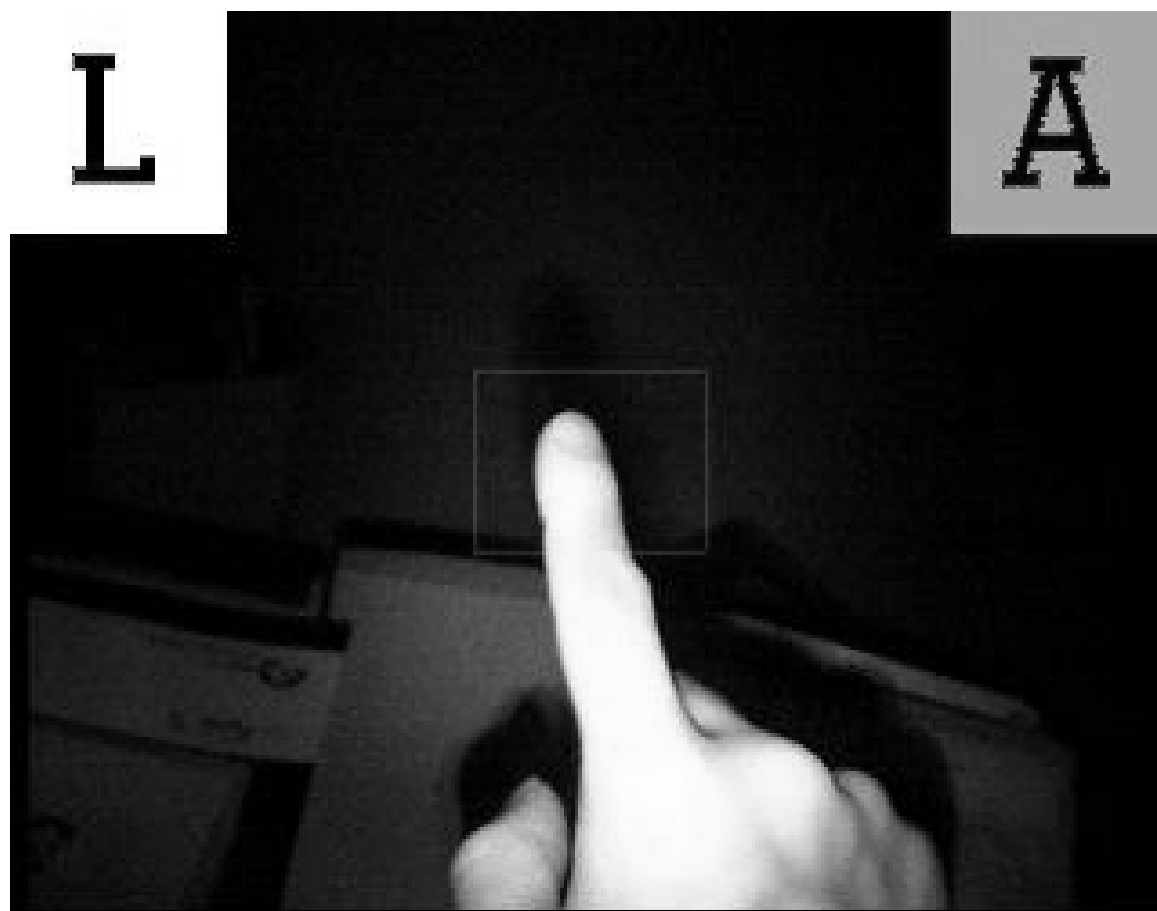


図 7.1: 赤外線ライト照明によって得られた画像

表 7.2: 赤外線照明と蛍光灯照明との認識率の比較

照明条件	成功数/入力数	認識率
蛍光灯	63/72	87.5 %
赤外線（レンズ上部 75mm）	57/72	79.2 %
赤外線（レンズ下部 35mm）	64/72	88.9 %

考察

実験の結果、レンズの位置が (1) の場合には、画像フレームの周囲では、指先が暗く表示され、ほとんど判別が難しくなることに対し、(2) の場合には、利用者の指先が画像フレームのどの位置にあっても、背景画像よりも極めて明るく映し出されることが確認できた。また、表 7.2 の結果からも、赤外線がビデオカメラレンズのより近くから照射された

場合は、本研究の手法を用いて、蛍光灯の場合と変わらない精度で文字入力の実現できることが分かる。

以上のことから、夜間に赤外線ライトを用いて本システムを用いる際には、赤外線の照射位置が極めて重要であり、画像フレーム中のどの位置でも、指先を追跡できるように赤外線ライトを設置した場合には、本研究で提案する一連の手法が有効に活用できることが確認できた。

7.3 歌声認識の実験

本論文で提案した歌声認識手法の有効性を実験的に確認するために、以下の一連の評価実験を行った。

- 曲検索システムのスループットの評価 (7.3.2)
- 曲データベース拡張性の評価 (7.3.3)
- 並列化した計算機上での曲検索システムの総合評価 (7.3.4)

以下では、順番に実験環境、実験結果と考察を述べる。

7.3.1 歌声認識の実験環境

歌声認識の手法を評価するために、ネットワーク上に曲検索サーバを 6 台の PC (CPU: *IntelPentiumTM III* processor 850-MHz、RAM: 256 MB) 上に実装した。

曲データベースには、標準 MIDI ファイルの形式で 21,500 曲 (1980 年～2002 年に発表された日本のポピュラーミュージックから 1,500 曲と、プログラムによって、ランダムな音符を 1,000 個並べて作成した 20,000 曲) を作成し、実験に用いた。

被験者は 20 代の男性 3 名、女性 1 名によって各々 400 回、合計 1,200 回の歌声の入力を行った。このとき、音声録音時間は 15 秒以内に制限したところ、平均音符数は 28.0 であり、被験者の音声以外には環境の音は録音されない指向性のあるマイクロホンを入力に用いた。音声は、サンプリング周波数 8kHz、8bits、モノラルの条件で録音し、FFT は、1 フレーム 512 点で処理し、フレームシフト幅は 128 点とした。

また、本実験には、独自に開発した WWW ブラウザのプラグインとして実装した曲検索クライアントを用いて、PC 上で行った。

7.3.2 曲検索システムのスループットの評価

本研究で提案した 2 つの中継サーバのうち、CGI サーバによって、曲検索サーバを並列化することで、システム全体のスループットが向上できることを数値的に確認するために、曲検索サーバ数とスループットとの関係性を評価する実験を行った。

実験

本実験では、5 台の PC 上に実装した各曲検索サーバの曲のデータベースに、上述した 21,500 曲をそれぞれ登録し、1 台の CGI サーバによって、利用者のクエリデータを分散させる実験を行った（図 4.3）。実験には、CGI サーバが次に利用者のクエリデータを中継することができる曲検索サーバの数を 1 台から 5 台まで徐々に増やし、各々の場合におけるシステム全体のスループットを計測した。なお、スループットは、1 秒間に処理できる利用者クエリデータの数によって計測した。図 7.2 に、実験結果を示す。

考察

図 7.2 では、曲検索サーバ数の増加に比例して、システム全体のスループットが向上できることが観察できる。この結果から、本研究で提案した 2 つの中継サーバのうち、CGI サーバによるデータ中継のモデルは、システム全体のスループット向上に役立つことが言える。

7.3.3 曲データベース拡張性の評価

本研究で提案した 2 つの中継サーバのうち、検索管理サーバによって、曲検索サーバを並列化することで、データベースの拡張性を実現できることを数値的に確認するために、曲データベースのサイズと検索時間の関係性を評価する実験を行った。

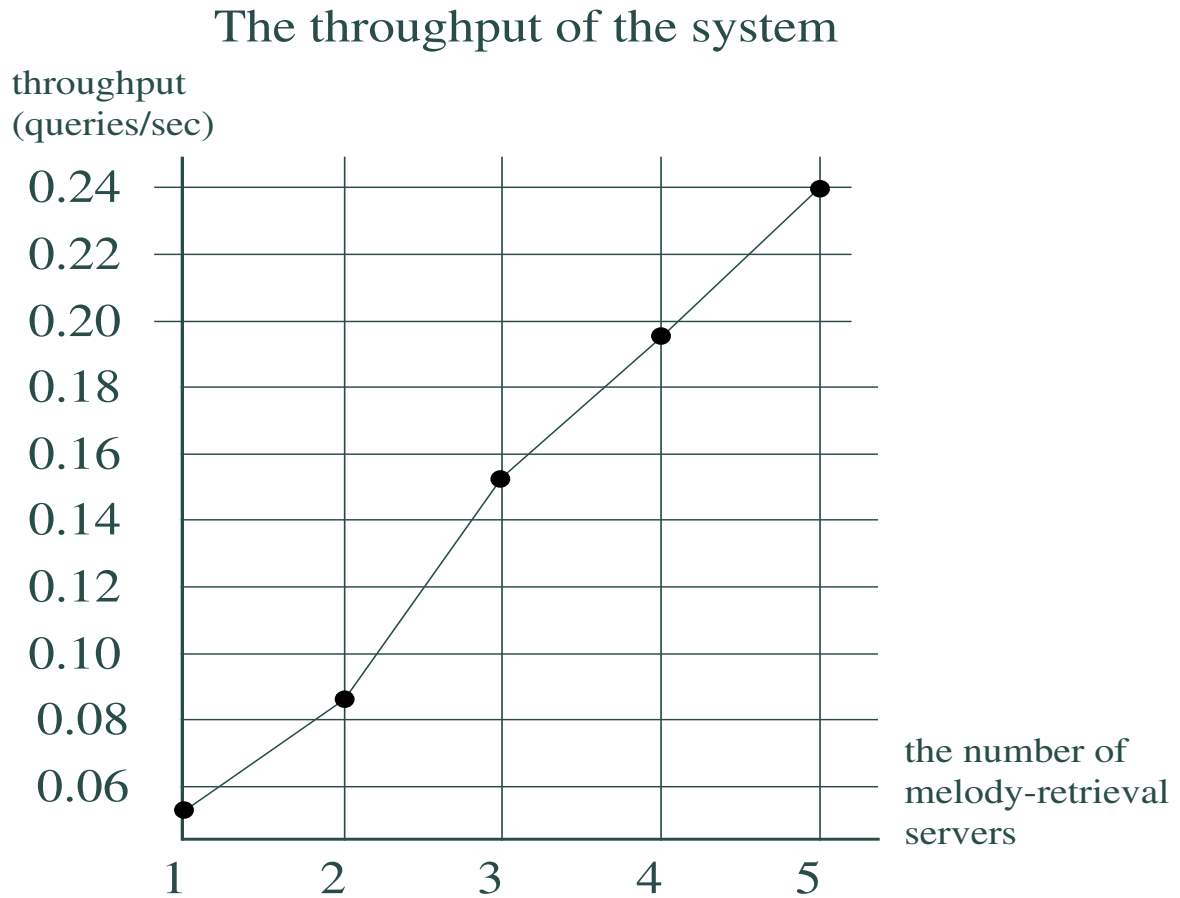


図 7.2: システム全体のスループットと並列化する曲検索サーバの数の関係

実験

本実験では、1 台の検索管理サーバと、異なる曲データベースを持つ 5 台の曲検索サーバによって行った。システム全体の曲データベースの曲数を 4,300 から 21,500 まで増加させることを、曲検索サーバの増加によって行い、各々の曲データベースのサイズにおける検索時間を計測した。検索システム全体の構成は、図 4.3-(3) のようにし、1 台の検索管理サーバで、複数台の曲検索サーバの結果を集計した結果を利用者側に送信した。また、この実験では、被験者 1 から得られた長さ 28 音符の検索クエリデータを用いて行い、検索時間は、検索管理サーバが検索クエリデータを受信した時間と、検索結果を送信した時間の差によって計測した。表 7.3 に実験結果を示す。この結果は、各データベースサイズにおいて検索を 5 回行った場合の平均時間を示している。

表 7.3: 曲データベースの拡張性の評価実験

parallel-ized servers (<i>DB</i> size)	the search time(sec)
1 (4,300)	4.62
2 (8,600)	5.22
3 (12,900)	5.44
4 (17,200)	5.45
5 (21,500)	5.48

考察

表 7.3 の結果から、曲検索システム全体の曲データベースサイズが増加した場合でも、検索時間をほぼ一定に保てることが確認できた。この結果から、本研究で提案した中継サーバのうち、検索管理サーバは、曲データベースの拡張性を実現する上で、有効であることが確認できた。

7.3.4 並列化した計算機上での歌声検索システムの総合評価

本研究で提案した歌声認識の手法が、多数の曲目の候補から、高速に精度良く正答の曲を特定するために有効であることを総合的に評価する実験を行った。

実験

システム全体で、21,500 曲のデータを保有する曲検索サービスを想定し、本実験では、前節の実験と同様に、1 台の検索管理サーバと 5 台の曲検索サーバによって、システムを構築し、各曲検索サーバには、4,300 曲の曲データを分散して保有させた。本実験では、曲検索正答率と検索時間を計測し、正答率は、被験者が検索を行った際に、意図する曲目が、検索結果の曲リストの最上位に位置している回数によって求めた。

1,200 回の曲検索結果を、表 7.4 に示す。

表 7.4: 並列化した計算機上での曲検索実験の結果

DB 中の曲数	21,500 曲
並列化した曲検索サーバの数	5
曲検索サーバ 1 台あたりの平均曲データベースサイズ	4,300 曲
実験に用いた検索回数	1,200
検索時の平均入力音符数	28.0
検索の平均正答率	70.1%
検索の平均検索時間	5.48 secs

考察

表 7.4 で得られた結果は、検索正答率が 約 70%、検索時間が約 5.5 秒という高い性能を示した。この結果は、本研究で提案した歌声認識の手法が、多数の曲目の候補から、精度良くかつ高速に正答の曲を絞り込むという問題に対して、極めて有効な手法であることを示している。

また、検索に失敗した約 30 % のクエリデータに共通して言える現象として、以下が観測された。

1. 男性の低音で声がかすれた場合に、クライアント側で、音符の区切れだと誤検出してしまい、正答以外の曲が上位になる現象。
2. 比較的ゆっくりとしたテンポの曲で、入力される音符数が 8 以下の場合に、上位の多くの曲で、入力旋律との DP マッチングの距離に差異が生じなかった現象。
3. 曲データベースの音符の出現位置と、利用者の記憶が大きく異なっていた。

1. の現象に対しては、本研究の手法では、歌声を入力する際の無音時間を、音符の区切りとみなしてしまうため、手法の改善が必要だと考えられる。この指針として、かすれ声の位置では、周波数領域のピーク位置が不安定に乱れることに対し、明確な音符の区切れでは、発音時に比較的、ピーク位置が安定して出現する現象を利用することが考えられる。

2. の現象に対する解決方法として、テンポのゆっくりとした曲に対しては、利用者によくの音符数を入力するようにうながすような、利用者インターフェースの導入が必要だと考えられる。

3. に対しては、検索に失敗した場合は、利用者に複数回、同じ曲目に対する検索を行ってもらい、各検索時に得られる DP マッチング距離を利用者クライアント側で管理し、その合計値によって、徐々に正答の曲目を絞る手法が挙げられる。

以上の問題は、すべて利用者の入力に起因しているため、今後は、曲検索クライアントの研究を深めることで、本手法の認識精度が大きく向上していくものと考察される。

7.4 コビト計算機の実時間動作の実験

空書認識と歌声認識の両手法を、同時にコビト計算機上で実施した場合のシステム全体の実時間動作の性能を評価するため、図 7.3 に示す試作機において、画像信号取得のスループットと遅延時間を計測した。また、音声信号取得時のデータ欠損について確認した。



図 7.3: 試作したコビト計算機

以下では、順番に実験環境、実験結果と考察を順に述べる。

7.4.1 実験環境

本実験では、空書認識の実験、歌声認識の実験の節で述べたものと同様のラップトップ型計算機 (CPU :*Intel(R) mobile pentium(R) III* 866MHz, RAM:384M,

OS:WindowsXP Professional Version 2002) を利用し、コビット計算機のデータ処理装置として用いた。

図 7.3 の試作機に示すように、デジタルビデオカメラ、スピーカ、マイクロホン、市販のぬいぐるみの中に設置し、ぬいぐるみを利用者の肩の上に固定することで、擬似的に装着型計算機を実現した。ここで、デジタルビデオカメラは、7.2 で用いたものより小型の USB カメラ (Creative LAB: CT6840) を用いて、映像は、360*240 のフレームサイズで白黒 256 階調画像でデジタル画像として記録した。一方、音声は 7.3 と同様に、サンプリング周波数 8kHz、8bits、モノラルの条件で録音した。

7.4.2 システムのスループットと遅延の評価

コビット計算機に対する計算機インターフェースとして、提案した空書認識、歌声認識の各手法が有効に機能することを評価する実験を行った。ここで、システムの実時間動作の評価を、画像処理のスループットとシステム反応時間の遅延時間によって計測することで行った。また、歌声認識の手法では、録音データを用いて検索クエリを生成するため、音声信号取得時のデータ欠損について確認する実験を行った。

実験

第 5 章で述べたモデルに従って、実験環境の計算機上に、コビット計算機システムを実装し、画像信号取得のスループットと遅延時間、音声信号取得後の録音データによって、コビット計算機のシステムの性能を評価する。

画像信号と音声信号は、C++ 言語で開発したプログラムによって取得した後に、通信ポート経由で、コビット計算機システムの信号イベント生成部分に送信し、空書認識、歌声認識部のプログラムに渡す。

このような実装を行った場合の画像処理のスループットを、10 分間計測した。このとき、スループットは、計測時間中に処理できたフレーム数によって求めた。また、画像処理によるシステムの遅延を計測するため、本システムの装着型カメラによって、ストップウォッチに表示される時刻を撮影し、本システムの利用者ディスプレイ上に表示させた様子を、別のデジタルスチルカメラで静止画として 20 回撮影した。この静止画フレーム内

には、実際のストップウォッチの時刻と、画像処理後に利用者ディスプレイに表示されるストップウォッチの時刻の 2 つが表示されるように撮影するため、利用者が感じる時刻の遅延を正確に計測することができる。

表 7.5 に実験結果を示す。

表 7.5: コビット計算機システムの画像信号取得のスループットと遅延時間

parallel-ized servers (<i>DB</i> size)	the search time(sec)
スループット	12.5 fps (10 分間の平均値)
遅延時間	2.1 フレーム (約 168msec)

一方、音声の録音について、録音時、あるいはネットワーク通信時におけるデータ欠損の有無を確認するために、コビット計算機の信号イベント生成プログラムに対して、音の信号を 30 秒間送信し、信号イベント生成プログラムで録音されたデータと録音時間から、データ欠損を計測する作業を 20 回行ったところ、録音データに欠損は観測されなかった。

考察

空書認識に対する画像処理の計測結果のスループットと遅延時間は、本システムを実時間で利用する場面においても、利用者に大きなストレスを与えないことが確認できた。また、人間が、画像フレーム内に描かれる指の動きを撮影した動画像を観察した場合でも、深刻なフレーム落ちはなく、描かれる文字の軌跡が判読可能であった。

一方、歌声認識においても、実験ではデータ欠損が観測されず、録音された音声データを用いて、7.3 の実験と同様に、ネットワーク上の曲データベースから曲目を特定可能であることを確認した。

以上の結果、本研究で提案したコビット計算機システムのモデルにおいて、空書認識、歌声認識の各手法が有効に機能することが確認できた。

7.5 むすび

本章では、本論文で述べてきた空書認識、歌声認識を定量的に評価した。以下に本章のまとめを述べる。

- 空書認識の評価実験では、本論文で提案した文字の表記法が通常のブロック体よりも高い認識率を達成することが示された。また、赤外線照明に対しても、可視光線照明と同様に、提案手法が有効であることが示された。
- 歌声認識の評価実験では、本論文で提案した曲検索サーバの並列化の手法が、曲データベース中の曲数、および利用者数の増加に対応できることが示された。また、本手法により、高い検索精度が達成できることが確認できた。
- 空書認識、歌声認識の両手法を試作したコビト計算機上で同時に動作させた場合に、実時間動作し、計算機インターフェースとして、有効に機能することが確認できた。

第 8 章

研究の今後の方向性

本章では、本論文で提案した一連の手法について、研究の今後の方向性を議論する。

8.1 まえがき

本研究で提案した空書認識、歌声認識、およびコビット計算機システムのモデルの手法に関する有効性を、第 7 章で定量的に評価した。その評価結果を踏まえ、各手法の現在の問題を議論し、各手法に関する研究の今後の方向性を示すことが必要である。

以下では、順に「空書認識 (8.2)」、「歌声認識 (8.3)」、および「コビット計算機 (8.4)」に関する研究の今後の方向性を考察する。

8.2 空書認識に関する研究の今後の方向性

本論文では、タッチパネル型のインターフェースにより、利用者の手の領域を特定し、輝度差分の重心位置で指先の軌跡を追跡し、連続 DP マッチングによる文字列を特定するという一連の手法を提案した。また、独自の英数字の表記法により、通常の表記よりも認識精度が 20% 以上向上することが確認できた。また、これらの手法を用いることで、赤外線画像処理に対しても有効なシステムを構築できた。

現在の実装では、予め用意された文字を利用者は入力するシステムであるが、空書される文字は軌跡が残らないという特徴を活かし、利用者が自由に設定した記号などによるパスワード入力などの応用システムにも、本手法は応用できるものと考えられる。

今後の研究の方向性として、連続文字入力や英数字以外の日本語や様々な文字を入力す

る手法を開発し、簡易なジェスチャを用いて、様々な文字入力可能な装着型計算機環境の研究を進める予定である。また、空書中の背景に移動物体が写っている場合の認識手法についても、研究を進めたい。

8.3 歌声認識に関する研究の今後の方向性

本論文では、多数の曲目の候補から、精度良く高速に正答の曲を特定するために、曲検索サーバの並列化の手法を提案した。本研究で提案した 2 種類の中継サーバを組み合わせることで、曲検索サーバを並列化でき、システムのスループットの向上とデータベースの拡張性を実現する手法が明らかとなった。また、検索正答率 70%、検索時間 5.5 秒という高い性能を実現できた。

第 7 章で評価した結果、本研究で提案した歌声認識が失敗する場合の要因は、本システムで想定した仕様において、利用者の録音が成功しなかったことであつた。このため、利用者の歌声の解析技術を高めるとともに、より良い利用者インターフェースを構築することが今後の課題として挙げられる。

また、現行のシステムでは、音響信号を記号化した、スタンダード MIDI ファイルを用いた実装となっているが、音響信号をそのまま用いて検索を行う手法も研究が進められている [Goto et al., 1999, Nishimura et al., 2001]。街中で聞く音を、歌うのではなく、そのまま認識することで、利用者に情報を提示できるシステムの実現は、本研究においても、極めて重要な課題であり、今後、研究を進めていく予定である。

8.4 コビット計算機に関する研究の今後の方向性

本研究では、コビット計算機概念を提示し、肩の上で動作する計算機のシステムモデルを提案した。また、人間とコビット計算機との交信手法として、空書認識と歌声認識に焦点を当てて議論し、統一されたシステム上で、マルチモーダルインターフェースとして実現する手法を明らかにした。さらに、人間の発する信号は、利用場面によって意味解釈が異なることに対応するために「動作仕様書」のモデルを提案し、利用場面に応じた信号処理を柔軟に行うことができることを示し、外界の信号を「信号イベント」に変換して処理するモデルを適用することで、新たな信号を、マルチモーダルインターフェースとして導入

できることを述べた。

本論文では、計算機のインターフェースを研究の中心として議論を進めてきたが、コビット計算機概念において定義したように、今後は、人間の発する信号だけでなく、外界の様々な信号を能動的に解釈し、人間の活動を補助するための認識技術の研究が必要である。

また、コビット計算機の応用範囲は限りなく広く、「盲導犬のように道案内をするシステム」や「電話通信によって利用者の代わりに他者と会話してくれるエージェントシステム」など、様々なアプリケーションに適用される。このような応用技術に対して必要な基礎技術も、同時に研究が必要となると考えられる。

さらに、本研究では「動作仕様書」は予め人間が作成しておく必要があったが、より気の効いた新しい計算機環境を構築するためには、コビット計算機が自ら利用者の状況を学習し、必要な動作仕様書を作成、改変、削除していく機能を付加する機能が必要であり、研究を進める予定である。

8.5 むすび

本章では、提案した空書認識、歌声認識およびコビット計算機に関する研究の評価結果を考察し、各々の研究分野における課題を掲げ、研究の今後の方向性を示した。以下に本章のまとめを述べる。

- 空書認識の研究では、今後、パスワード入力などの応用システムの実現をはじめ、連続文字列や英数字以外文字の認識、背景に移動物体が写っている場合の認識等を実現する方向性が挙げられる。
- 歌声認識の研究では、利用者の音声の解析の精度を高めるとともに、記号化された歌声の処理だけでなく、音響信号処理を用いた認識技術の実現を目指す方向性が考察される。
- コビット計算機の研究では、本論文で述べた利用者インターフェースの実現の課題に加え、環境から得られる信号に対して、能動的に反応し、利用者の活動を補助・拡張するという目的の実現が、方向性として考えられる。また、多くの応用技術を実現するための基礎的研究や、学習によってコビット計算機が、利用者の状況に応じて

動作仕様書を作成・変更していく技術の研究が必要である。

第9章

結論

本章では、本論文の研究成果のまとめを述べる。

本論文では、利用者の肩の上で動作する計算機として、コビット計算機概念を提案した。また、人間とコビット計算機との通信手段として、人間の発する自然なコミュニケーション手段を用いたマルチモーダルインターフェースの実現を目的として、空書認識と歌声認識の実現手法を明らかにした。また、これらの認識手法を、コビット計算機の統一環境で用いるためのシステムモデルを提案し、視聴覚信号をはじめ、様々な信号を計算機インターフェースとして用いる手法を明らかにした。さらに、空書認識と歌声認識の有効性を評価実験により明らかにし、従来は実現が困難であった計算機インターフェースを構築できることを実証した。

これまでの各章で述べた内容の要旨は、次のとおりである。

第1章では、本研究の目的、背景、意義を明らかにした。本研究は、肩の上で動作する装着型計算機に対するマルチモーダルインターフェースの実現を目的としている。本研究は、空書認識、歌声認識を利用した計算機インターフェースの研究分野に位置づけらる。空書認識の研究分野においては、接触型、あるいは固定カメラによって撮影された従来の認識技術とは異なり、非接触型でかつ装着型計算機での認識を実現するものである。歌声認識の研究においては、従来の技術では実現が難しかった、高精度かつ高速な歌声検索を、多数の曲を保有する曲データベースに対して実現するものである。本研究は、新しい計算機概念を提案する意義とともに、各認識技術は、画像処理、音声認識、情報検索、ロボティクスなど、幅広い工学的分野に貢献する意義がある。

第 2 章では、コビット計算機概念を定義した。コビット計算機は、人間の肩の上で動作し、利用者や外界の環境から得られる感覚信号を認識し、人間にとって自然なコミュニケーション手段で交信して、人間の活動を補助・拡張することを要件とした計算機概念である。

第 3 章では、空書認識の手法を提案した。この手法では、利用者が空中で描く文字を特定するために、様々な照明条件のもとで、画像フレーム中の利用者の手の領域を抽出し、画像フレーム間の輝度差分の重心位置を手がかりとして、空中での利用者の手の動きを追跡する。この手法と連続 DP マッチングを併用することで空書される文字を、特定することが可能となる。

第 4 章では、歌声認識の手法を提案した。この手法では、歌声検索を行うサーバを並列化することで、高い楽曲検索精度を保ちながら、高速な検索が実現できるようになる。また、本手法では、2 種類のデータ中継サーバを提案し、曲データベース中の曲数や利用者数に応じた、拡張性の高いシステムを構築できる。

第 5 章では、コビット計算機のシステムモデルを提案した。コビット計算機は、視聴覚情報などの感覚情報を取得するセンサー、その情報を処理するプログラム、感覚情報を解釈した結果である「信号イベント」を生成するプログラム、信号イベントによって計算機を制御する「動作仕様書」、信号イベントを処理し、動作仕様書の命令を実行するプログラムと、動作命令から構成される。

第 6 章では、空書認識と歌声認識の各手法を、コビット計算機のシステム上に具現化する手法を提案した。信号イベント生成モデルと動作仕様書のモデルを導入することで、これら異なる 2 つの認識技術を統一環境で具現化することができる。また、視聴覚情報以外の信号に対しても、コビット計算機のモデルは適用でき、計算機インターフェースを構築する上で有効であることが示唆される。

第 7 章では、空書認識と歌声認識の各手法について、定量的な評価実験を行った。空書認識は赤外線照明のもとで有効に動作し、本研究で提案した英数字の表記は、ブロック体のものよりも精度高く認識することが検証された。歌声認識は、本研究で提案した 2 種類のデータ中継サーバは、曲検索サーバのスループットの維持とデータベース拡張性に有効であることを実証し、精度が高く、高速な検索を実現できることが確認できた。空書認識、歌声認識をコビット計算機上で同時に実施した場合にも、有効に動作することが確認で

第 9 章 結論

きた。

第 8 章では、空書認識、歌声認識およびコビット計算機の課題を掲げ、各分野の研究の方向性を示唆した。空書認識では、パスワード入力などの応用システムの実現をはじめ、連続文字列や英数字以外文字の認識、背景に移動物体が写っている場合の認識等を実現する研究の方向性が挙げられる。歌声認識では、利用者の音声の解析精度の向上を目指すとともに、音響信号処理を用いた認識技術の実現を目指す研究の方向性が考察される。コビット計算機の研究では、利用者から得られる信号に加え、環境から得られる信号に対して、能動的に反応する技術を実現し、コビット計算機が利用者の状況に応じて、動作仕様書を生成・変更していく技術の達成が目標とされる。

本研究により、肩の上の計算機に対して、空書認識、歌声認識を用いたマルチモーダルインターフェースの実現方法が明らかになった。

そして、肩の上の計算機が、利用者や外界の環境から得られる感覚信号を認識し、人間にとって自然なコミュニケーション手段で交信して、人間の活動を補助・拡張する、そのようなコビット計算機社会の光景が、ほんの少し見えた。

謝辞

本論文は、私が早稲田大学 大学院理工学研究科 村岡洋一研究室に在籍していた期間に行った研究をまとめたものです。本研究を進める過程でご指導、ご支援を賜りました多くの方々に深謝致します。

村岡洋一教授、深澤良彰教授、中島達夫教授、山名早人教授には、本論文の審査を行って頂きました。

村岡洋一教授には、私が教授の研究室に所属して以来、6年間に渡ってご指導を賜りました。先生は常に世界的視野で、科学の在り方、研究者としての在り方、をお示しになりました。また、自らそれらを実践されるお姿は、私の研究者としての在り方、人間としての生き方を、強く決定付けるものでした。先生は、常に高い目標を設定すること、一流を見て一流として通用する人間になることという、大きな課題をお示しになり、私がその達成に向けて歩む過程で幾度となく方向を見失う度に、正しい道を明るく照らし、決して人生に妥協しないように奮い立たせてくださいました。先生のご指導、ご支援なしに、本研究はあり得ませんでした。

深澤良彰教授には、本研究のコビット計算機概念や手法に関する貴重なご意見・ご助言を頂きました。また、先生の授業で学んだソフトウェア工学の知識は、本研究の歌声認識を事業化し、実社会で実践する上で、極めて重要なものでした。

中島達夫教授には、本研究の将来目指すべき方向や、技術的な課題、可能性に関して、実社会に照らし合わせた貴重なご意見・ご助言を賜りました。先生のご示唆されたコビット計算機インターフェース概念の組み立て方は、本研究の今後の方向性を確認する上で、大変、意義深いものでした。

山名早人教授には、本研究の第一歩であった歌声検索システムに着手する前から、検索システムに関する数多くの最先端の技術を、授業の中で教えて頂きました。山名先生の検

索エンジンに関する研究成果は、本研究の歌声認識で用いた曲検索システムを事業化する、きっかけとなりました。

上田和紀教授、大石進一教授には、本論文の審査の場において、極めて有用なご助言・ご示唆を頂きました。また、授業を通して学んだ知識情報処理の考え方や情報エントロピーを用いた情報の評価手法は、本研究のコビット計算機や視聴覚インターフェースの構築には欠かせないものでした。

村岡研究室の諸氏には、議論、助言、示唆という形で、また、研究室での生活の場で、言葉に言い尽くせないほど、お世話になりました。

後藤真孝氏には、研究者としてどうあるべきかという、大切な姿勢をはじめ、歌声検索における主要な技術に関する御助言、音楽情報科学、音声認識などの最先端の研究、論文作成、学会発表の方法など、私が研究を遂行するために必要な大部分について、御指導を頂きました。また、多岐に渡る最先端の研究成果を、常に実践してお示しになるお姿は、私を含む研究室諸氏の目標となりました。同氏の「ビートトラッキングシステム」の研究成果である、軽快な音楽に合わせて踊る「CG ダンサー」を拝見したことが、私の研究者としての出発点でした。

貝塚智典氏には、歌声認識の基本技術となる音声波形解析、曖昧な情報の照合技術について、多くのご指導を頂きました。同氏の「歌声の旋律情報と歌詞情報をキーとした曲検索システム」が本論文の歌声認識の出発点でした。

福盛秀雄氏には、私の研究に対して数多くのご助言を頂いただけでなく、日進月歩する最先端の技術を、あらゆる角度から捕らえるという、研究者として生命線となる姿勢をお示しになりました。また、本研究を6年間、暖かくご支援して頂きました。

田村健人氏には、日本で最初の検索エンジン「千里眼」について、ご指導を頂き、本研究の歌声検索エンジンの前身となった文字列検索エンジンを実現することができました。

首藤一幸氏には、本研究のシステムを構築・実装する上で、極めて多くのご助言を頂きました。また、並列分散処理に関する知識をはじめ、本研究の基盤となる重要な技術を御指導・御示唆頂いたばかりでなく、本論文の執筆にあたり、多くの情報をご提供頂きました。さらに、家庭を大切にされながら、博士課程での研究活動を一心に取り組まれる御姿は、同じように家庭を持った私のお手本となりました。

興梠正克氏には、本研究の空書認識における画像処理に関して、多くのご助言を頂きまし

第 9 章 謝辞

た。また、拡張現実感の実現をはじめ、画像処理分野における最先端の研究成果をお示しになり、本研究における装着型計算機を考えるきっかけを与えてくださいました。同氏の「パノラマベースドアノテーションシステム」を拝見したことが本論文の空書認識の出発点でした。

根山亮氏には、並列分散プログラムの考え方や研究当時では確立されていなかった Java 言語の技術について、ご指導頂き、本研究を完成させることができました。また、研究室での日々の活動の中で多岐に渡るご指導を頂きました。

内野聡氏、水野裕識氏、三浦敏孝氏、河野洋一氏、伊藤克信氏、松本英明氏には、本研究の初期段階で、多くのご指導・ご助言を頂きました。また、研究室の卒業生として、あるいは社会人として、私を含む多くの学生の研究に役立つ情報を、数多くご提示頂き、研究を遂行する上で、大きな刺激となりました。

秋岡明香氏には、並列処理の技術を始め、本研究に関する多くの議論を行って頂き、また、私の研究室での毎日の活動を支えて頂きました。

岩村誠氏には、研究室在籍以前から、企業において一緒にソフトウェア開発に取り組み、たくさんの刺激を受けました。また、本研究に対して、鋭い視点で、数多くのご助言を頂いただけでなく、歌声検索システムを具現化する上で、多くの技術供与を頂きました。

池長俊哉氏、清水佳奈氏には、1 年以上の長い期間に渡り、本研究の歌声認識の研究を、あらゆる面で支えて頂き、また、一緒に開発に取り組むことができました。また、本研究を事業化する上で、あらゆる面でご支援を頂き、数々の御智恵を頂きました。

柴田祐助氏、宮之前謙一氏、久和祐介氏には、本研究に多くの助言を頂いただけでなく、本論文で述べた各システムの開発に直接関わって頂き、本研究の発展に多大な貢献を頂きました。

森村賢一氏とは、画像処理を用い、ジェスチャを認識する研究を、共同で取り組むことができました。また、私の研究活動を支えて頂きました。

畠山勇樹氏、増保利行氏には、本研究を遂行してきた中で、多くの解決すべき課題を、あらゆる角度から指摘して頂き、本研究の発展に、道標を与えてくださいました。また、私の研究活動を支えて頂きました。

加藤浩一氏には、研究と実社会を結び付ける大きなきっかけを、様々な形で頂きました。これは、私が研究者として経済・社会を再考し、企業経営者として研究を見つめ直す

大切な機会となりました。

高橋里恵氏は、私を含めた村岡研究室諸氏の活動を常に支えて下さいました。

楠本修二郎氏、入川秀人氏には、本研究の歌声検索を事業化する上で、経営者としての在り方をご指導して頂きました。また、アカデミックな場を離れた実社会において、国内外の多くの方々に、本研究を評価して頂く機会を設けて頂きました。さらに、本研究の事業化にあたり、施設面、精神面において、ご支援を賜りました。

本研究の一部は、日本学術振興会特別研究員の奨励費および科学研究費補助金、早稲田大学 DC 奨励金によって遂行されました。

他にも、多くの方々からご支援、ご助言を頂きました。私の研究活動を可能にするために、「肩」をお貸しくださった、すべての方々に深く感謝致します。

最後に、私の研究生活を見守り続けてくださった福岡県と佐賀県の家族に、そして、私を信頼し支え続けてくれる妻の綾香と、本論文執筆中、妻のお腹の中で、私の歌声に元気に応えて励ましてくれた娘の紫子^{ゆかりこ}に、心から感謝致します。

参考文献

- [Merton, 1965] Robert K. Merton: *On the Shoulders of Giants*, A Shandean Postscript 31, 1965.
- [BOLT, 1987] Richard A. BOLT, *The Integrated Multi-Modal Interface*, 電子情報通信学会誌 D, VolJ70-D, No.11, pp.2017-2025, 1987.
- [Hirai et al., 1998] Hirai,Hirose,Haikawa and Takenaka: *The Development of Humanoid Robot*, Proceedings of the 1998 IEEE International Conference on Robotics and Automation, pp.1321-1326,May 1998.
- [Yamasaki et al., 2002] Yamasaki F., Endo K., Asada M., Kitano H.: *Energy-Efficient Walking for a Low-Cost Humanoid Robot, PINO*, AI magazine, Vol 23, No.1, pp.60-61, Spring, 2002.
- [Hashimoto, 2001] 橋本 周司:
ヒューマノイドロボットー人間と機械の新しい関係を目指してー, 映像情報メディア学会誌、Vol. 55, No. 12, pp.1547-1555, 2001.
- [Masahiro et al., 1997] Masahiro Fujita and Koji Kageyama:
An Open Architecture for Robot Entertainment Proc. of Autonomous Agent 97, pp.435-442, 1997.
- [Inoue et al., 2001] Inoue,H., Tachi,S., Nakamura,Y., Hirai,K., Ohyu,N., Hirai.S., Tanie,K., Yokoi,K., and Hirukawa,H: *Overview of Humanoid Robotics Project of METI*, Proc. Int. Symp. Robotics, 2001.
- [Ono et al., 2000] 小野哲雄, 今井倫太, 江谷為之, 中津良平:
ヒューマンロボットインタラクションにおける関係性の創出, 情報処理学会論文誌, Vol. 41, No. 1, pp. 158-166, 2000.
- [Jebara et al., 1997] Tony Jebara, Cyrus Eyster, Josh Weaver, Thad Starner and Alex Pentland: *"Stochasticks: Augmenting the Billiards Experience with Probabilistic Vision and Wearable Computers"*, Proceedings of the International Symposium on Wearable Computers, October 1997.
- [Kourogi et al., 1999] 興梠 正克, 蔵田 武志, 坂上 勝彦, 村岡 洋一: パノラマ画像群を位置合わせに用いた注釈つき映像の実時間提示システム, 信学技報, PRMU-99-64, pp.1-8, September 1999.
- [Kurata et al., 2000] 蔵田 武志, 興梠 正克, 栗田 多喜夫, 村岡 洋一, 坂上 勝彦: 気の利いたウェアラブルビジョンシステムの実現に向けて, 信学技報, PRMU-99-270, pp.65-71, March 2000.
- [Kurata et al., 2001] 蔵田 武志, 大隈 隆史, 興梠 正克, 坂上 勝彦: ハンドマウス: ビジュアル

- ウェアラブルズが可能にする拡張現実環境に適したインターフェイス, 信学技報, PRMU-2000, pp.69-76, 2001.
- [Kato et al., 2001] 加藤 丈和, 蔵田 武志, 坂上 勝彦: VizWear-Active - 記憶補助のための顔画像検出, 追跡, 登録 - 映情学技報, VIS2001-102, Vol.25, No. 85, pp.41-46, 2001.
- [Kato et al., 2002] 加藤 丈和, 蔵田 武志, 坂上 勝彦, 向川 康博, 尺長 健: 実環境下における自律的な顔画像登録とその応用, 情処研報, 2002-CVIM-131, pp.45-52, 2002.
- [Murakami 2001] Taisuke Murakami, *Contact Water*, Art Garalley of SIGGRAPH 2001, 2001.
- [Kurata et al., 2000-2] 蔵田 武志, 坂上 勝彦, 興梠 正克, 村岡 洋一, : 気が利いたウェアラブルビジョンシステムの実現に向けて - 状況把握に有効な要素機能の実時間実装, 画像ラボ, 2000.6.
- [Yamamoto et al., 1999] 山本 吉伸, 椎尾 一郎: 空気ペン - 空間への描画による情報共有 -, 情報処理学会第 59 回全国大会 3ZA-3, 1999.
- [Katagiri et al., 2001] 片桐 雅二, 杉村 利明: ビデオカメラを用いた空中署名による個人認証の試み, 信学技報, PRMU2001-34, pp.9-16, June 2001.
- [Yoon et al., 1999] Ho-Sub YOON, Jung SOH, Byung-Woo MIN and Hyun Seung YANG: Recognition of Alphabetical Hand Gestures Using Hidden Markov Model, IEICE TRANS. FUNDAMENTALS, VOL.E82-A, NO.7 JULY 1999.
- [Sonoda et al., Mar.2001] 園田 智也, 村岡 洋一: 空中での手書き文字認識システム - ウェアラブルコンピュータ環境での入力インターフェース -, 情報処理学会インタラクション 2001 講演論文集 pp.3-10, 2001.3.
- [Sonoda et al., Sept.2001] 園田 智也, 村岡 洋一: 赤外線カメラで撮影した画像による空中での手書き文字入力システム, 情報処理学会第 63 回全国大会 6J-04, 2001.
- [Kubota et al., 1998] 窪田 進, 堀 修: こう配法の誤差モデルとオプティカルフローの定量的な信頼度, 電子情報通信学会論文誌: D-II, Vol.J81-D-II, No.8, pp.1760-1767. 1998.
- [Takahashi et al., 1994] 高橋 裕信, 関 進, 小島 浩, 岡 隆一: ジェスチャー動画像のスポッティング認識, 電子情報通信学会誌, Vol.J77-D-II, No.8, pp.1552-1561, 1994.
- [Mukai et al., 1998] 向井 理朗, 高橋 裕信, 遠藤 隆, 中沢 正幸, 松村 博, 岡 隆一: マルチモーダルヒューマンインタフェースのノートパソコンへの実装, 信学技報, PRMU-98-70, pp.69-75, May 1998.
- [Oka et al., 1999] 岡 隆一, 西村 裕一, 向井 理朗: しぐさで伝える, 電子情報通信学会誌, Vol.82, No.4, pp.332-339, 1999.
- [Takahashi et al., 1994] 高橋 勝彦, 関 進, 小島 浩, 岡 隆一: ジェスチャー動画像のスポッティング認識, 電子情報通信学会論文誌, D-II, vol.J77-D-II, No8, pp.1552-1561, Aug. 1994.
- [Imano et al., 1995] 今野 潤, 八木正紀, 山内英明, 安村通晃, ジェスチャー入力によるマルチモーダルインタラクティブシステムの試作, 情報処理学会ヒューマンインターフェイス研究会, Vol.60 No.9, pp.65-72, 1995.
- [Hasegawa 2000] 長谷川修: マルチモーダル研究の現状と展望, 信学技報, PRMU2000-106, pp.47-52., November 2000.
- [Kawahara et al., 2000] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 伊藤克亘, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏: 日本語ディクテーション基本ソフトウェア (98 年度

- 版) , 日本音響学会誌, Vol.56, No.4, pp.255-259. 2000.
- [Yamamoto 1988] Yamamoto, N. : *Music Database*, Journal of Information Processing Society of Japan, 29(6), pp.107-111, 1988.
- [Kageyama et al., 1993] T.Kageyama, K.Mochizuki, and Y.Takashima: *Melody Retrieval with Humming*, ICMC Proc., pp.349-351, 1993.
- [Ghias et al., 1995] Asif Ghias and Jonathan Logan *Query By Humming – Musical Information Retrieval in an Audio Database*, ACM Multimedia 95, Electronic Proc., 1995.
- [Sonoda et al., 1998j] 園田智也, 後藤真孝, 村岡洋一: WWW 上での歌声による曲検索システム, 電子情報通信学会論文誌: D-II, VOL.J82-D-II, No.4, pp.721-731. 1998.
- [Sonoda et al., 1998] Tomonari Sonoda, Masataka Goto, Yoichi Muraoka: *A WWW-based Melody Retrieval System*, pp.349-pp.352, ICMC'98 Proc., 1998.
- [Smith et al., 1998] Lloyd A. Smith, Rodger J. McNab, and Ian H. Witten: *Sequence-Based Melodic Comparison: A Dynamic-Programming Approach*, pp.101-pp.116, COMPUTING IN MUSICOLOGY 11, 1997-98.
- [Bainbridge 1998] David Bainbridge: *MELDEX: A Web-based Melodic Locator Service*, pp.223-pp.229, COMPUTING IN MUSICOLOGY 11, 1997-98.
- [Goto et al., 1999] Masataka Goto, and Satoru Hayamizu : *A Real-time Music Scene Description System: Detecting Melody and Bass Lines in Audio Signals*, Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis, pp.31-40, August 1999.
- [Kosugi et al., 2000] Naoko Kosugi, Yuichi Nishihara, Tetsuo Sakata, Masashi Yamamuro, and Kazuhiko Kushima : *A Practical Query-By-Humming System for a Large Music Database*, ACM Multimedia 2000.
- [Sonoda et al., 2000] Tomonari Sonoda and Yoichi Muraoka: *A WWW-based Melody Retrieval System -An Indexing Method for a Large Database-*, ICMC2000 Proc., 2000.
- [Nishimura et al., 2001] T. Nishimura, H. Hashiguchi, J. Takita, J. Xin Zhang, M. Goto and R. Oka, : *Music Signal Spotting Retrieval by a Humming Query Using Start Frame Feature Dependent Continuous Dynamic Programming*, Proceedings of the 2nd Annual International Symposium on Music Information Retrieval (ISMIR 2001), pp.211-218, 2001.
- [Sonoda et al., 2002] Tomonari Sonoda, Toshiya Ikenaga, Kana Shimizu and Yoichi Muraoka: *A Melody-Retrieval System on Parallel-ized Computers*, Proc. of IWEC2002, pp.261-268, 2002.

著者研究業績

論文

種別	題名	発行年月	掲載誌名	著者
論文誌				
1 ○	WWW 上での歌声による曲検索システム	1999 年 4 月	電子情報通信学会論文誌: D-II, Vol.J82-D-II, No.4, pp.721-731	園田智也 後藤真孝 村岡洋一
2 ○	空中での手書き文字入力システム	掲載決定	電子情報通信学会論文誌: D-II	園田智也 村岡洋一
査読付 国際会議				
3 ○	A WWW-based Melody Retrieval System	1998 年 10 月	International Computer Music Conference (ICMC1998)	Tomonari Sonoda Masataka Goto Yoichi Muraoka
4 ○	A WWW-based Melody Retrieval System - An Indexing Method for A Large Database -	2000 年 9 月	International Computer Music Conference (ICMC2000)	Tomonari Sonoda Yoichi Muraoka
5 ○	A Melody Retrieval System On Parallelized Computers	2002 年 5 月	International Workshop on Entertainment Computing (IWEC2002)	Tomonari Sonoda Toshiya Ikenaga Kana Shimizu Yoichi Muraoka
6	A Whole-Body-Body-Gesture Input Interface with a Single-View Camera - A User Interface for 3D Games with a Subjective Viewpoint	2002 年 7 月	The International Conference on Imaging Science, Systems, and Technology 2002 (CISST'02)	Kenji Morimura Tomonari Sonoda Yoichi Muraoka
7 ○	THE DESIGN METHOD OF A MELODY RETRIEVAL SYSTEM ON PARALLELIZED COMPUTERS	2002 年 12 月	Web Delivering of Music (WEDELMUSIC2002)	Tomonari Sonoda Toshiya Ikenaga Kana Shimizu Yoichi Muraoka

講演

種別	題名	発行年月	掲載誌名	著者
研究会 8 ○	WWW 上での歌声による曲検索システム	1998 年 2 月	情報処理学会音楽情報科学 研究会研究報告 98-MUS-24-4, Vol.98, No.14	園田智也 後藤真孝 村岡洋一
9	「ICMC 2000 レポート」	2000 年 12 月	情報処理学会音楽情報科学 研究会研究報告 2000-MUS-38-1, Vol., No.14	平田圭二 石川修 鈴木健嗣 園田智也 瀧陽一郎 松田周 米澤朋子
一般講演 10 ○	歌声による曲検索システム	1997 年 9 月	情報処理学会第 55 回全 国大会講演論文集	園田智也 後藤真孝 村岡洋一
11 ○	仮想空間中のマルチ ユーザ環境を利用した WWW 上での曲検索シ ステム	1999 年 9 月	情報処理学会 第 58 回全 国大会演論文集	園田智也 根山亮 後藤真孝 村岡洋一
12 ○	DP マッチングを用い る検索の DB 索引作成 方法と応用-大規模曲検 索への応用-	2000 年 9 月	情報処理学会 第 60 回全 国大会演論文集	園田智也 村岡洋一
13 ○	空中での手書き文字認 識システム - ウェア ラブルコンピュータ環 境での入力インターフ ェース -	2001 年 3 月	情報処理学会インタラク ション 2001 講演論文集 pp.3-10	園田智也 村岡洋一
14 ○	赤外線カメラで撮影し た画像による空中での 手書き文字認識システ ム	2001 年 9 月	情報処理学会 第 63 回全 国大会演論文集 6J-04	園田智也 村岡洋一

その他

種別	題名	発行年月	掲載誌名	著者
国内特許 15 ○	歌声による曲検索システム	1998 年 3 月	平成 10 年特許願第 111273 号	園田智也
16 ○	ネットワークを用いた歌声による曲検索システム	1998 年 12 月	平成 10 年特許願第 378084 号	園田智也
17 ○	空中で手描きした文字・記号入力システム	2001 年 3 月	特許出願 2001-105219	園田智也
米国特許 18 ○	WWW-based Melody Retrieval System	1998 年 3 月	Serial Number 09/272,211	Tomonari SON-ODA
欧州特許 19 ○	WWW-based Melody Retrieval System	1998 年 3 月	No.99302171.6	園田智也

特記事項

- 論文 8. により、情報処理学会山下記念賞 受賞 (1997 年 9 月)。
- 論文 10. により、情報処理学会第 55 回全国大会奨励賞 受賞 (1999 年 3 月)。
- 歌声認識の研究成果により、早稲田大学 Zaiya.com 「ベンチャーコンテスト 2001」 準グランプリ 受賞 (2000 年 7 月)。
- 歌声認識の研究成果を事業化し、うたごえ有限会社設立 (2001 年 1 月)。
- 歌声認識の研究成果により、日経新聞・VEC 主催「ベンチャーコンテスト 2001」 準優勝 受賞 (2001 年 5 月)。