動画像処理を用いた
新しいマンマシンインタフェースの研究

間　瀬　健　二

目　次

1　序論 **1**

1.1　視覚を用いたインタフェース 1

1.2　ヒューマンイメージリーダ 4

1.3　従来の研究と関連する研究 5

1.3.1　人間が送るメッセージとインタフェース 6

1.3.2　人間からのメッセージ抽出に関する研究 7

1.3.3　メッセージの統合 9

1.3.4　他の分野との関連 9

1.4　本論文の概要と構成 10

2　頭部動作の抽出と理解 — ヘッドリーダ — **12**

2.1　はじめに 12

2.2　頭部動作検出の原理とアルゴリズム 13

2.2.1　頭部の投影変換モデル 13

2.2.2　回転・移動パラメータの抽出 16

2.3　画像処理アルゴリズムとインプリメンテーション 17

2.3.1　頭部領域の抽出 17

2.3.2　顔領域の抽出 19

2.3.3　インプリメンテーション 19

2.4　動作検出実験と考察 21

2.5　リアルタイム頭部動作認識合成実験システム 22

2.6　頭部動作の理解と応用 25

2.6.1　コマンド認識 26

2.6.2　アプリケーション事例 27

2.7　むすび 30

3　表情の認識 — フェイスリーダ — **34**

3.1　はじめに 34

3.2　表情とオプティカルフロー 35

3.2.1 顔の表情に関する研究 35
3.2.2 顔と表情の画像処理 36
3.2.3 FACS(Facial Action Coding System) 37
3.2.4 オプティカルフロー抽出アルゴリズム 39
3.2.5 表情筋 41
3.3 筋肉モデルによる表情の記述 41
3.3.1 表情筋の動き推定 41
3.3.2 筋肉モデルと FACS の関係 43
3.4 表情記述の実験結果 44
3.4.1 筋肉モデルによる記述 44
3.4.2 FACS に基づく筋肉モデルの記述 47
3.4.3 筋肉モデルによる感情の識別 47
3.5 感情を表わす表情の識別 50
3.5.1 顔の表情と感情 50
3.5.2 特徴ベクトルの決定 50
3.5.3 k- 最近傍法による特徴識別 52
3.6 感情識別の実験結果 53
3.6.1 データ収集 53
3.6.2 特徴ベクトルの決定（学習） 53
3.6.3 識別結果 53
3.7 むすび 56

4 発話動作の認識 — リップリーダ — 59
4.1 はじめに 59
4.2 計算機による読唇の研究 59
4.3 筋肉モデル 61
4.3.1 発声に関わる筋肉 61
4.3.2 口唇画像のオプティカルフロー 62
4.4 英数字認識システム 65
4.4.1 画像入力と前処理 65
4.4.2 単語検出とマッチング 67
4.5 実験結果 69
4.6 むすび 69

5 物体フローの推定と歩行者の計数 — ピープルリーダ — 73
5.1 はじめに 73
5.2 時空間画像の解析と歩行者計数 73

5.2.1 時空間画像解析によるフロー抽出 73
5.2.2 歩行者計数技術の動向 75
5.2.3 画像処理による歩行者の計数 75
5.3 直交断面像による物体フロー推定 75
5.3.1 動物体の非エピポーラ面画像 76
5.3.2 直交2断面法によるフロー推定 78
5.4 物体フローの推定実験 81
5.4.1 シミュレーション実験 83
5.4.2 歩行者画像による実験 83
5.5 1断面法による歩行者計数 85
5.5.1 斜交投影像からのフローの方向判別 86
5.5.2 歩行者計数システム 86
5.5.3 領域数の逐次数え上げを目的としたラベリング法 88
5.5.4 モーメント計算 90
5.5.5 速度の推定方法 90
5.5.6 人数の計数法 90
5.6 実験システムの構成と実験結果 91
5.7 むすび 93

6 結論 **95**

謝辞 **98**

参考文献 **99**

研究発表一覧 **104**

図一覧

1.1 人間が伝えるメッセージ 3
1.2 ヒューマンイメージリーダ (Human Image Reader) 5

2.1 頭の回転にともなう頭部と顔部の形状と重心の変化 13
2.2 頭部の 3 次元モデル 14
2.3 y 軸回りの回転による重心の移動 (a) 水平断面上の顔領域の投影像と振舞い (b) 回転角と重心の変位の関係 14
2.4 x 軸回りの回転による重心の移動 (a) 垂直断面上の領域の変化 (b) 回転角と重心の変位の関係 15
2.5 動き抽出のための特徴量 (a) 特徴パラメータ (b)z 軸回りの回転と物体座標系 17
2.6 ヘッドリーダの画像処理ブロック図 18
2.7 画像処理結果（頭部輪郭と顔部領域の抽出） 20
2.8 ヘッドリーダの操作メニュー 20
2.9 CG による顔画像の例 21
2.10 CG 顔画像に対する回転量の推定結果 (a) x 軸回り (b) y 軸回り 22
2.11 実画像を使った動きの計測 23
2.12 リアルタイム頭部動作認識合成実験システム 24
2.13 認識合成実験システムのプログラム構成 25
2.14 数値データのフォーマット 25
2.15 ハードウェアテクスチャマッピングを使った頭部の表示例 26
2.16 頭部動作によるウィンドウ選択 28
2.17 電子秘書のシステム構成 31
2.18 電子秘書の動作 32

3.1 顔の表情筋 41
3.2 表情筋の動作検出窓 42
3.3 表情筋の動作推定例 (幸福) 42
3.4 オプティカルフロー抽出結果 (女性；幸福の表情) 45
3.5 筋肉の動き推定量の時間変化 (前図の女性) 46

3.6 筋肉の動き推定量の時間変化 (次図の男性) 48
3.7 別の人物のオプティカルフロー抽出結果 (男性、幸福、口は閉じたまま) 49
3.8 別の表情（嫌悪） 49
3.9 表情動画像の学習サンプル 52
3.10 特徴ベクトルを構成する点 54
3.11 識別実験に用いた表情動画像の例 55

4.1 入力画像と動き検出窓 61
4.2 口唇まわりのオプティカルフロー 63
4.3 窓内の速度成分の平均値による動きデータ 64
4.4 特徴量 $O(t)$ と $E(t)$ 65
4.5 オプティカルフローを用いた読唇システム 66
4.6 /one/ から /four/ までの数字の連続発声の時の $O(t)$ と $E(t)$ 67
4.7 認識辞書として用いたテンプレートパターン 67
4.8 単語のマッチング手順 68
4.9 実験データの特徴量 70

5.1 レンジファインダによる歩行者計数システム (市販されているシステムの構成図 (推定)) 74
5.2 動画像の時空間表現 76
5.3 直交 2 断面と物体の斜交投影像 77
5.4 典型的な直交 2 断面の配置 77
5.5 スリットの配置と斜交投影像の関係 79
5.6 直交断面法による物体フロー 80
5.7 物体フロー推定シミュレーション結果の例 82
5.8 シミュレーションにおける推定誤差 82
5.9 フロー推定実験にもちいた歩行者画像の一部 83
5.10 歩行者画像のフロー推定結果 84
5.11 斜交投影像の傾きと方向の正負 85
5.12 歩行者計数システム基本処理ブロック図 87
5.13 計数実験対象画像の例 91
5.14 領域抽出・計数結果 92

表一覧

1.1 情報入力用デバイス、機能と、人間のメッセージ発出メディア 6

2.1 押下キーとシステム動作の割当 29

3.1 FACS のアクションユニットの一覧 (AU 番号と名称) 38
3.2 筋肉窓とアクションユニットの関係 44
3.3 感情認識結果 58

4.1 単語認識結果：被験者 1 の辞書とのマッチング (被験者 1-[A,B,C,D],2,3) 71

5.1 歩行者計数結果 (10 分間における計数結果, in/out) 91

第 1 章

序論

やさしさに包まれたならきっと，目に映るすべてのことはメッセージ
——— 荒井由美　“やさしさに包まれたなら”

1.1　視覚を用いたインタフェース

画像処理による情景の認識理解は，自然言語理解や音声認識と並んでコンピュータをより知的にして，作業の能率をあげたり人間にできない作業を肩代りしてもらうことを容易にするために不可欠な機能である．今日，コンピュータの性能は驚くほど向上し，高度な数値計算やシステム制御を瞬時に行なうことが可能である．また一方で装置が超小型化され，電話器やファクシミリなどの通信機器，あるいは電子レンジやテレビのような家電製品や自動車にまで使われるようになっている．しかしながら，それらの人間に対するインタフェース (Human Interface) はコンピュータを使う人間にとって必ずしも快適とはいえない．これらのコンピュータを操作するには，マウスやキーボード，50 ものボタンのついたリモコンなどが頼りとなり，操作ミスやストレスにより，メッセージの伝達がスムーズにいかないことが，しばしば起こる．これは操作盤などのデザインに問題があることが多い[Norman, 1988]が，操作のとき非日常的なアクションを要求したり，ユーザの状況に対応できない固定的なインタフェースとなっていることも一因である．

人間にとって**快適なインタフェース**とはどんなものであろうか？例えば，[末永, 1991]は，“はやい (迅速)，やすい (容易)，うまい (的確)” の条件を満たす必要性を提案している．そのためには，メッセージの入力時には (i) 人間の発するメッセージをすばやくとらえ，(ii) 特別の負担のない，自然な動作，行動，反応の中で発するメッセージを直接とらえ，(iii) 的確な処理を行なうことができ，一方でそのメッセージに対してシステムが，(i) すばやく反応し，(ii) 結果を自然で受け入れやすくかつ認識しやすい方法で返し，さらに (iii) 必要な結果がコンパクトにあたえられることが必要であろう．この条件を満たす 1 つの方法が，我々が日常で会話をする際に用いている，音声やボディランゲージを使ったメッセージ伝達である．すなわち，非接触，無侵襲のためすぐ使えて自由度が高く，日常的であるため自然で負担をかけない，受け入れやすい手段である．

また，人間対人間の日常会話で使われる身ぶり，手振り，表情などは，言葉だけでは表現できない

メッセージを伝えたり，言葉と組み合わせて的確に情報伝達するのに役だっている．例えば，まだ言葉を知らない子供でも，うなずいたり首を振って，「はい，いいえ」のメッセージを相手におくることができるし，視線の方向を調べれば何に興味があるのかもわかる．そこで，コンピュータに言葉だけでなく，その環境，特に人間をみる視覚とその映像を理解する知性を与えれば，マンマシンインタフェースがずっと改善されて，快適な環境でコンピュータを自由に使うことができるようになると考えられる．すなわち，技術的には，コンピュータにテレビカメラを接続してこれをコンピュータの“目”にして，外界の映像を取り込み画像処理を行なって視覚の機能を達成することをめざす．これは20年間以上も画像処理やコンピュータビジョンの研究者がたどってきたシナリオであるが，人間の作業代替手段としてではなく，**人間にやさしいインタフェースを作るために視覚の機能を利用する**ことを考える．すなわち，人間の動作を理解する目の機能を画像処理で実現することによって，人間にとって快適なマンマシンインタフェースを持ったコンピュータを将来作り上げることが可能となるはずである．

ところで，人間の動作を計算機に入力することが目的であるとき，視覚だけにたよることはない．マウスやタブレットを操作すれば，2次元的でしかも作業場所は限られるが，手のある一点の動きを入力することは可能である．最近では光ファイバを使った曲げ検出器と磁気センサによる位置検出器を組合わせたデータグローブ，データスーツなどと呼ばれる人間の3次元的な動作測定装置もある．これらデバイス出力を解析して動作理解をしてマンマシンインタフェースを構築する方法もある．視線や頭部の動きも特殊なヘルメットや特殊な眼鏡をかければ比較的容易に検出することができる．しかしながら，このような特殊な装置を身につけることは，特定の応用分野を除いては歓迎されない．1日中それを身につけているわけにはいかないし，ケーブルによって接続されれば行動範囲が限定されてしまう．人間の視覚が行なうようにテレビカメラなどでとる映像情報からかなりの情報は収集できるはずであり，**非接触，無侵襲なインタフェース**を提供することは非常に大事なことである．また，これらのセンサを装着したり，そこへ手をのばす時間なども省かれ，迅速なインタフェースが提供できる可能性がある．

身ぶり手振りによる人間の動作を理解するには，動き情報を抽出して処理するという動画像処理による動作認識をする必要がある．しかも，インタフェースとして使うためには実時間で動作しなければ使う場所が限られてしまう．一方で，これまでに提案されている画像処理やコンピュータビジョンのアルゴリズムの多くは膨大な計算量を必要とし，実時間での動作が可能なものは非常に限られている．したがって，現時点では，困難な課題を複雑なアルゴリズムによって解決しても実時間での動作が不可能となったり，一方で実時間動作を優先すると単純な機能しか提供できない，という矛盾におちいる．視覚を使ったマンマシンインタフェースの研究はまだ歴史が浅く，どの様な機能を与えれば，どの様にインタフェースが改善されるかを実際に体験できる環境を作らなければならない．たとえ機能は低かったりロバスト性に欠けていても，現時点で実時間動作する環境が必要である．これら観点から，高度な機能を実現する手法を検討する一方で，単純な画像処理の計算手法を組み合わせて実時間動作するシステムを開発しいくつかの実験を行なうというように，両面から視覚によるマンマシンインタフェースの問題の解明に取り組むことが必要である．

本研究は，以上の考えに基づいて，非接触で可能な動画像処理を使って，人間の動きを認識し，その動きからメッセージを抽出することによってインタフェースの向上をはかることを目指して検討する．

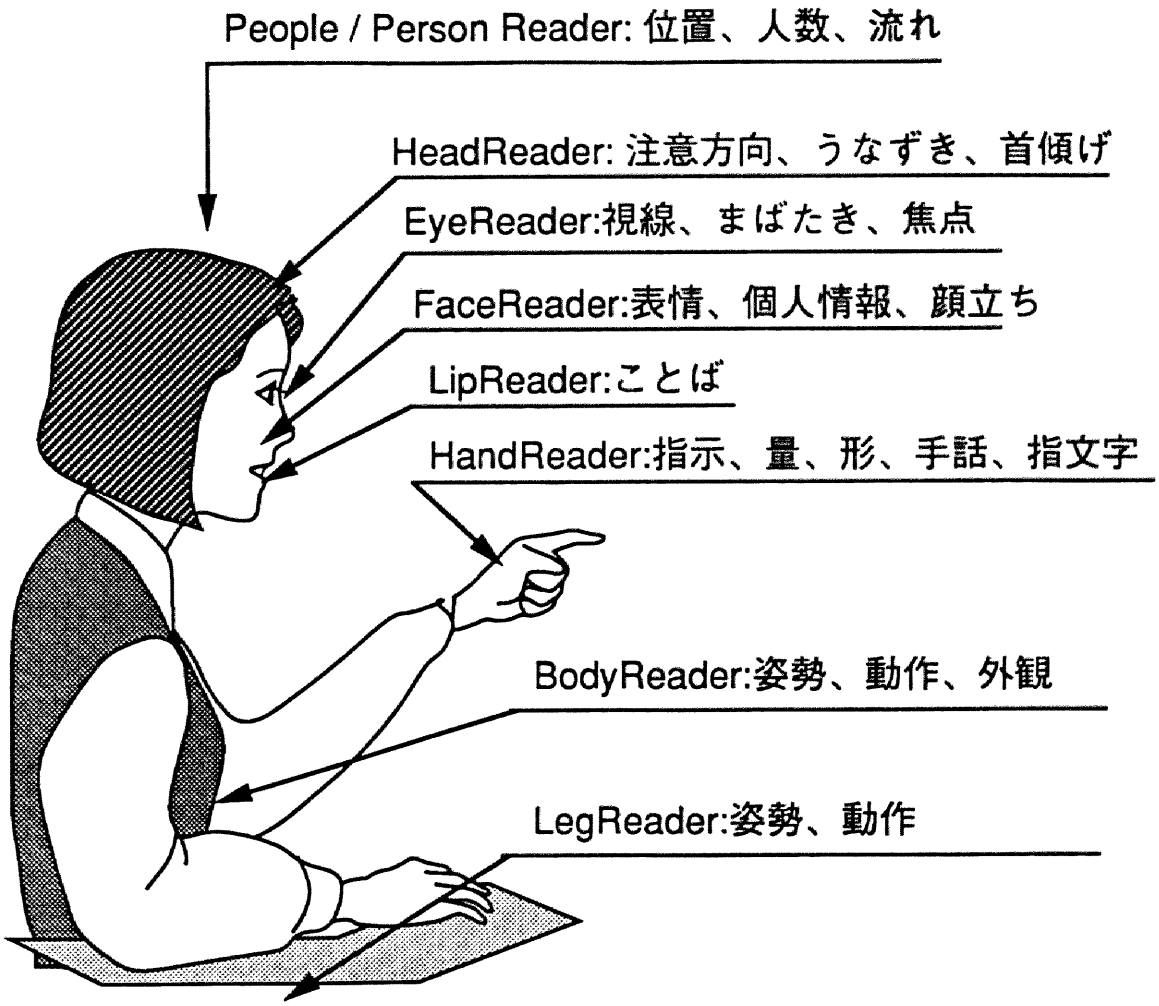


図 1.1: 人間が伝えるメッセージ

1.2 ヒューマンイメージリーダ

人間同士のコミュニケーションでは言語メディアと非言語メディアを使って，メッセージを伝達する．非言語メディアを分類すると，主に動作，距離，外観，声（ことばになっていないもの），におい，接触の6つのメディアがあり，動作をさらに細分すると顔の表情，目，ジェスチャ，姿勢に分類される[本名 信行, 1981, p.237]．我々は言葉や手振りなど複数メディアを使ったマルチチャネルによる会話をしており，チャネル同士が互いに情報を強調しあったり，他チャネルのメッセージの切れ間を補完して，コミュニケーションを円滑にしている．例えば，あるメッセージを伝えるために，表情や音声，ジェスチャなどのいくつかの異なるチャネルを使ってメッセージを強調することがある．時には指示動作のように1つのチャネルだけを使ったりする．したがって，我々は複数のメディアからメッセージを抽出したあとで，それらを統合して相手の意図を解釈して，スムーズなインタフェースを実現している．

同様に機械が人間とスムーズにやりとりするためにも上に列記したメディアに反応してメッセージ抽出を行なう機能を揃えて，さらにそれらを統合する機構が必要である．本論文では視覚を用いた総合的な非言語メディアインタフェースをヒューマンイメージリーダ **(Human Image Reader)**[末永ほか, 1992] [1]とよび，その構築を目指す．ヒューマンイメージリーダは人間の各パーツ対応およびメディア対応にサブシステムを持ち，各サブシステムからの情報を総合してメッセージ解読を行なうシステムである．必要なサブシステムとして，頭，顔，目，口，手（腕），胴，脚の各部に対応する，ヘッドリーダ，フェイスリーダ，アイリーダ，リップリーダ，ハンドリーダ，ボディリーダ，レッグリーダなどが考えられる．また，視覚を用いたインタフェースであるから，知覚可能なメッセージメディアとしてはそれぞれのパーツの動作，距離，外観などがある．例えば，人物像からは人物の在・不在，人数，歩行方向，顔の向き，顔の動き，視線，目の開閉，口の開閉，表情，指さし方向，身ぶり，手振り，誰であるか，もしくは本人であるかどうかの確認，など計り知れない情報が得られる[黒川, 1988]．図1.1はその図解と各パーツが主に伝達するメッセージの例である．また，個々の人間ではなく，群れとして人間の動きを読みとるためのピープルリーダが考えられる．

このようなメッセージのインタフェースは，図1.2に示すように人間が発する信号をとらえるセンサ部 (sensor)，とらえた信号から原メッセージを抽出する解釈部 (interpreter)，およびメッセージを統合してシステム (あるいはアプリケーションプログラム) におくる統合部 (integrator) を持っていると考えることができる．たとえば，頭の動作の場合には，センサ部は映像をとりこみ頭部の回転角などを計算する．解釈部は回転角と頭部の位置から頭部の方向による視線を決定しロケータ (locator) デバイスとしてのメッセージを作成するとともに，回転角が周期的に変化していることから［はい］などのセレクタ (selector) デバイスとしてのメッセージを抽出して統合部におくる．(デバイスの分類については次節で説明する．) さいごに，統合部は，いろいろな解釈部からあがってきたメッセージを統合して，システムが期待しあるいは受け取ることのできるメッセージをつくってシステムに送る．ここで統合部

[1] ヒューマンイメージリーダ (Human Image Reader) は，動作を行なっている人間の画像からメッセージを読みとるシステムという意味を込めて命名された．その語源は口唇の動きを読み取り，言葉を理解するという読唇術 (Lipreading) に由来している．

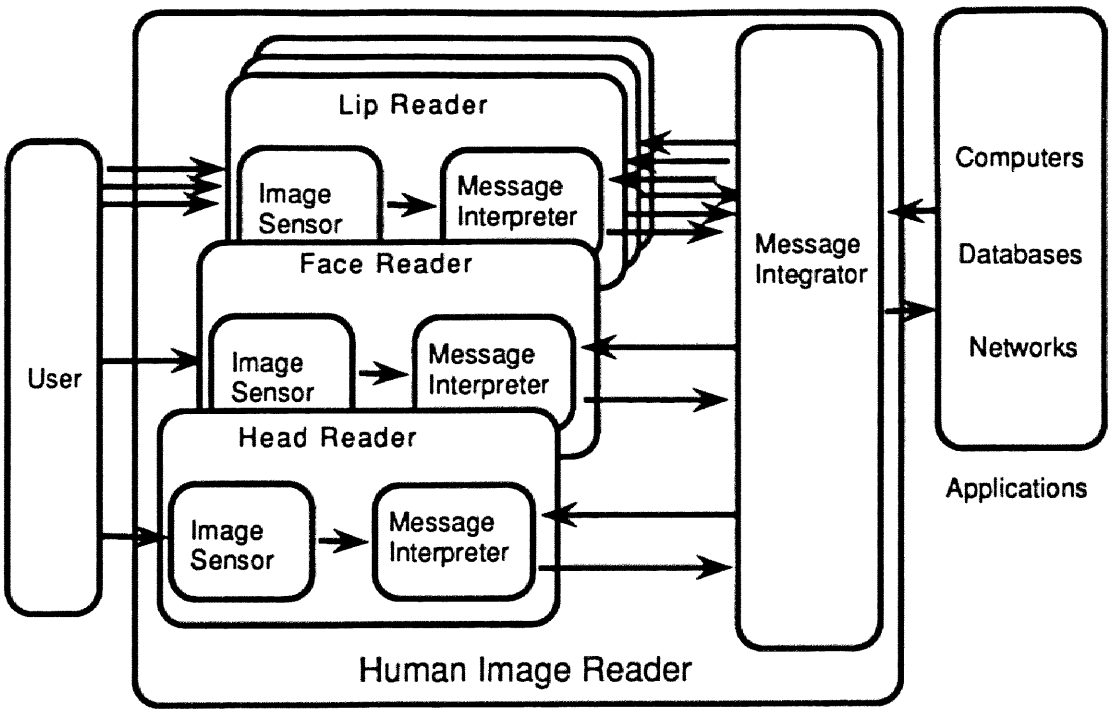


図 1.2: ヒューマンイメージリーダ (Human Image Reader)

はシステムとのインタラクションが必要であり，また，メッセージの的確さをきめる知的な処理が必要となる部分である．

本研究はこのようなマルチメディアをつかったインタフェース (マルチモーダルインタフェース[Bolt, 1987]ともいう) 構築の第 1 歩として，動作メディアからのメッセージの抽出を行なうために，センサ部と解釈部にあたる機能を実現する．なお，対象を動作に限っても，感情を表わす表情や，発声と同時に起こる顔の表情，視線やまばたきなどの目の動き，頭，首，手足，および胴を使ったジェスチャと姿勢など，多くのパーツが関係する．

1.3 従来の研究と関連する研究

インタフェースとは，お互いに相手のことを完全には知りつくしていない人間とコンピュータの間にたつメッセージの通訳者に相当する．言い換えると，インタフェースの機能は「人間 (あるいはコンピュータ) がメッセージとして発する様々な信号を検出・解釈し、コンピュータ (あるいは人間) が利用可能な形態に変換する」ことである．その変換がいかにスムーズで自然かということが、インタフェースの良さを決める．

良いインタフェースの入力側の条件を整理すると，つぎのようになる．

- はやい (quick)— 処理が早い，短いメッセージで十分，メディアが手近
- やすい (cheap,easy)— メッセージにあったメディアの選択自在，手軽，
  メッセージが覚え易い／覚える必要がない

表 1.1: 情報入力用デバイス，機能と，人間のメッセージ発出メディア

デバイス	例	機能	人間からのメッセージ発出メディア, メッセージ例
キーボード (keyboard)	キーボード	文字 (symbol) 入力	音声 (ことばの発声), 指文字, 他
セレクタ (selector)	ボタン, メニュー等	メニュー選択	ことば, 手話, 在不在, 個人特徴, 他
ロケータ (locator)	マウス, タブレット等	位置指定	指による指示, 視線, 頭部回転, 自分の位置, 他
バリュエータ (valuator)	ダイアル等	量的指示	手の回転, 両手による大きさの指示, 音声, 他
サンプラ (sampler)	TV カメラ,A/D 等	データの標本化	ものまね
イメージャ (imager)	特徴抽出器等	概念, 感性入力	手の形, 表情, 姿勢, 外観, 顔立ち, 他

- うまい (delicious,intelligent)— 的確な通訳

1.3.1 人間が送るメッセージとインタフェース

ここで，良いインタフェースを構成するには，人間がどのようなメッセージを送っているか，どんなメッセージが重要で必要性が高いかを，まず知る必要がある．そこで，マウスのように，既存のコンピュータインタフェースのデバイスとして確立しよく分類されたものとの対比で考えることにする．表 1.1は，現在よく用いられる入力デバイス（の一般的名称）とその機能を分類して，さらに各デバイスが送るメッセージを，本来は人間がどのような方法で送っているかを割り当てたものである．

表のうち，最後のイメージャを除くデバイスは，ロケータに対するマウスやバリュエータに対するダイアルのように，それぞれの種類のメッセージを効率よく入力するデバイスが多く開発されている．とくに今日よく使われるマウスは，基本はロケータであるが，ソフトキーボードやプルダウンメニューなどインタフェースプログラムの進歩により，キーボードやセレクタの機能も果たすことができるようになっている．また，人間の動作に密接に関連するものとして，データグローブやデータスーツといった，各部の位置や関節の角度データを入力できるデバイスが市販されるようになってきた．これらのデータも高度に解釈すると，ロケータやバリュエータの機能を果たすことができる[大西ほか, 1990].

これらのメッセージのいくつかは，人間に対しては言葉で伝えたり，指で指示したり，うなずいたり，相手の聴覚や視覚にうったえて伝えることができる．なかでも，顔が果たすことができる役割は多く，視線，うなずき，顔の個人特徴など，それぞれのデバイスに代わってメッセージを伝達できるものがおおい．良いインタフェースは，あるメッセージを伝えるのにいろいろな手段があって，そのときの状況で最も便利なデバイスやメディアを使えるものでなければならない．多くの代替手段を提供するこ

とはけっして無駄とはならない．

人間は場所や個人の名前といったメッセージのほかに，感情や雰囲気，あるいは形などといった，概念や心象に分類できるメッセージを伝達している．このようなメッセージを伝達できる独立したデバイスは，いまのところ存在しない．ここでは，このような**概念や心象にかかわるメッセージを伝えるデバイスをイメージャと呼ぶ**ことにする．顔は上記のようなロケータやセレクタにあたるメッセージより，このイメージャのメッセージを送るのが得意である．また，現存するデバイスがないため，映像で表現される顔をメディアとするイメージャの実現に期待がかかる．

1.3.2 人間からのメッセージ抽出に関する研究

人間からのメッセージ送出には**静的なメディアと動的なメディア**が使われる．静的なメディアには，個人の特徴としての顔の違い，頭の向き，目の向き，腕の位置，姿勢，手の形などがあり，動的なメディアには頭の動作，顔の表情，発声にかかわる口の動き，まばたき，目の動き，腕の動きなどが使われる．パーツごとにどのようなメッセージが抽出できるかを，過去の研究を概観しながら整理する．

頭の向きと動作： 表情や目の動きは頭部全体の動きのなかで決定されるので，頭部の向きと動作は顔のメッセージの基本といえる．顔の識別などのためにも，向きを正規化するときに必要となる．

頭部の方向はロケータとして使われる．頭部の方向を決定するには，目や口などの造作の特徴点を抽出してその位置とあらかじめ測定した特徴点の距離から計算したり[大村ほか, 1989]，動きと造作の配置モデルをつかって回転角を計算する[崔ほか, 1991]．本論文では，造作の抽出が要らない顔と頭の領域の変化から方向を推定する方法を提案する[Akita, 1984][間瀬ほか, 1991b]．両目の対象性から向きや傾きを推定する[角, 太田, 1989]ことも可能であり，最近は頭部の 3 次元データが入手可能となり，その対象軸を求めることで方向を推定する方法[阿部ほか, 1990]もある．

これらのうち，目鼻口などの造作を抽出する必要のある方法では，これらの安定した抽出法が，顔の識別処理のための特徴抽出[Goldstein *et al.*, 1972][Bruce, 1988]として，長い間の研究テーマとなっている．多くの中で現在受け入れられているのは，線画像に対する周辺分布を使う[坂井ほか, 1973]，あるいは色相情報を使う[佐々木ほか, 1991b]などして領域候補を抽出したあと，造作位置や配置に関するヒューリスティックを使う方法[秋本ほか, 1990]である．また，モデルの当てはめを行う方法もあるが初期値の良さに依存するという問題がある[Kass *et al.*, 1987][Yuille *et al.*, 1989]．

頭の連続動作には「はい，いいえ」のような意味をもつものがある．これはセレクタデバイスに相当する．本論文のヘッドリーダ[Mase *et al.*, 1987]がそのようなメッセージを抽出する．

表情： 表情は「楽しい」とか「悲しい」というような言葉に代わるメッセージを伝えるので，そのようなメニューをもったセレクタデバイスとして位置づけることも可能であるが，人間の感情はもっと複雑で，メニューを選ぶように表情を見せるのは難しい．すなわち表情はもっと概念的で，イメージャとなる．本論文では，オプティカルフローからの表情筋の動作を検出[Mase, 1990]して表情からの基本感情[間瀬ほか, 1991a]の推定を行なう．類似の目的で，部品の変形パラメータを記述した報告[崔ほか,

1991]などがある．残念ながら，積極的にインタフェースのデバイスとして利用するまでには至っていない．

口： 顔の表情は感情の表出結果であるが，発声時の調音動作によっても顔の表情は変化する．そこで，その表情の変化をもとに話している言葉をある程度理解することが可能となる．読唇術（Lipreading）と呼ばれる技術がこれで，音声認識と同じくキーボードあるいはセレクタデバイスに相当する．さきに述べた表情変化の抽出法を，口の動きによる発声単語の認識という問題に特化して適用することができる．画像処理による読唇には，口形あるいは口唇輪郭の解析[Petajan and Bodoff, 1988][松岡ほか, 1986][栗田ほか, 1988] [田村ほか, 1989]がある．本論文では，口唇まわりの動きの解析[間瀬, ペントランド, 1990]による方法を示す．

視線とまばたき： 「目は口ほどにものをいう」のことわざがある．視線の機能には，ロケータ[Bolt, 1987][伴野ほか, 1989][青山, 河越, 1988]，まばたきによるセレクタ[畠山, 1989]，目の焦点によるバリュエータ(興味の度合など)またはイメージャ(落ち着き)の機能があろう．重度身障者でも，目の機能を残しているひとにとっては，重要なメッセージ伝達メディアである[2]．視線検出器やまばたきセンサでとったメッセージをつかってワードプロセッサをコントロールすることさえできる[畠山, 1989]．

個人識別： 顔は個人が誰であるかという名前との対応づけを行うセレクタデバイスとしての機能と，どんな顔立ち（例えば，丸い，シワの，穏和な）かというメッセージを送るイメージャの機能がある．個人の名前との対応づけは，インタフェースにおいて必要不可欠である．また，個人の雰囲気をつたえる顔立ちは服装や姿勢とならんで，会話の際の共通基盤を構成するのにおおきな役割を果たしている．しかし，このようなイメージャを実現する研究はまだ行われていない．

画像処理により個人を識別するための研究はすでに20年近い歴史がある．まず顔の造作抽出を行い，その造作の位置関係を特徴パターンとする識別[坂井ほか, 1973]や，造作を使って位置合わせをして濃淡情報で識別する方法[佐々木ほか, 1991a]などがある．

手の動き： 手の位置や指の方向は，「あれは何か？」という時の指さしのように，ロケータとして働く[Bolt, 1980][福本ほか, 1991]．手でダイアルを回す動作のように，バリュエータとしても使える．手をもっとも有効に利用しているのは，指文字を含む手話である[Tamura and Kawasaki, 1986] [高橋, 岸野, 1990]．これらはシンボルをあらわすキーボード，単語や文章をあらわすセレクタ，さらに感情や感性などのイメージャ[Kruger, 1983]としての機能を果たすことができる．

なお手は，メッセージ伝達だけでなくマニピュレーション(manipulation)することもできる．仮想現実感の世界では，上記のメッセージをデータグローブからコマンドとして送るのと同時に，世界のなかの物体をつかんだり投げるといったマニピュレーションの動作を利用している[服部, 1991]．

[2] 重度肢体障害者にとっては目に限らず，機能がはたらく限り，顔のすべてをメッセージ伝達デバイスとして使うことになる．

体全体： 体全体では体型，姿勢(踊りのポーズ)，動作(踊り，活力)などいずれも感性に関わる情報を表現している．わずかに，体の向きと手の動きを使った交通整理などのセレクタに相当する姿勢がある．踊りの姿勢やポーズを記述する研究はあるが[小野, 黒川, 1985]，目的が記述にとどまっており，サンプラに相当するといえよう．

集団： 1人1人の場所を抽出するとロケータとなる．複数の人の位置および動きを知ることによって，人の流れや，行動パターンが察知できる．歩行者の計数もピープルリーダの1つである[間瀬, 1990]．

1.3.3 メッセージの統合

多数のメディアからメッセージが得られるようになると，それらを統合解釈してシステムに伝達する機能が重要になってくる．図1.2における統合部の役割である．これについては十分な研究は進んでいない．Boltは，統合の利点として，"unburdening(気がる)", "summation(情報の集約)", "redundancy(あいまいさの除去)"をあげた．さらにその解釈の仕方として，視線を例にあげ，interest, attention, referenceの3つの場合があることをしめしている[Bolt, 1987]．これは，視線のロケータメッセージを，場面の状況と，音声および動作のメッセージによっては，興味の対象か，注意を払っているのか，あるいは参照の一部かを区別できることを述べている．このような，興味とか注意というメッセージを抽出するには，統合部はかなり知的な機能を備えている必要があり，まだ困難な課題である．

1.3.4 他の分野との関連

この節では本研究の進展に強く関連している3つの分野についてごく簡単に述べておく．

人工知能

上記で述べたように，ヒューマンイメージリーダは人間から機械へのメッセージの通訳者であり，知的な作業が期待される．本論文ではおもに個々のサブシステムからの原メッセージの抽出について述べるが，本当に使いやすいインタフェースは，統合部の知的な能力に関わる．状況を解釈し，どうゆうメッセージをシステムが必要とし，人間が発するメッセージから必要なメッセージをいかに選ぶかという問題を解決するには，システム，ユーザ，習慣，文化などに関する知識を利用する必要があり，人工知能の研究に依存するところがおおきい．

とくに通訳者の作業を3つ(センサ，解釈，統合)のエージェンシーに分割したときに統合部に知的な能力が求められ，この部分のモジュールの詳細化とエージェントの単機能化を進める必要があろう[Minsky, 1985]．

コンピュータグラフィックス

冒頭で述べたようにインタフェースにはメッセージの入力と出力があり，分かりやすい出力も欠かせない．カーソルの表示の仕方や表示速度[間瀬, 末永, 1985]，あるいはシステムの結果のデータをいかに

分かりやすく表示するかというビジュアライゼーションなど，コンピュータグラフィックス技術はインタフェースをよくする大きな要因である[Foley and Van Dam, 1982]．我々人間の姿かたちを有するCG人間[Waters, 1987][間瀬ほか, 1990]として現われ，リアルな音声と豊かな表情や身振り手振りをもって我々に語りかけてくれるとしたら，データだけでなくコンピュータの感性をも伝わり自然な感じのインタフェースを実現できるに違いない．

知的符号化 (モデルベース符号化)

人間の発するメッセージを抽出して，それを伝送し，上記のCG人間にメッセージを再現すると，いわゆるテレビ電話やテレビ会議[安田, 1988]と同様のサービスが可能となる．これまで，動画像の伝送は信号レベルで，ベクトル符号化器や高能率の符号化圧縮を目指してきたが，圧縮率の低減化が限界になっている．そこで冗長性のないメッセージにメディア変換して伝送する知的符号化が検討されている[相澤ほか, 1989]．分析や合成に対象（この場合は人間）に関する知識やモデルを使うので，モデルベース符号化などと呼ばれる．これらのメッセージを直接送るのではなく，信号レベルの符号化能率をさらに上げるために分析合成を利用する方法も検討されている．いまのところ，通信回線の使用料は高価であるが，近い将来光ファイバを用いた低使用料のB-ISDN(高帯域ディジタルネットワーク) が利用できるようになる[寺田, 1991]．そのとき知的符号化は単に高効率符号化のためだけでなく，冗長性の低い，意味をもったメッセージにメディア変換することによって，ネットワークや通信装置がユーザのコミュニケーションに協力して円滑にする必要がある．そのためには，意味をもったメッセージの抽出と使いやすいインタフェースへの利用の検討が重要なのである．

1.4 本論文の概要と構成

本論文は前節のヒューマンイメージリーダの構築をめざして，そのサブシステムのうち頭や顔など1部のパーツの動きを読みとるサブシステムを実現する方法を検討する．具体的には，頭部の動作検出を行なうヘッドリーダ (Head reader) ，顔の表情を読みとるフェースリーダ (Face reader)，口の動きで言葉を認識するリップリーダ (Lipreader，読唇器) とピープルリーダの一つとして歩行者の動きを抽出するための方法とそれを利用した歩行者計数器 (p-counter) の各サブシステムを構築する．

前述したように，実時間性の追求と機能の追求という2つの考え方のどちらかをサブシステムごとに当てはめて検討を行なう．実時間性を追求する際には撮影条件を限定して，簡易なアルゴリズムでシステムを実現し，ワークステーション相当の処理能力で実時間で操作できる環境を作ることを目指す．また，機能を追求する際には，実時間性より機能を重視してアルゴリズムを構築し，シミュレーションやバッチ処理による実験で動作を確認する．

第1章は序論であり，本研究の意義，目的，従来の研究の流れと課題および本論文の構成，について述べている．第2章から第5章までが本研究の中核であり，第6章で結言を述べる．

第2章では頭部の動作を読みとるヘッドリーダの構築をはかる．頭部3次元動き量と「はい，いいえ」などの意味のある動作の認識を行なう．ヘッドリーダは実時間で動作するシステムを目指し，単純な

画像処理手法を適用する．そのため，髪の毛と顔の見え方から頭部動作を抽出するためのモデルをつくり，実時間で取得可能な特徴パラメータから3次元動作量を導く式を導出する．システムの動作を確認する実験をするとともに，ワークステーション上に実現した実験システムを使って，いろいろな応用システムを検討する．この応用システムの試用を通じて視覚によるマンマシンインタフェースの可能性と問題点について議論する．

第3,4章では顔の表情を読みとる2つのサブシステムを検討する．すなわち，感情をあらわす表情そのものを読みとるためのフェイスリーダと，発話動作によって起こる表情を言葉に結びつけるリップリーダを検討する．ここでは，表情の変化を抽出するためにコンピュータビジョンで提案されているオプティカルフローアルゴリズムを適用する．オプティカルフローの計算は現時点では実時間処理には高性能ハードウェアの利用が必須になる．本論文では実時間性は重点におかず，表情や言葉の読み取り機能を実現することを主眼におく．

第3章では表情を読みとるフェイスリーダをトップダウンとボトムアップの2つのアプローチで検討する．まず表情に関する研究を概観し，心理学などで使われる表情の記述法について概説する．また，第3，4章で用いる動きデータのもとになるオプティカルフローのアルゴリズムを解説する．次にトップダウンの認識として，筋肉モデルによる表情の記述をゴールとした手法を検討し実験を行なう．またボトムアップで表情を認識する手法を，古典的パターン認識の手法を応用して構築する．オプティカルフローの統計量を特徴パターンとして，4つの基本的感情を示している表情の識別を行なうシステムを構築し実際の表情動画像の分類識別実験の結果を示す．

第4章は言葉を読みとるリップリーダの検討を行なう．そこではまず計算機による読唇の研究を調べ，解決すべき問題点を明らかにする．次に表情認識で用いたオプティカルフローによる表情筋の動き抽出を，口唇まわりの動き抽出に問題を特化してオプティカルフローに基づく特徴ベクトルを決定する．さらに単語認識実験システムを構築して連続発声した英数字の認識実験を行い結果を示す．

第5章は人のパーツの動作ではなく，人間全体の動きとして流れの抽出を行なうピープルリーダを検討する．本章では，まず時空間画像解析に基づき，物体のフローを推定する方法を提案する．さらにフローの推定法を応用して人物の計数を行なうp(pedestrian)-カウンタを構築する．具体的には，まず時空間中の非エピポーラ面画像の性質をしらべ，直交2断面を使った物体フローの推定法を提案し，実験でその有効性を確認する．さらに直交2断面法を簡単化した直交1断面法で，物体フローの移動方向を判別できることを示し，これが歩行者の方向別計数に応用できることをしめしてp-カウンタシステムを構築する．システムをワークステーション上で実現して，実時間動作実験システムを作成し，歩行者の計数実験の結果を示し本手法が有効であることをしめす．

第6章は，本研究で展開した視覚によるインタフェースを実現する動きメディアからメッセージを抽出するヒューマンイメージリーダのサブシステムの構築とそれによる実験結果について総括し結言とする．

第 2 章

頭部動作の抽出と理解 — ヘッドリーダ —

2.1 はじめに

本章では，意図を伝える動作としてジェスチャーの一部である頭部の動きの検出に注目する．表情や目の動きも頭部全体の動きのなかで決まるものであるから，頭部の動きは動作の基本といえる．ここでは，頭部の動きを抽出し動きの意味を理解するヘッドリーダ (Headreader) と呼ぶサブシステムを構築する．

頭部の動きを画像処理で検出・認識した例には，顔領域の重心等の変化から顔の向きを 2 次元ベクトルの方向として検出[Akita, 1984]する手法が提案されているが，ここでは 3 次元的な動作への変換は行なわれなかった．また，頭の剛体性を仮定して，数点の特徴点の動きを使って動作を計測し，指示入力へ応用[大村ほか, 1989]する方法や，知的符号化のための動き分析に応用するための手法も提案されているが，画像面に平行な動きの抽出に限定した[金子ほか, 1988]り，剛体仮定による動き解析のための特徴点を人手で入力する[相澤ほか, 1989]など，汎用性と自動化が望まれている．

本論文では，高速で比較的安定した，かつ補助手段をまったく用いない自動的な動作検出法として，頭と顔の領域の面積と重心の情報をもとにした頭の 3 次元動作の検出法[間瀬, 末永, 1985][Mase *et al.*, 1987][間瀬ほか, 1988][間瀬ほか, 1989][Mase *et al.*, 1990]を示す．これは Akita らの手法に対し，動きの 3 次元的解釈を追加したこととみなすことができる．本章の概要は以下のとおりである．

まず 2 節では，2 次元画像から得られる特徴量から 3 次元の移動・回転量への変換について詳しく述べる．頭部の簡単な幾何モデルを想定して回転量への変換関数の形を調べ，特別な場合には 1 次変換で近似できることを示す．

3 節では 2 次元特徴である面積及び面積重心を求めるための画像処理手法について説明し，しきい値処理を主体とした具体的なインプリメンテーションについて述べる．汎用ワークステーションを使ったシステムにおける処理時間について検討する．

4 節では，まずコンピュータグラフィックスで作った頭部のモデルを使って動作測定のシミュレーションを行い，変換関数形を確認，修正する．また，実際の頭部動画像から動きを抽出した例を示す．

5 節では，動作認識システムをグラフィックスワークステーションと接続して，動作を認識合成するシステムを構成したので紹介する．

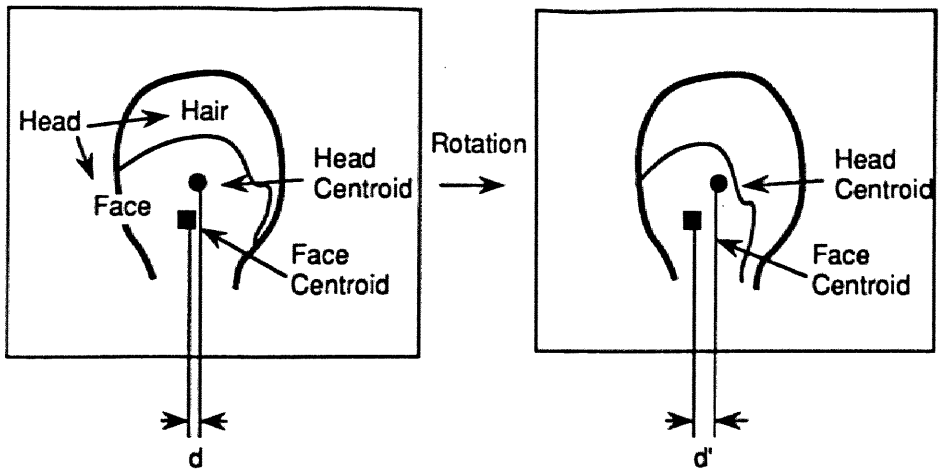


図 2.1: 頭の回転にともなう頭部と顔部の形状と重心の変化

6 節では，頭部動作抽出の応用として，「はい・いいえ」などの意味を認識理解する手法と応用システムを検討する．電子秘書とウィンドウ選択の実験システム例を示して，マンマシンシステムにおける頭部動作認識の有効性を考察する．

2.2 頭部動作検出の原理とアルゴリズム

我々は，ある人が誰なのか判らないほど遠くに立っている場合でも，その人がどちらを向いているのか，おおよそ見当をつけることができる．すなわち目や口を識別せずに顔の向きを知ることができるのである．これは見えている顔の部分を髪の毛がどのように覆っているかという情報にもとづいて認識していると考えることができる．

2.2.1 頭部の投影変換モデル

例えば，透視投影した頭が回転すると，画像平面上で顔の幾何学的特徴が変化することがわかる．頭と顔をそれぞれ領域としてとらえ，各領域の面積重心 (1 次モーメント) を考えると，図 2.1に示すように，回転とともにお互いの重心の位置関係が変化する． Akita[Akita, 1984]はこれに注目して顔の向きを 2 つの重心が作るベクトルで表現した．本節では，これらの重心が与える情報を 3 次元的な回転量に変換することを中心に考える．特に正面の重心の位置を基準にして，変化分を用いて動作量を計算する．また，平行移動については，画像平面上の頭部の重心の動きおよびその領域の面積の変化が 3 次元空間での平行移動に対応していることは，ほぼ自明であるが，具体的な変換式を示す．

図 2.2のような球体による理想的な 3 次元頭部モデルを考える．髪の毛は球の表面に張り付いていると考える．まず，y 軸回りの回転の場合について考える．この頭部モデルの $x-z$ 平面に平行な面 P で切断した断面は図 2.3(a) のようになる． P 上の顔領域の中心角を $2a$ とする．これを画像平面 VP に透視投影すると断面部分の投影像が図に示すようになり，顔領域の重心（ = 投影像の中点 m ）が決まる．このモデルを y 軸回りに θ だけ回転すると，中点の位置がずれる ($m \rightarrow m'$)．このずれによる重心

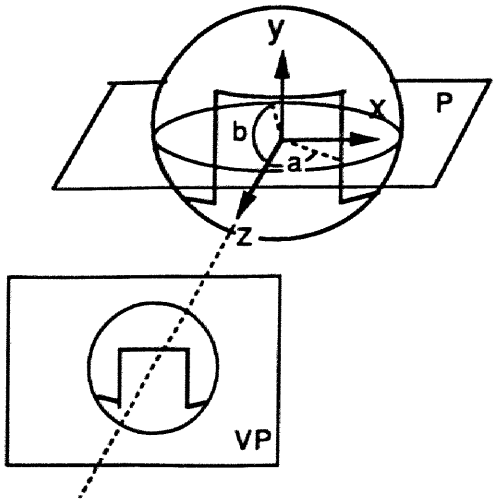


図 2.2: 頭部の 3 次元モデル

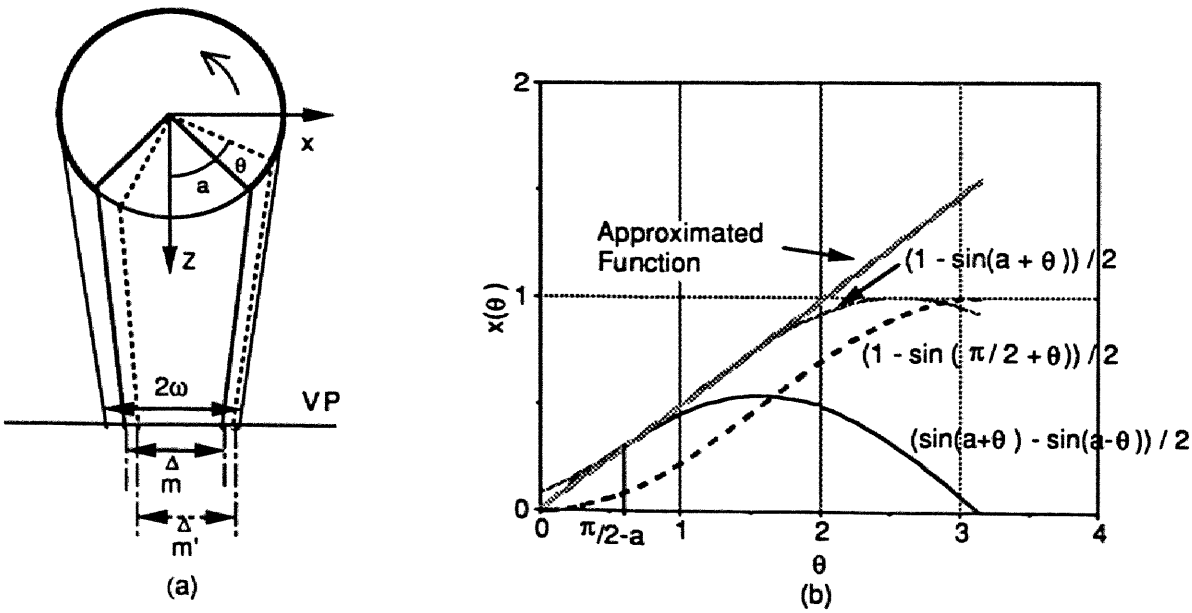


図 2.3: y 軸回りの回転による重心の移動 (a) 水平断面上の顔領域の投影像と振舞い (b) 回転角と重心の変位の関係

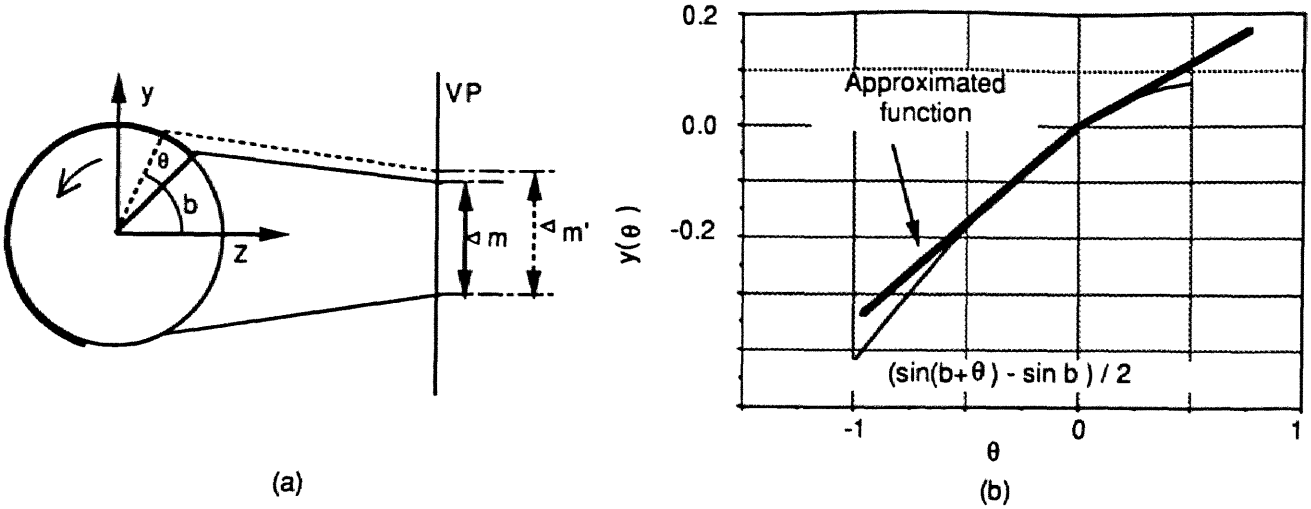


図 2.4: x 軸回りの回転による重心の移動 (a) 垂直断面上の領域の変化 (b) 回転角と重心の変位の関係

座標 $x(\theta)$ は，回転角 θ に対して，

$$x(\theta) \approx \begin{cases} w\frac{\sin(a+\theta)-\sin(a-\theta)}{2} & \text{if } 0 \le \theta < (\frac{\pi}{2}-a) \\ w\frac{1-\sin(a-\theta)}{2} & \text{if } (\frac{\pi}{2}-a) \le \theta < \frac{\pi}{2} \end{cases} \tag{2.1}$$

となる．ただし $2w$ は頭部全体の投影長である．

式の導出には，|頭部大きさ / 物体と画像平面の距離| $\ll 1$ と仮定して，中心角 a のときの顔部（端）の投影位置 x が頭部の投影位置 w に対して，$x \approx w\sin(a)$ となることを利用した．ただし $a+\theta$ が $\frac{\pi}{2}$ を越えるところでは顔の右端は θ にかかわらず頭の右端に一致し一定となる．

$x(\theta)$ をプロットすると，図 2.3(b) のように，a が小さいかあるいは θ が小さいところでは線形近似可能であることが判る．従ってこのモデルでは，回転角は重心の変位にほぼ比例し，投影長に反比例すると言える．θ が負の場合も同様に求まる．ただし，a が大きくなって $\frac{\pi}{2}$ に近づくと波形は 3 角関数に近づく（図 2.3(b) の点線のグラフは $a = \frac{\pi}{2}$ のときを表す）．

x 軸周りの回転は同じモデルを使うと，図 2.4(a) のようになり，

$$y(\theta) \approx w\frac{\sin(b+\theta)-\sin(b)}{2} \tag{2.2}$$

となる．b は垂直断面における髪の位置の仰角である．このとき $\theta > 0$，$\theta < 0$ の場合を別々に考え，それぞれ近似的に直線をあてはめることができる (図 2.4(b))．

実際は，面積重心を使うので，これらの変化を全ての断面について加重平均することになる．すなわち，顔部の開口角 a および b は各断面で一定でなく，また個人によっても異なる．しかし，θ があまり大きくない範囲については，全体としても，回転角と重心の移動量は近似的に比例関係にあるということができる．ただし，断面の多くで a が $\frac{\pi}{2}$ に近い場合には，精度良く比例関係が成立する θ の範囲はごくわずかである．

2.2.2 回転・移動パラメータの抽出

以上の議論から，重心および領域面積の変化分による，正面位置に対する3次元の回転量 (R_x, R_y) および平行移動量 (T_x, T_y, T_z) への変換式は以下のようになる．

$$R_x(t) = g_{rx}\left(\left(\Delta y_h(t) - \Delta y_f(t)\right)/\sqrt{S_h(t)}\right) \tag{2.3}$$

$$R_y(t) = g_{ry}\left(\left(\Delta x_h(t) - \Delta x_f(t)\right)/\sqrt{S_h(t)}\right) \tag{2.4}$$

$$T_x(t) = g_{tx}\left(\Delta x_h(t)/\sqrt{S_h(t)}\right) \tag{2.5}$$

$$T_y(t) = g_{ty}\left(\Delta y_h(t)/\sqrt{S_h(t)}\right) \tag{2.6}$$

$$T_z(t) = g_{tz}\left(\sqrt{\Delta S_h(t)/S_h(0)}\right) \tag{2.7}$$

ここで $(x_h(t), y_h(t)), (x_f(t), y_f(t))$ はそれぞれ時刻 t の頭部および顔部の，画像座標系における面積重心の位置であり，Δ は時刻 $t = 0$ に対する変分を示す（透視変換を考慮して，例えば，$\Delta x_h(t) = x_h(t) - x_h(0)\sqrt{\frac{S_h(t)}{S_h(o)}}$ のように求める）．また，$S_h(t)$ は頭部の面積で，その平方根 $\sqrt{S_h(t)}$ が投影長 w に比例すると仮定している．なお，各軸周りの回転角の計算にそれと直交する軸方向の座標変化（x 軸に対して y 座標）を使っていることに注意していただきたい．$\{g\}$ は変換のためのマッピング関数で，回転については前述の議論から，理想的な場合は線形1次関数となる．ただし，式(2.3)の g_{rx} は変数の符号で傾きが異なる．式(2.5),(2.6),(2.7) は透視変換の関係から明らかで，g_{tx}, g_{ty}, g_{tz} は線形1次関数である．また，一般には式(2.3),(2.4)の g は1次の逆3角関数で表現される[1]．

首の中点の位置が精度よく求められる場合には，次式を使って z 軸回りの回転も計算することができる．

$$R_z(t) = tan^{-1}\frac{x_h(t) - x_n(t)}{y_h(t) - y_n(t)} - tan^{-1}\frac{x_h(0) - x_n(0)}{y_h(0) - y_n(0)} \tag{2.8}$$

ここで，$(x_n(t), y_n(t))$ は首の中点の座標で，これらを図示すると図2.5(a)のようになる．このとき $R_x(t)$ および $R_y(t)$ の計算には式(2.3,2.4)の代わりに次式，

$$R_x(t) = g_{rx}\left(\left(\Delta y'_h(t) - \Delta y'_f(t)\right)/\sqrt{S_h(t)}\right) \tag{2.9}$$

$$R_y(t) = g_{ry}\left(\left(\Delta x'_h(t) - \Delta x'_f(t)\right)/\sqrt{S_h(t)}\right) \tag{2.10}$$

のように画像上の頭部座標系の重心の変化分を用いる．すなわち，適当な点の周りに $R_z(t)$ だけ回転した座標系における各重心位置 (x'_h, y'_h) および (x'_f, y'_f) を使って $\Delta x'_h(t) = x'_h(t) - x_h(0)\sqrt{\frac{S_h(t)}{S_h(0)}}$ のように計算する．ここで，
$\Delta x'_h(t) - \Delta x'_f(t) = (x'_h(t) - x'_f(t)) - (x_h(0) - x_f(0))\sqrt{\frac{S_h(t)}{S_h(0)}}$ なので回転の中心は各時刻について任意に計算することができる．このようすを図2.5(b)に示す．

[1] 厳密には $\Delta\theta = \theta_t - \theta_0 = \arcsin(x_f(t) - x_h(t)) - \arcsin(x_f(0) - x_h(0))$ となる．公式 $\arcsin a + \arcsin b = \arcsin(a\sqrt{1-b^2} + b\sqrt{1-a^2})$ から a, b が十分小さいとき以外は変数は単純な和とはならない．しかし個人についてマッピング関数をテーブル化できるなら，関数の厳密性は特に重要とはならない．

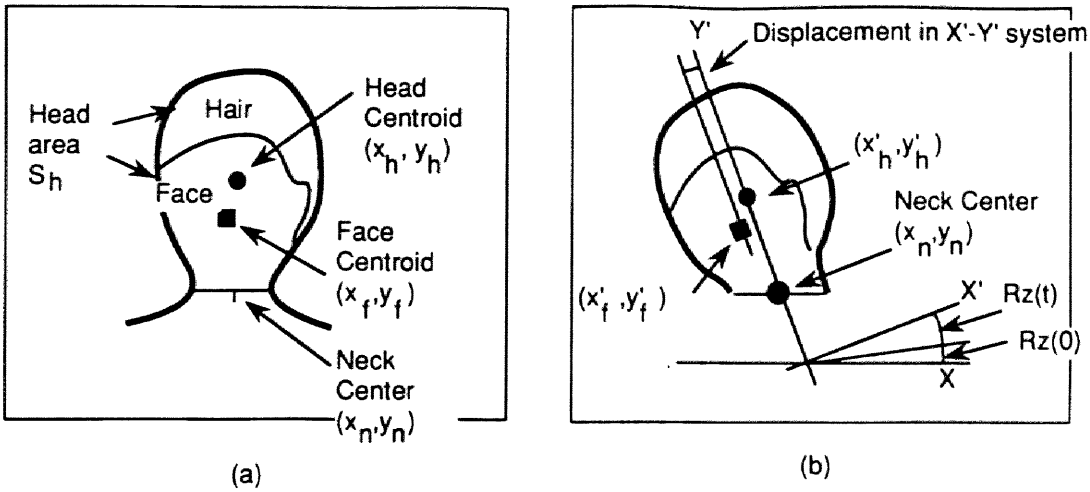


図 2.5: 動き抽出のための特徴量 (a) 特徴パラメータ (b)z 軸回りの回転と物体座標系

2.3 画像処理アルゴリズムとインプリメンテーション

前述の変換に必要な各重心位置を求めるための画像処理アルゴリズムを説明する（図 2.6参照）．原理的には，適当なしきい値を使った 2 値化処理による領域抽出を行う．処理を簡単化するため，ここで，以下の 5 つの条件を定める．

1. 撮影条件（照明，レンズ焦点距離，カメラ位置，方向）は一定である．
2. 人の映っていない背景画像を 1 枚用いる．また背景は変化しない．
3. 頭部の動画像として，肩から上の画像が常に得られる．
4. 髪と膚の輝度にコントラストがある．
5. 顔の回転に伴って，頭髪が顔の領域の形状を変化させる．

これらの条件のうち，1–3 はよく用いられる仮定である．条件 4,5 は領域の重心をつかうこの手法特有の制限となるが，一方で眼鏡の有無や目まで伸びた髪については制限がなく，問題なく処理できるという利点がある．

2.3.1 頭部領域の抽出

まず，頭部の抽出には，背景画像 $B(x,y)$ と処理画像 $f_t(x,y)$ の差分をとってシルエット画像 $s_t(x,y)$:

$$s_t(x,y) = |f_t(x,y) - B(x,y)| \tag{2.11}$$

を作る．ただし t は時刻を示す．つぎにこのシルエット画像を走査し，しきい値処理により頭部の領域 $H(t)$ を抽出する：

$$H(t) \equiv \{(x,y)|s_t(x,y) > T_h\}. \tag{2.12}$$

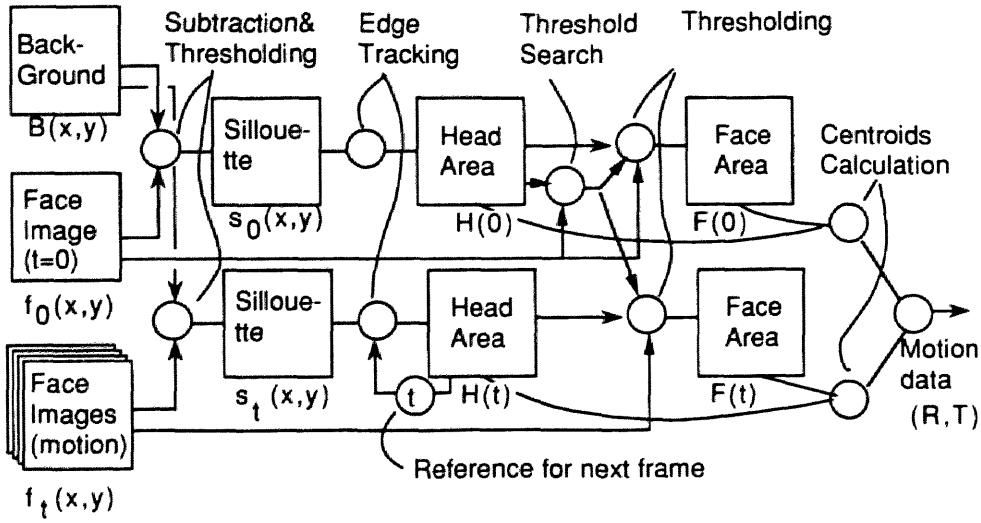


図 2.6: ヘッドリーダの画像処理ブロック図

ここで T_h は，あらかじめ与えられた頭部領域の抽出のしきい値である．条件 1,2 からは原理的に $T_h > 0$ さえ満たせばよいが，実際には TV カメラのオートアイリスや，雑音，照明のちらつきなどに応じて撮影条件は完全に一定とはならないので，適当な値を設定する必要がある．

このシルエット化と頭部領域の抽出は画面の上部から 1 走査線ずつ同時に行い，画面の両側からエッジとなる画素をさがし，最長のランをもとめる．M 画素だけ離れた 2 つの画素の組を順次調べ，両方の画素が頭部領域の条件式を満たすとき，すなわち $s_t(x,y) > T_h$ かつ $s_t(x+M,y) > T_h$ のとき，その外側の画素を頭部領域の一方のエッジ画素とする．これを画面の両側からそれぞれ探索すると，左右のエッジ画素が決まり，それらを端点とするラン－セグメントが求まる．M は通常 2 から 5 である．プログラム上定義される頭部の領域 $\hat{H}(t)$ は，首の位置までのラン－セグメントの集合であり，式 (2.12) で定義した H_t と多少異なる．首の位置は，頭頂から十分離れた範囲で，ラン－セグメントの長さが極小になるかゼロになる位置とする．

頭部の各特徴，面積および重心は，

$$S_h(t) = \sum_{(x,y)\in\hat{H}(t)} 1, \tag{2.13}$$

$$x_h(t) = \frac{1}{S_h(t)} \sum_{(x,y)\in\hat{H}(t)} x, \tag{2.14}$$

$$y_h(t) = \frac{1}{S_h(t)} \sum_{(x,y)\in\hat{H}(t)} y \tag{2.15}$$

となる．

なお，$t = 0$ のときの画像を正面画像 $f_0(x,y)$ としてとりこんで，顔領域の抽出に用いるしきい値 (T_f) の設定と，正面の方向の基準となる $S_h(0), x_h(0), y_h(0)$ を決定する．抽出処理の手順は同様である．しきい値 T_f は，$\hat{H}(0)$ の濃度ヒストグラムをとり，判別分析法[Otsu, 1979]によって自動設定

する．

2.3.2 顔領域の抽出

顔もしきい値処理によって抽出する．顔領域 $F(t)$ を次のように定義する．

$$F(t) \equiv \{(x,y)|f_t(x,y) > T_f, (x,y) \in \hat{H}(t)\} \tag{2.16}$$

（ただし，髪より顔の輝度が高い場合．逆の時は符号の向きは反対．）顔部の面積および重心は，

$$S_f(t) = \sum_{(x,y)\in F(t)} 1, \tag{2.17}$$

$$x_f(t) = \frac{1}{S_f(t)} \sum_{(x,y)\in F(t)} x, \tag{2.18}$$

$$y_f(t) = \frac{1}{S_f(t)} \sum_{(x,y)\in F(t)} y \tag{2.19}$$

となる．

通常の室内照明では，顔がうつむくにつれて，顔の部分は暗くなってしまう．動き量の検出精度の一貫性がとくに重要ではなく，とにかく下をむいたことを知りたい場合は，検出された顔部の面積が異常に小さいときに，しきい値 T_f を下げて再度顔部の抽出を行うことも考えられる．

2.3.3 インプリメンテーション

以上の特徴抽出アルゴリズムと，動き計算アルゴリズムをワークステーション (SUN4/260) 上に C 言語で実現した．画像入力装置 (NEXUS) を使って，128 × 120 画素（各 8bit ）の白黒画像を取り込み，画像を主記憶上に持ってきたのち，頭部領域の抽出，顔部領域の抽出，各領域の面積および重心計算，首の中点決定を行い，2.2.2 節に従って回転角および移動量を計算した．図 2.7は頭部と顔部の抽出結果である．

図 2.6に示したように，まず背景画像をとりこみ記憶しておく．これは以降の処理で頭部領域の抽出に使われる．次に，移動回転の基準となる正面像を 1 枚取り込む．これが $f_0(x,y)$ として使われる．以上の準備の後に，頭部の動きを連続画像で取り込み，処理を行う．図 2.8は，SUN のウインドウ上の操作メニューである．

処理速度は約 0.14 ～ 0.16 秒 / フレームとなり毎秒最大 7 フレーム処理できた．このうち画像の取り込み (画像メモリーにフリーズして主記憶に送る処理) に 0.12 秒前後かかっており，実際の画像処理にかかる時間ははほぼフレーム周期と同じ程度であることがわかった．ところで，処理速度と処理結果の確かさは一見無関係のように見えるが，動画像処理の場合はフレーム間の連続性を利用することが可能である．この例でいえば，頭部の抽出のための探索領域として，前フレームの処理結果の少し広い範囲だけを考えると，処理速度を上げることができ，さらにフレーム間の連続性を強められるという，好循環になる．プログラムにはこのような工夫もこらした．

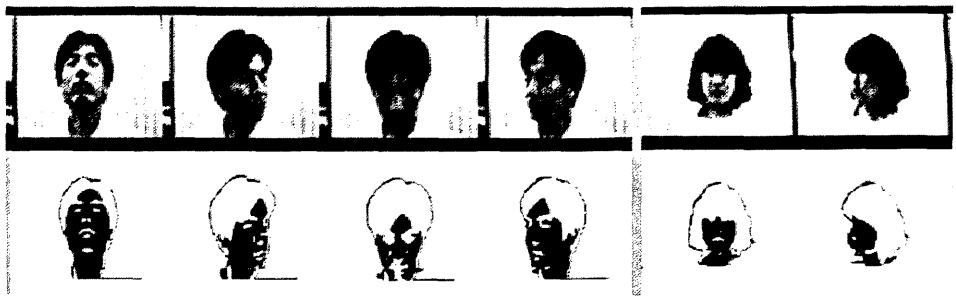


図 2.7: 画像処理結果（頭部輪郭と顔部領域の抽出）

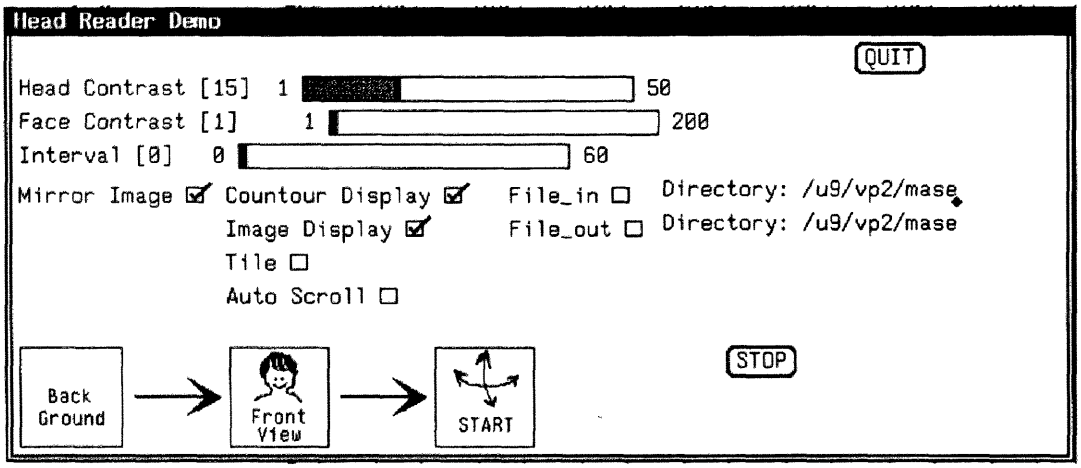


図 2.8: ヘッドリーダの操作メニュー

なお，プログラミングにあたってはシステムの互換性を上げるために，処理時間のかかる画像処理の部分も C 言語で作成した．実際この処理プログラムは当初 VAX11/780(VMS/C) で開発し[間瀬, 末永, 1985; Mase *et al.*, 1987]，その後，パーソナルコンピュータ (NEC-PC9801VM/MS-C) に移植[間瀬ほか, 1988]後，現在に至っている[間瀬ほか, 1989; Mase *et al.*, 1990; 間瀬ほか, 1991b] [2]．それぞれの環境では，画像入力部が異なっており，移植のための修正が画像入力部とコマンドインタフェースに必要であったが，その他はほとんど必要なかった．また，この間に計算機の CPU 能力も向上し，労せず処理速度の向上を図ることが可能であった．画像処理部については専用のハードウエアを利用することも考えられるが，ハードウエアに特化したプログラム部分は数年のうちに陳腐化してしまうので，この様な戦略は重要である．

[2] さらに現在 IRIS ワークステーションに移植中である

図 2.9: CG による顔画像の例

2.4 動作検出実験と考察

まず，コンピュータ・グラフィックス（CG）で作った顔画像をつかって，画像処理アルゴリズムと動作抽出アルゴリズムの動作を確認する．この CG モデルは人物の頭を計測して作った顔形状に渡部ら[Watanabe and Suenaga, 1989]の方法で髪の毛を 1 本ずつ発生させて作ったものであり，幾何学的には実際の人物と同じ形状であると考えることができる．図 2.9はその CG の顔画像である．この画像を上下，左右に回転させ重心位置の変化をプロットしたものが図 2.10である．図 2.10(a) が x 軸回り，図 2.10(b) が y 軸回りの回転のときの，CG の頭の実際の回転角に対する，推定回転角度である．

これらを見ると，図 2.10(a) からは，x 軸周りの回転角が直線で近似できることが分かる．また，この CG 人物像の場合，図 2.10(b) からは，y 軸周りの回転角 R_y は，直線では精度よく近似できないが，単調増加傾向を確認できる．これによって，x 軸周りと同様に線形変換を仮定して角度を推定しても，しきい値処理などにより，左右，上下などの単純な動きの有無を判別することができる．

なおこのように耳に髪をかけたようなヘアスタイルでは，髪の開き角度 a が多くのスライスで $\frac{\pi}{2}$ に近くなり，θ の大きい範囲にわたってまで，精度よく一次近似することは困難である．このため，回転角の高精度推定を行う際には，変換テーブルを作成するか，別の関数形を導入する必要がある．これらは今後の課題である．

次に図 2.10(b) の R_x の値 (図中，Rx(uncalibrated)) に注目すると，回転角の絶対値が大きくなるにつれて，値が変化している．これは頭部の髪型が上下対象ではないため起こる．すなわち，式 (2.9),(2.10) のようには各軸周りの回転の計算が独立になっていない．従って，この変化を打ち消すような補正項が必要である．式 (2.3) または式 (2.9) に代わる，新しい，$R_x(t)$ の計算式は，

$$R_x(t) = g_{rx}\left((\Delta y'_h(t) - \Delta y'_f(t))/\sqrt{S_h(t)} \right) + c_{ry}\, R_y(t) \tag{2.20}$$

となる．補正関数 c_{ry} は髪型に依存する．図には $c_{ry} = 0.4$ で補正した $R_x(t)$ を示した．より完全な補

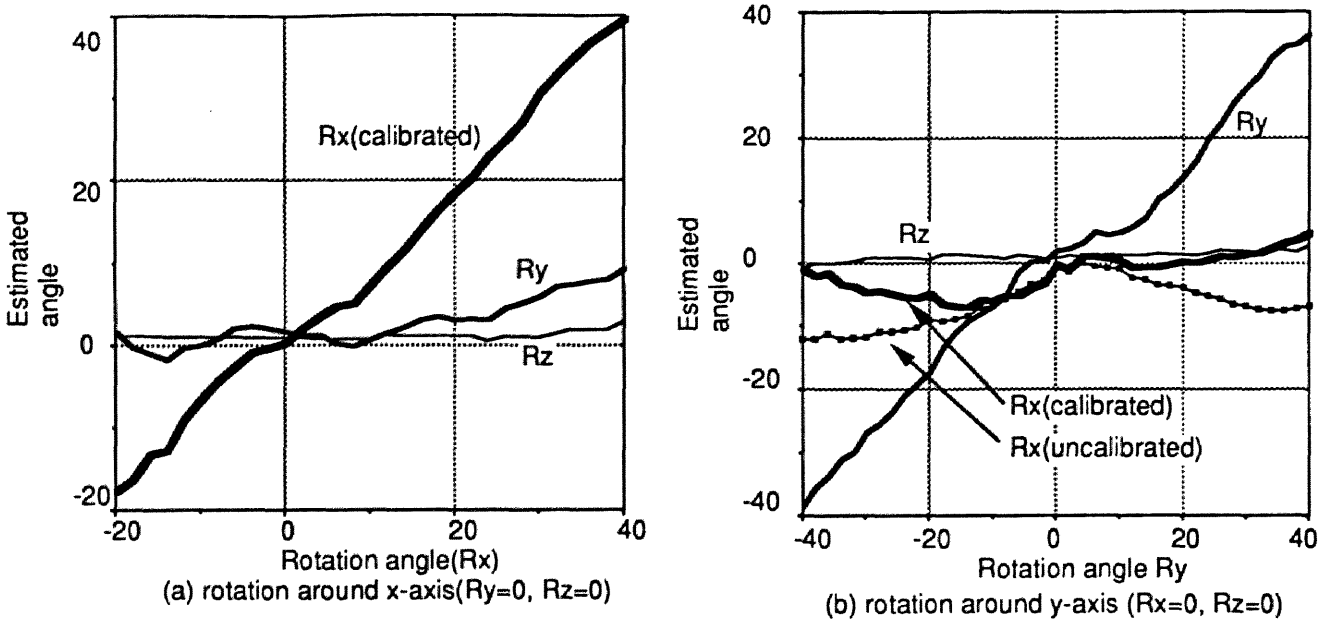


図 2.10: CG 顔画像に対する回転量の推定結果 (a) x 軸回り (b) y 軸回り

正をするためには，補正なしの $R_x(t)$ を使って，補正テーブルを作る必要がある．ここにおいて，各軸周りの回転角の計算は，R_z, R_y, R_x の順番に行う必要があることが分かる．

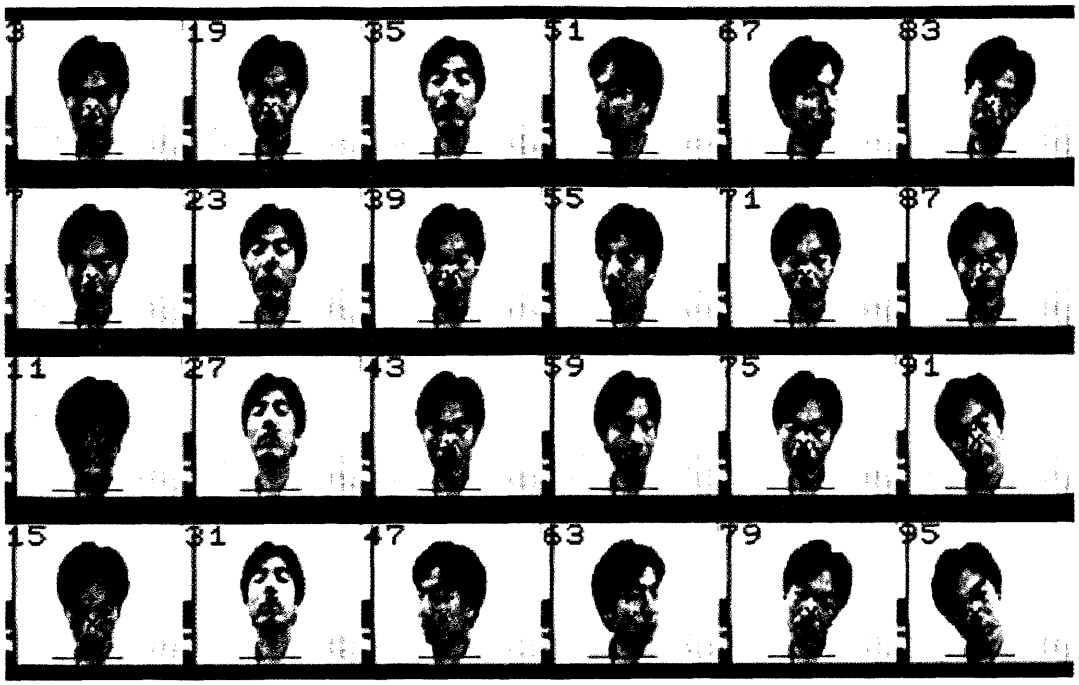
図 2.11は実際の画像で動きを抽出した例である．入力画像と画像処理による重心抽出結果（同図 (a)）および回転角の変化をグラフにしたもの（同図 (b)）である．良好に動きが抽出できている．ここでは実際の回転角度を測定していないので，推定精度については，不明であるが首を振る動作をうまくとらえているといえる．なお，動きの計算には補正項をもつ式 (2.20) をもちいた．

2.5 リアルタイム頭部動作認識合成実験システム

この動作認識システムを別のグラフィックスワークステーション (GWS，シリコングラフィックス社製 IRIS-4D/70GT) と接続し，頭部動作の認識合成の実験を行った．図 2.12が装置の構成図と外観である．簡易テレビ電話の TV カメラとモニタの機能を利用して顔を撮影し，検出した動き情報を GWS に送って，3 次元モデルとして持っている頭部データを回転移動させて，実時間で，原画像の動作を CG のアニメーションとして再現するシステムである．頭部の 3D モデルは約 600 個の三角パッチで構成され，GWS で実時間表示が可能である．認識 WS(SUN4) と合成 GWS はイーサネット (最大 10Mbit/sec) で接続して 3 次元の座標変換パラメータを送っている．

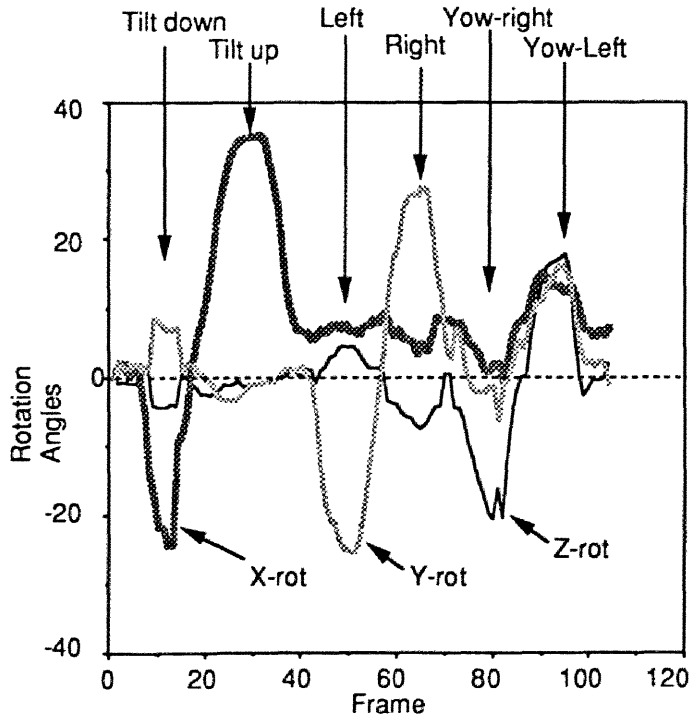
図 2.13はこの実験システムのプログラムの構成を示す．このシステムを動作させるには，まずデータ受信プログラムと表示プログラムを起動してデータ入力待ちの状態にする．その後認識プログラムを起動して，データ送信プログラムを介してデータの送信を行う．データ送信プログラムは無手順で，順次，文字コード化された数値データを送る．数値データのフォーマットは図 2.14の通りである．

このシステムは，頭部動作の認識結果を視覚的に確認するためのシミュレータとして作成したが，同時に将来の知的画像通信システムの一形態を示している．ネットワーク上で通信しているデータは 3 次



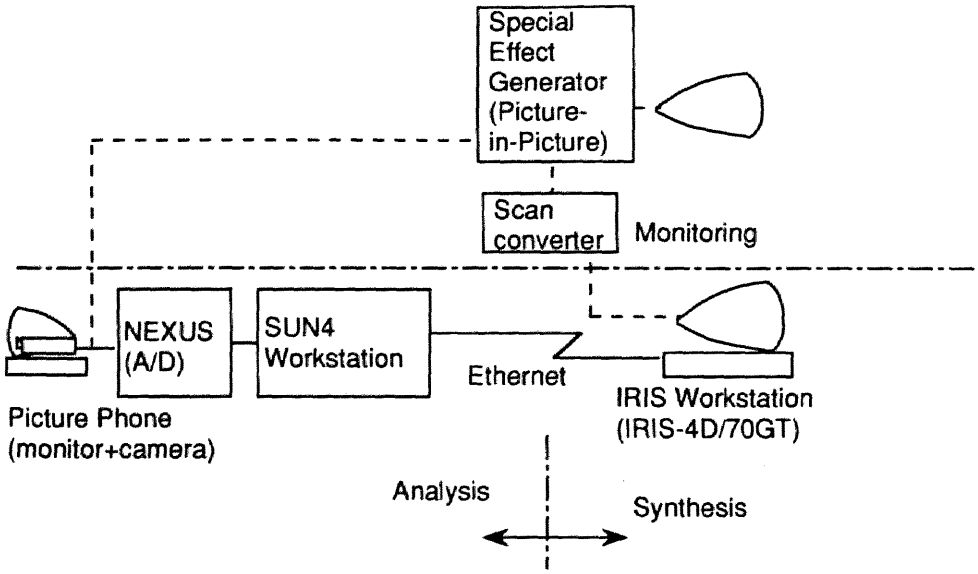
+ : Head Center　　x : Face Center

(a) 原画像の一部と重心抽出結果 (数字はフレーム番号)

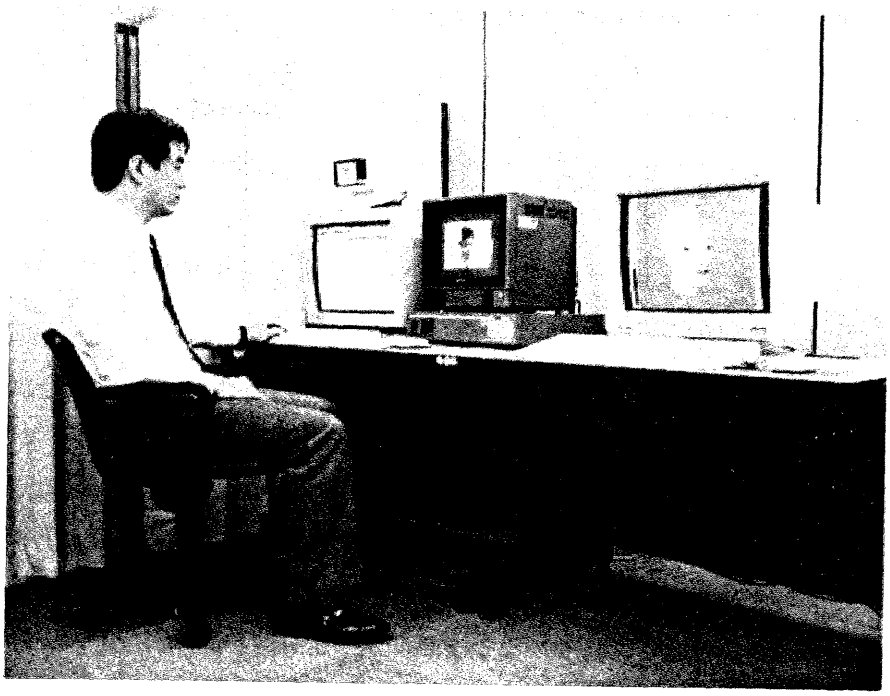


(b) 計測された回転角

図 2.11: 実画像を使った動きの計測



(a) システムブロック図



(b) 実験装置外観

図 2.12: リアルタイム頭部動作認識合成実験システム

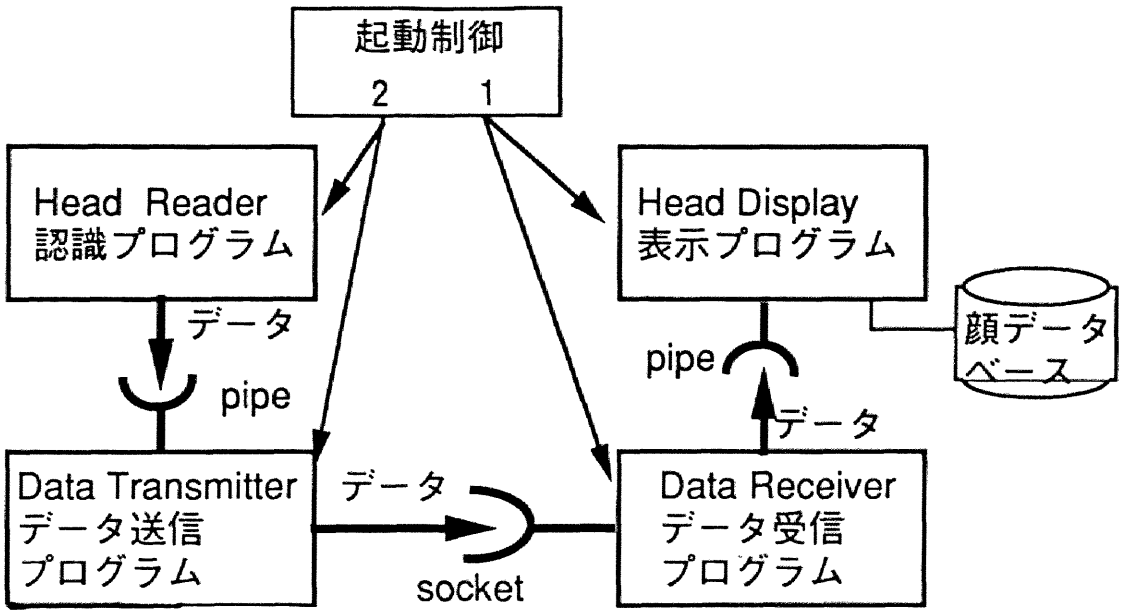


図 2.13: 認識合成実験システムのプログラム構成

```
識別子
 ↓
R Rx, Ry, Rz, Tx, Ty, Tz <cr>
  +---------+   +---------+
    回転角        平行移動
  (各 4byte)    (各 4byte)
```

図 2.14: 数値データのフォーマット

元座標変換のデータのみである．データの通信には特別な符号化を行わず，前述の通り 3 次元の座標変換パラメータを文字列で送信した場合，実測で約 2,500bps であった．将来はこれに表情を記述するデータを併せて送ることになろう．

また，現在は合成側の処理能力が低いので，図 2.12(b) に示すような顔画像が実時間で表示する限度であるが，将来ハードウェアの進歩により，髪の毛や皮膚のテクスチャをもった顔を表示できるようになると考えられる．なお図 2.15は，テクスチャマッピングを高速に行うハードウェアを備えた GWS で，頭部を表示した例である．このデータはカラー情報とレンジ情報を同時に入力できる回転式のレンジファインダ[末永, 渡部, 1990]を使って入力したものである．このように，受信側に頭部の形状および色データがあれば，任意の方向を向いた頭部を表示することはほぼ実時間でできる状況にある．

2.6 頭部動作の理解と応用

2.2 から 2.4 節で構成した頭部動作認識システムを使って，頭部動作抽出処理のマンマシンインタフェースへの具体的応用を検討する．非言語メッセージとしての頭部の動作には，注視（関心）によるロ

図 2.15: ハードウェアテクスチャマッピングを使った頭部の表示例

ケータの機能をはたすものと，「はい・いいえ」などの簡単なメッセージを送出するものなどがある．以下，メッセージの抽出手法と，応用として電子秘書とウィンドウ選択の実験システム例を示して，マンマシンシステムにおける頭部動作認識の有効性を考察する．

2.6.1 コマンド認識

視方向の識別は頭部の回転角を適当に量子化して行う．カーソル位置をコントロールしたり，運動視差をつけるために頭部の動きを利用するときは，連続に変化する回転角が必要であるが，ディスプレイの“右”を見ているとか“下”を見ているというようなことの判定は回転角 θ の

$$\theta > th^{+}, th^{+} > \theta > th^{-}, th^{-} > \theta$$

のような 3 値化により可能である．

また「はい・いいえ」のメッセージは x,y 軸回りの回転量の変化で判断することができる．例えば，それぞれの首振り動作を行ったときの回転量の変化の代表的なパターンを学習して，DP マッチングなどを行うことが考えられる．ここでは，処理の簡素化のため回転量の局所統計量を使って，「はい・いいえ」の動作の有無を判定した．すなわち，次式に示すように，過去数フレーム (実験では 6 フレーム，1 秒弱) にわたる回転角とその時間微分係数の分散と平均を用いた．

$$mean_{rx}(t) = \frac{1}{t_c}\sum_{T=t-t_c}^{t} R_x(T), \tag{2.21}$$

$$var_{rx}(t) = \frac{1}{t_c}\sum_{T=t-t_c}^{t} (mean_x(t) - R_x(T))^2, \tag{2.22}$$

$$mean_{\Delta rx}(t) = \frac{1}{t_c}\sum_{T=t-t_c}^{t} (R_x(T) - R_x(T-1)), \tag{2.23}$$

$$var_{\Delta rx}(t) = \frac{1}{t_c}\sum_{T=t-t_c}^{t}(mean_{\Delta rx}(t) - (R_x(T) - R_x(T-1)))^2 \quad (2.24)$$

ここで，t_c は統計量計算対象のフレーム数である．また，y 軸回りの統計量も同様に計算する．

さらに次式を用いて「はい・いいえ」の有無を判定する．

$$if(var_{rx} > th_1 \wedge var_{\Delta rx} > th_2 \wedge mean_{rx} < th_{down} \wedge mean_{rx} > th_{up}) \Rightarrow \text{はい} \quad (2.25)$$

$$if(var_{ry} > th_3 \wedge var_{\Delta ry} > th_4 \wedge mean_{ry} < th_{right} \wedge mean_{ry} > th_{left}) \Rightarrow \text{いいえ} \quad (2.26)$$

次には，これらのコマンド識別を使ったアプリケーションのプロトタイプとなる実験システムを紹介する．

2.6.2 アプリケーション事例

(a) ウィンドウ選択

背景： ワークステーションのモニタ上に複数のウィンドウを開いて作業をしていると，たびたびマウスに手を伸ばして，作業ウィンドウを切り替える作業をしなければならないときがある．キーボードのブラインドタッチができる人にとっては，せっかくのホームポジションから手を動かして，また，その場所を確かめて戻さないといけないことになり非常に億劫になる．こんなとき，作業しようとしているウィンドウに，自動的にそちらにタスクが切り替わったら便利だと思うことがある．実際，これと同じ発想で，キーボードの中央にジョイスティックを立てて，ホームポジションから移動せずにカーソルを操作しようとする研究報告もある．

実験システムの基本的考え方： この作業ウィンドウを切り替えたいというメッセージは，キーボードやマウスやジョイスティックによるか，あるいは，音声か視線の方向で伝えることが可能である．そこで，ここではヘッドリーダの出力を使って，ちょっと顔を向ければ作業ウィンドウを切り替えるようにウィンドウシステムを変更した．ディスプレイ上のカーソルの動きを，マウスとヘッドリーダの動きの両方で制御できるようにした．図 2.16は，このシステムのイメージを示すイラストである．この図にあるように，作業モニタ画面が小さい (視野角が小さい) 場合には，実は頭部の動きは必要でなく，視線の方向だけを動かすことで操作しようとし，画面が大きいときに，頭部を動かして，さらに視線を使って操作しようとするであろう．本研究の範囲では視線の検出を行うべきアイリーダの検討をおこなっていないので，ここでは，頭部の動きだけでウィンドウ選択をするとどうなるか試みた．

インプリメンテーション： ウィンドウ切り替えで問題となるのは，頭部の動きのうち，どれが切り替えの意志を反映しているものであるかを認識して選択することである．これは，実は，この応用だけでなく，ノンバーバル言語インタフェースが共通して抱えている大きな問題である．可能性のある方法としては，(1) メッセージを伝える動作に非日常的な動作を割り当てて，ほかの曖昧な動作と識別し易くする，(2) 多くの動作が理解できるシステムをつくり，さらにその上に文脈を考慮する機能をいれて意

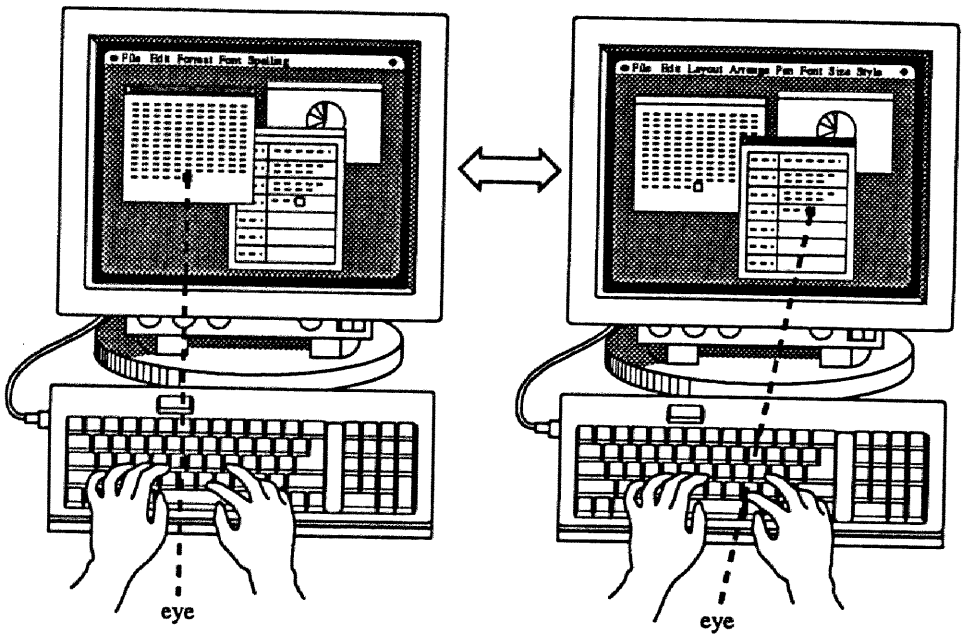


図 2.16: 頭部動作によるウィンドウ選択

表 2.1: 押下キーとシステム動作の割当

押下キー	システム動作
ctrl	カーソルを頭部方向に移動
ctrl+shift	カーソルが位置するウィンドウを close(アイコン化)
ctrl+meta	カーソルが位置するアイコンを open(ウィンドウを開く)

味のある動作かどうか判定する，(3) 音声などのほかのメディアを使ったマルチモーダルなシステム，などがあるが，最も単純な方法として (4) キーボードの特殊キーを押すことによって動作メッセージを伝達する方法がある．これは，Emacs と呼ばれる画面エディタで 1 つのキーに複数のコマンドを割り当てることを拡張してできるので，ホームポジションを移動せずに操作可能である．実験システムでは，左の 3 つの特殊キー（例えば，ctrl キー，meta キーと shift キー）が押されたときだけ，頭部の動作に意味をもたせるようにした．さらに，ウィンドウのアイコン化，アイコンになっているウィンドウを開くコマンドも作成した．表 2.1がその割当てを示す．

実験結果と考察： 上記の機能を，SUN ワークステーションの SUNwindow 上に作成し，作業場所の変更やウィンドウの開閉を行なってみた．頭部方向を検出するヘッドリーダのプログラムも同じワークステーション上で走行させている．

ほぼ思う通りの動作が実行されるが，使用感はスムーズとはいえず，改善する余地があることがわかった．大きな問題としては，

1. ヘッドリーダのサンプリング周波数が低い (画像処理のせいであるが，毎秒 6–8point/sec.) ので，思うようにカーソルが動いてくれない．
2. 位置の精度が低く，小さいアイコンを選択するのが困難．
3. 小さいスクリーンに対してオーバーアクションになって，かえって疲れる．

などが，あげられる．特に上記 1 が，快適性を大きく左右すると思われ，画像処理速度に制限のある現時点では，改善が困難である．これは今後の課題である．

以上のように，システムの反応速度に難点はあるものの，視方向を計算機側がみて，興味のある場所を推定するインタフェースは将来いろいろな場面で使われることと思われる．

(b) 電子秘書

背景： 序論でも述べたように，首を縦や横に振って，「はい」，「いいえ」というメッセージを伝えることは，小さな子供でも行なうし，万国共通性が高い．また，言葉を話すことに障害のある人でも首を振って意志を伝えることがある．首振りによる意志伝達は，ノンバーバル言語のもっとも基本的な動作である．

システム構築： そこで，ヘッドリーダの頭部動作検出機能に高次の動作判定機能を追加して，「はい，いいえ」の判定を可能にした．そして，この機能を対人間インタフェースのメタファとして考案した電子秘書のシステムに組み込んで動作を確認する．ここで，電子秘書は，

1. 人（ユーザ）を検知して，セッションをスタートする．
2. メールの既達をチェックして知らせる．
3. 作業環境メニューを提示して，選択された作業を行なう．
4. 作業の中断の指示を受け取り，実行する．
5. 人の不在を検知して，セッションをクローズする．

という機能をもっていることを仮定して，プログラム化された手順を実行するシステムを構築する．

インプリメンテーション： そこで，ヘッドリーダがもつ機能を活用して，電子秘書の目として，以下の役割を実現する．

1. 人の在不在を検知する．— 頭部検出部で，背景から抽出した頭部領域の面積があるしきい値を越えた場合，人がいると判定する．
2. 視線（頭部方向）を調べて，選択メニューを推定する．— 後述するように4つのメニューをディスプレイ上に作成したので，［右上，右下，左上，左下］の4つの量子化した状態だけを出力する．量子化した状態の判定は2.6.1に述べた通りである．
3. ［はい，いいえ］をコマンドとして受け取りメニュー選択やプログラム中断を実行する．— はい，いいえの検出は2.6.1に述べたアルゴリズムで行なう．

実験システム全体の構成を図2.17に示す．ヘッドリーダの部分はSUNワークステーション，電子秘書のコントロールモジュールとメニュー表示，さらにビデオメールの表示，ビデオ表示をIRISワークステーション，音声の出力をパーソナルIRIS(p-IRIS)でおこなった．ビデオメール，ビデオ番組はあらかじめ録画しておいたものを，リモートコントロール可能なVTRにおき，IRISから再生コマンドを送ってIRIS上の動画表示機能を使ってディスプレイ上に表示した．電子秘書の音声も，あらかじめ定型の文章を録音しておき，p-IRISの音声モニタで再生した．システムの動作は図2.18のフローチャートによる．

考察： キーボードやマウスといった従来のインタフェースデバイスをまったく使わない，計算機と対話する環境のメタファとしての電子秘書を，現在のヘッドリーダが持っている機能だけをつかって構成した．このシステムは未評価であるが，将来これにほかのサブシステムや，音声認識器を接続することを検討中である．

2.7 むすび

簡単な画像処理を使って，人物頭部の動きを検出する手法を提案し，具体的なアルゴリズムと，実際のシステムによる実験結果を示した．画像処理の部分ではしきい値処理により頭部と顔部の領域を求め

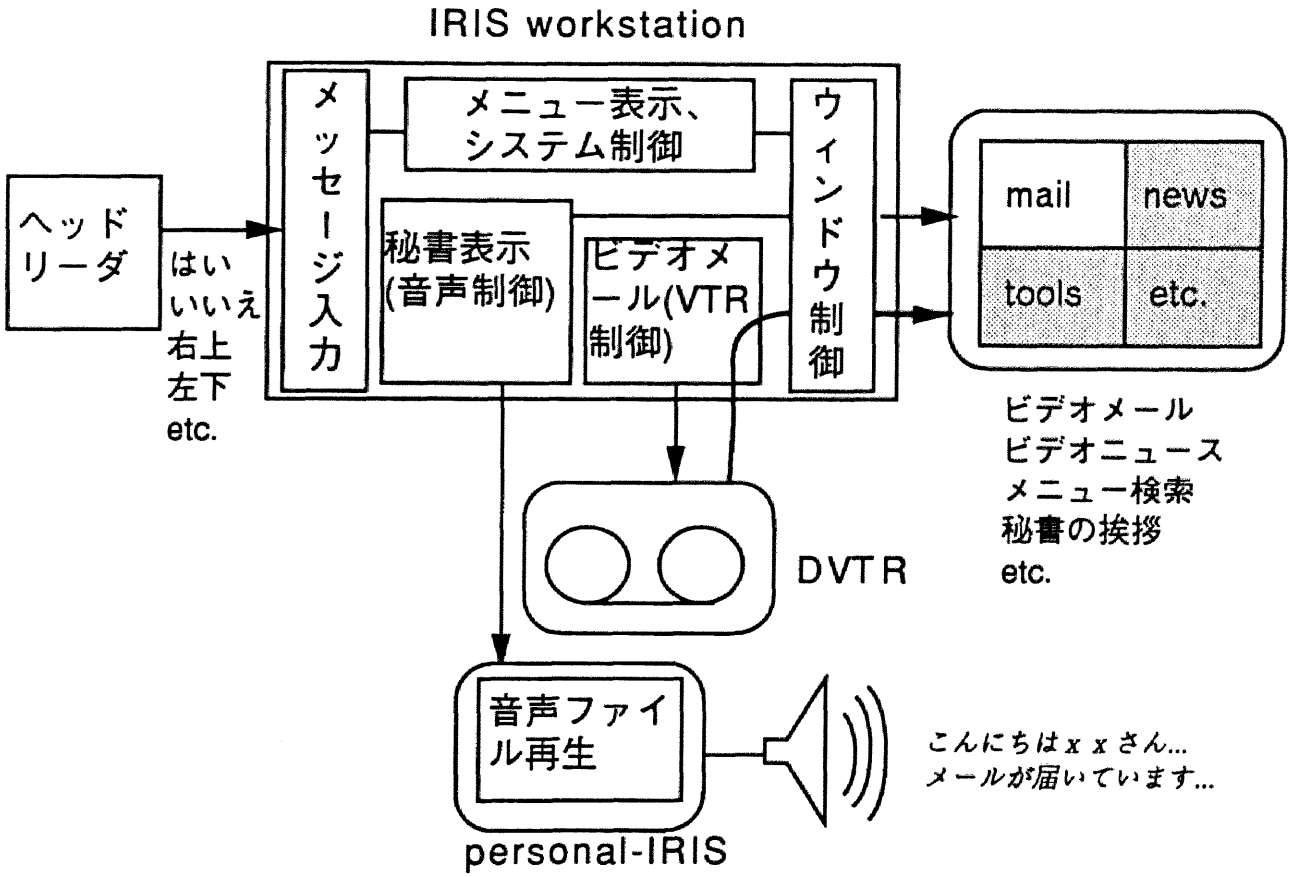


図 2.17: 電子秘書のシステム構成

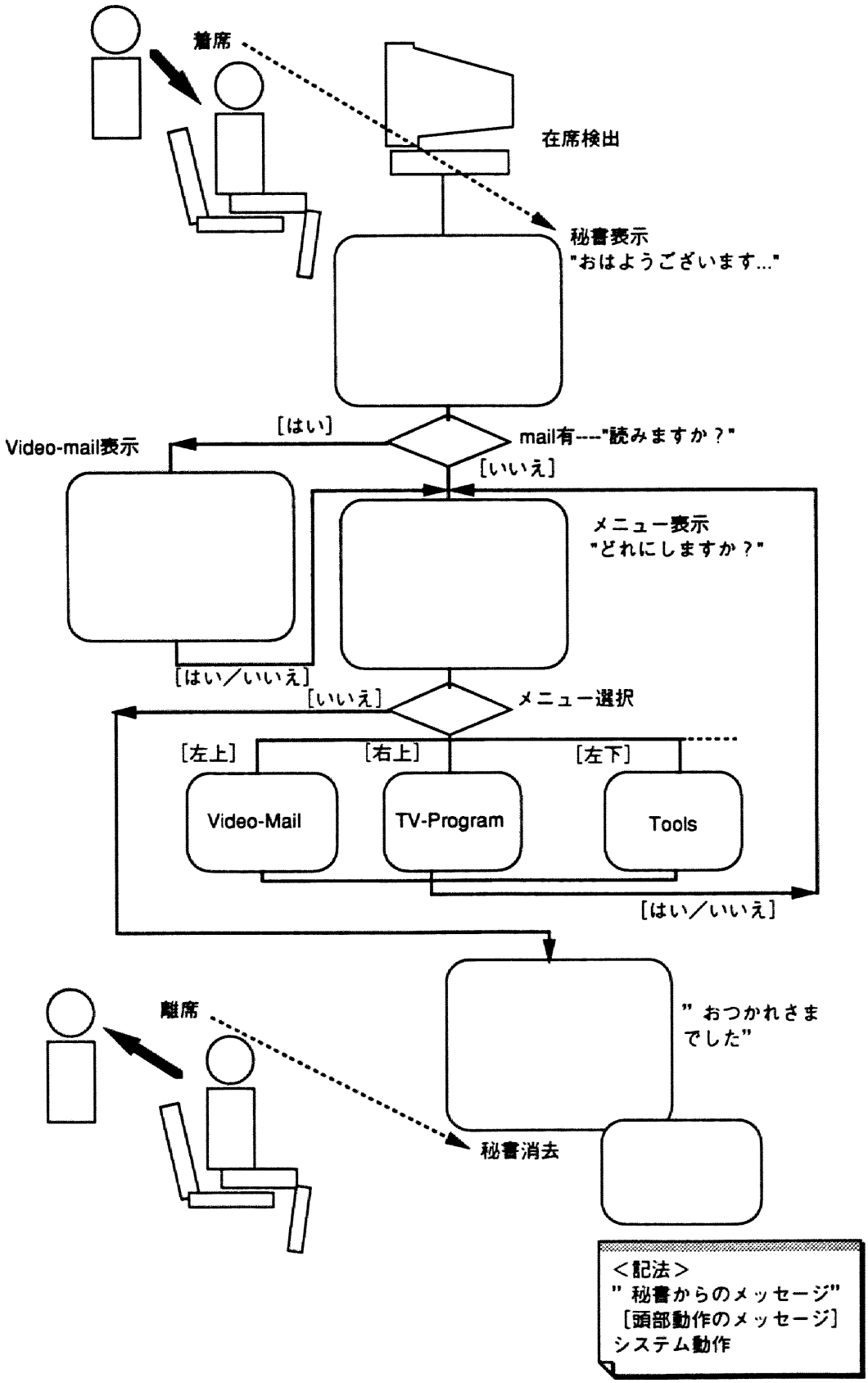


図 2.18: 電子秘書の動作

る手法を示し，それから得られる画像平面上の特徴から3次元動きの解釈をする方法を示した．さらにヘッドリーダのシステムだけを使った，マンマシンインタフェースの例として，いくつかの応用例を検討した．

本章では，重心のずれから回転角への変換関数を，頭部の3次元モデルを使って推定すること試みた．そこでは，一般的な関数形を前提としつつ，1次近似することを中心に述べ，精度よく近似可能な条件を示した．また，実際のシステムも線形関数でインプリメントした．任意の髪型や，回転角に対する近似精度のよい変換関数形を求めることは，今後の課題である．対象が限定される場合には，それに応じて変換テーブルを作成するなどの方法もある．　ここで提案しているのは，髪と顔のみえかたというヒューリスティックを利用した動き解釈の手法であり，対象範囲が限定されることになる．しかしながら，これを前処理として用いれば，後段においてモデルを用いた動き理解におけるマッチング空間の削減に役立つと考える．また，画像を用いた人物の動作理解が，計算機とのインタフェースにどのように貢献できるかを試す1つの実験台として，実時間で動作するシステムは意義深いと我々は考えている．動作量の測定手段として使用できるほどの精度を期待することは現時点で困難であるが，「ハイ・イイエ」のような単純な動作の識別ならば十分に応用可能である．

領域の重心を主な画像特徴としたことにより，以下のような性質があることを知っておかなければならない．まず領域抽出の段階における雑音によって結果が大きく左右されないという長所がある．この方法ならば眼鏡や髭があっても，動作することを強調しておきたい．短所としては各軸に関する情報が独立に求められないことである．さらに，頭部の重心位置は髪型に大きく左右されることも明記しておく．顕著な例としては，ポニーテールのように後ろで髪をしばると，見かけ上頭部形状が前後に細長い物体となり，横を向いたときの頭部の重心が大きくずれることになる．これを防ぐ一つの方法として，楕円領域をあてはめるなどして，注目する領域として髪のシルエットではなく，髪の根元を考えることができよう．

動き検出の精度は，画像の解像度に関係があり，直感的にはこれらは比例すると考えられる．また，実際に使う場合には個人の顔形状や髪型の特徴が異なるため，決められた数点の向きを入力して，変換の関数形を決定する必要がある．個々人でこれらがどう異なるかは，今後の検討課題である．

知的符号化のための頭部動作抽出としては，本手法だけでは高い精度の動作量を得ることは困難である．しかしながら，計算速度が早いことを利用して，後段のモデルベース動作抽出法の前段としてマッチング領域を減少させる処理として有効と思われる．

6節では，ヘッドリーダの出力を利用して新しいインタフェースを探ることを目的に2つの応用を検討し，実験システムの作成を行なった．非接触，無装着の自然なインタフェースをもった実験システムを構築できたが，ヘッドリーダの処理能力の限界から，動作による意図のスムーズな伝達の確認は，現時点では不可能であった．これは今後の課題である．

第 3 章

表情の認識 — フェイスリーダ —

3.1 はじめに

コミュニケーションにおける動作メディアはたくさんあるが，インタフェースを考える上で，顔の表情は欠かせないものである．人間対人間の会話においては顔の表情は相手の状態を知るのにかなり役に立つ．マンマシンインタフェースにおいても，人間の表情からその感情を知ることができれば，コンピュータ側にそれなりの対応をさせることが期待できる．例えば，コンピュータを操作する状況を考えると，操作が判らなくて困った顔をしている人には，システム側が自動的により初歩的，説明的なガイダンスを与えることも可能となるだろうし，疲労の表情を表しているときには，自動的にフォールトトレラントな環境に移行することもありうる[1]．表情は多くのメッセージを伝達している[2] ので，マンマシンインタフェースとしてでなくとも，多くの応用が考えられる．たとえば，テレビ番組の視聴率測定システムや商品監視システムなどのモニタリングシステムにおいて，関心度を表情から読み取るシステムが考えられる．そこで，本章は表情を読み取るフェイスリーダ (Face reader) について検討する．

ここでは，まず表情を認識するための基礎となる顔の変形を測定する方法を検討する．顔の変形が筋肉の収縮にもとづくものであることに注目して，筋肉の動きを，オプティカルフロー計算によって検出する．オプティカルフローを使うと，頬のような特別の特徴点のないところでも，皮膚がつくるテクスチャによって動きを検出することができ，個々の筋肉の動きをとらえることが可能となる．そこで，本章では，オプティカルフローのデータをもとに，トップダウンとボトムアップの 2 つのアプローチで表

[1] 自動車の居眠り防止に表情読み取りを使ったらどうだろうか．

[2] 実際，言語に比べてどれくらいの割合のメッセージを伝達しているか興味のあるところであるが，統計的なデータを示した報告はない．“ノンバーバルコミュニケーション[本名 信行, 1981]” の訳者である本名らは訳書の「あとがき」で “バードウィステルは伝達で言語がしめる割合は 30-35% にすぎないと計算している．メーラビアンはもっと極端で，メッセージの 7% が言語，38% が準言語（トーン，イントネーションほか），そして 55% が顔の表情によって伝達されると判断している ([本名 信行, 1981]からの引用，p.237)” と，述べている．しかし実際は，バードウィステル (R. L. Birdwhistell) は著書 “Kinesics and Context [Birdwhistell, 1970]” のなかで，“Our present guess is that in pseudostatistics probably no more than 30 to 35 per cent of the social meaning of a conversation or an interaction is carries by the words.” と述べているに過ぎず，これらの数値の根拠は工学的観点からは甚だ危うく，注意が必要である．しかしなお，表情が多くのメッセージを伝達するメディアであることに異論を唱えるものは少ないであろう．

情動画像を解析する．すなわち，まずトップダウンアプローチでは筋肉モデルを仮定した表情の記述方法を示し，次にボトムアップアプローチとして動画像のメッシュ特徴からパターン分類を行う感情識別法を示して，表情認識の手法を検討する[Mase, 1991].

なお，この章では表情と感情を区別し，表情とはある感情の表出動作のことをさす．したがって，感情の識別とは感情に関わる表情を識別することである．例えば，笑いと怒りという2つの感情に対して，それぞれの表情を動作として記述することが1つの課題であり，もう1つの課題は，その表出動作である笑いの表情と怒りの表情を区別することである．人間は ポーカーフェースなどと呼ばれるように，感情を表さないこともできる[Ekman, 1985]が，ここでは感情がストレートに表出されている場合を考える．

本章の構成は以下のようである．

まず2節で表情に関する研究を概観し，心理学などで使われるFACSについて紹介する．また，本章と次章で用いる動きデータのもとになる，オプティカルフローのアルゴリズムを解説する．さらに表情筋の配置を示して，つづく節の準備とする．

3節はトップダウンの認識として，筋肉モデルによる表情の記述をゴールとした手法を示し，FACSとの関係について整理する．

4節は実際の表情動画像から筋肉モデルによる記述を行なった実験結果を示す．また，筋肉モデルに基づくFACS記述を試みた例を示す．

5節はボトムアップで表情を認識する手法を，古典的パターン認識の手法を応用して構築する．オプティカルフローの統計量を特徴パターンとして，4つの基本的感情を示している表情の識別を行なうシステムを示す．すなわち，アルゴリズムを概観すると，空間的にブロック化した領域でのオプティカルフローの2次までの統計量データから，Fisherの判別基準を利用して次数の低い特徴ベクトルを生成し，さらにk-最近傍法で学習データとの距離を基準に識別を行う．これは，トップダウンに特徴を記述する従来の手法に対して，ボトムアップにデータを統計処理して判別する方法と見ることもできる．

6節は上記のシステムで実際の感情を示す動画像の分類識別実験の結果を示す．4種類の感情について，20の学習サンプルと30のテストサンプルを使って感情識別の初期実験を行った結果を示す．それによれば，約85%の識別結果を得て，有意な判別が可能であることを述べる．

3.2 表情とオプティカルフロー

3.2.1 顔の表情に関する研究

表情を認識あるいは分析する研究は，人間が行う感情の表出の起源を探ろうとしたCharles Darwin [Darwin, 1872][Ekman, 1973] の研究以来100年以上もの長い歴史がある．近年Ekman とFriesenは感情の表出に関連した顔面の動きを測定して記述する目的で，表情をコードとして表現するFACS(Facial Action Coding System)[Ekman and Friesen, 1978]を作った．これは現在，心理学や臨床など多方面で使われているが，顔面筋の位置を考慮した，長年の心理学的な実験や調査に基づく表情の記述として，1つのまとまった体系をもっているため，最近コンピュータグラフィックスなどの工学

分野でも利用されつつある．また，Ekman らは FACS が完成する以前に表情と感情を結び付ける研究を行い，驚き，恐怖，怒り，嫌悪，悲しみ，および幸福の感情が顔の造作の変化とどう関連づいているかを整理している[Ekman and Friesen, 1975]．（こういった努力は心理学などの立場で各種なされているようで，似たようなものとして，Izard[Izard, 1979]による MAX というシステムもある．）

3.2.2 顔と表情の画像処理

このような，表情の分析を自動化したり，顔による人物自動識別を行なって，他の用途に用いるためには顔の画像処理技術が不可欠であり，長い間の研究課題である．コンピュータによる顔と表情認識に関連する研究は，顔の形状の記述，顔の識別を主目的とした特徴抽出，表情の分析と記述，さらに表情からの感情認識がある．

人間の顔は目，鼻，口などいくつかの際だった造作があり，表情を認識するにはこれらの動きや形状の変化を手がかりとすることが考えられる．これらの造作を抽出し特徴となる点をさがす手法は，顔画像処理による個人識別[Goldstein *et al.*, 1972][Bruce, 1988][Kaya and Kobayashi, 1971]のために各種提案されている．[3]坂井ら[坂井ほか, 1973]は顔画像にラプラシアンフィルタをかけてエッジ抽出を行い，エッジ画像の周辺分布と顔の構造を利用して部品の抽出を行った．金子ら[金子ほか, 1988]は，口や目の抽出のために 2 値化した部分画像にモデルをあてはめた．Kass ら[Kass *et al.*, 1987]の Snakes は，エネルギー最小化問題によって spline 曲線をエッジに漸近させるモデルであり，エッジがはっきりしていて，適当な初期値が与えられるときに効果があり，唇（口紅を塗った）の抽出を 1 例として示した．唇の形状の抽出は計算機による読唇[Petajan, 1985][間瀬，ペントランド, 1990][田村ほか, 1989]を目的として 2 値化やエッジ抽出，spline フィッティングなどが行われている．しかし，唇は変形があるうえ我々が感じているほど頬とのコントラストが大きくないためこれらの画像処理では抽出が困難である．そこで，モデルを使ったり，口紅の助けを借りるなどの必要がある．色空間のセグメンテーションにより領域を抽出する方法も多種提案されている（たとえば，[佐々木ほか, 1991b]）．しかしながら，照明や撮影条件を工夫・限定しないと安定してこれらの領域を抽出することは困難である．

表情を画像処理に基づいて分析する研究は，画像分析合成による知的符号化 (モデルベース符号化とも呼ばれる．[相澤ほか, 1989]などを参照) を中心に行われている．原島らのグループは FACS を利用した表情合成システムを作成[崔ほか, 1990]した後，合成規則を利用して表情をアクションユニット (Action Units,AUs) で記述することを試みている[崔ほか, 1991]．すなわち標準顔と表情顔の間の部品の変形量を AU に対応づけて表情を記述した．部品の自動抽出が前提になっており，先に述べたように安定した抽出法の検討が必要である．Terzopoulos & Waters は Snakes を応用して表情を分析してコンピュータグラフィックスで顔を合成した[Terzopoulos and Waters, 1990]．しかしながら Snakes は基本的に会話処理を前提としており，初期値の設定やパラメータの調整が必要[Waite and Welsh, 1990]であるほか，原画像のエッジを際だたせるために過度のメーキャップを必要とし，追跡も不得意であるという欠点があった．これを回避するために[上田ほか, 1991]は Snakes をエラスティックモデルとして拡

[3]顔画像処理による個人識別については多くの手法が提案されている．最近の動向をみるためには[Turk, 1991][南, 1991]などを参照のこと

張し追跡のできる輪郭モデルを提案している.

表情筋の動きをシミュレートして顔の表情をコンピュータグラフィックス(CG)のアニメーションで表現する研究もCGの初期のころから行なわれている[Platt and Badler, 1981][Parke, 1975]. Waters[Waters, 1987]はFACSの考え方に基づき, いくつかのAUを再現し, その組合せで"怒る", "笑う"などの表情の表現に成功した.

表情から感情を認識する過程は,

1. 顔の造作の抽出
2. 顔の変形の検出
3. 変形の表情要素による記述
4. 表情の解釈(基本感情の認識)
5. 複雑な感情の認識

の5段階に分けることができる. 崔らの表情分析は, 上記の造作の抽出を前提として, 2番目の変形と表情要素の結合を2種類の方法で試みた.

ここでは, 顔の変形が筋肉の収縮にもとづくものであることに注目して, 造作の抽出を行なわず, 動画像から筋肉の動きを直接推定して上記の3.までを行なうとともに, 動きの情報をつかって, 直接表情の解釈を試みる.

すなわち, 3節および4節では表情筋の動きの抽出と, その表情筋モデルに基づいたAUによる記述を行なう. さらに5,6節では, 動きの特徴から直接, 各表情がどのような感情を表しているかを識別する. 表情変化を画像処理して得られた特徴に基づく感情の識別を行った研究は少なく, わずかに石井らによる報告[石井, 岩田, 1984]が見受けられる程度である. 石井らは眉の角度, 目の開き具合い, 口の曲率, および口の縦横比の4つの幾何パラメータの変化を計算して, それらが4つの感情を表す画像でどう変化するかを調べたが, 詳細な報告がなされていなかった.

動きの推定には, 後述のオプティカルフローの情報を利用する. オプティカルフローからの表情筋の動作検出は, 動画像の連続するフレームから動きを検出することによって(1)特に標準顔を必要としない, (2)特徴点の抽出が不要であるという利点がある. とくに, 頬のように皮膚のテクスチャだけが頼りとなるところでも動きを抽出でき, 造作を使わずに筋肉の動きを抽出できるという利点がある.

3.2.3 FACS(Facial Action Coding System)

近年Ekman と Friesenが提案したFACS(Facial Action Coding System)[Ekman and Friesen, 1978]は前述のように, 感情の表出に関連した顔面の動きを記述することのできるシステムである. 表層的な皮膚や造作の変形だけでなく顔面筋の位置を考慮しているので, 表情の生成過程に適合した記述法の1つである.

FACSでは, 表情はアクションユニット(Action Unit, AU)と呼ばれる表情単位の組合せで記述される. アクションユニットは, 表3.1のように顔の上部の筋肉, 顔の下部の筋肉による動作と, その他の動作, 頭部と目の位置に分類された約60個がある. これらを使って, 顔の画像(時には動画像)を調

表 3.1: FACS のアクションユニットの一覧 (AU 番号と名称)

顔の上部	顔の下部	その他の動作	頭と目の位置
1— 眉の内側を上げる	9— 鼻にしわを寄せる	8— 唇同士を接近させる	51— 左を向く
2— 眉の外側を上げる	10— 上唇を上げる	19— 舌を見せる	52— 右を向く
4— 眉を下げる	11— 鼻唇溝を深める	21— 首を緊張させる	53— 頭を上げる
5— 上瞼を上げる	12— 唇端を引張り上げる	29— 下顎を突き出す	54— 頭を下げる
6— 頬を持ち上げる	13— 唇端を鋭く上げて頬を膨らます	30— 下顎を横へずらす	55— 左へ傾ける
7— 瞼を緊張させる	14— えくぼを作る	31— 歯をくいしばる	56— 右へ傾ける
41— 瞼を力なく下げる	15— 唇端を下げる	32— 唇を噛む	57— 前へ出す
42— 薄目	16— 下唇を下げる	33— 息を吹きかける	58— 後ろへ引く
43— 瞼を閉じる	17— 下顎を上げる	34— 頬を息で膨らます	61— 左を見る
44— 細目	18— 唇をすぼめる	35— 頬を吸い込む	62— 右を見る
45— まばたく	20— 唇を横に引っ張る	36— 舌で頬や唇を膨らます	63— 上を見る
46— ウィンクする	22— 唇を突き出す	37— 舌で唇をなめる	64— 下を見る
70— 眉が見えない	23— 唇を固く閉じる	38— 鼻孔を開く	65— 斜視
71— 目が見えない	24— 唇を押さえつける	39— 鼻孔を狭める	66— 内斜視
	25— 唇を開く（顎は下げない）		
	26— 顎を下げて唇を開く		
	27— 口を大きく開く		
	28— 唇をかむ（吸い込む）		

べて，どのAUが現われているかを判断し，表情を複数個のAUで記述することになる．文献[Ekman and Friesen, 1978]は，どのAUがあるかを判定するための指針の集合である．たとえば，写真を見て眉の内側が上がっていれば，AU1が計測されることになるが，詳しくは，次に示すように解説されている([Ekman and Friesen, 1978]より一部掲載).

#——#

Action Unit 1 – Inner Brow Raiser(眉の内側をあげる)

額のところの大きな筋肉が眉を上にあげる．

- AU1による見かけの変化

  1. 眉の内側を上に引っ張る．
  2. たいてい眉が斜めになったり，ハの字形になる．
  3. ひたい（特に中心）に皺ができる
  4. 外側はあまり動かない．AU2に比べ内側に向かう．

- AU1の作り方 – 意識的に作るのは難しい動きである．

  - 眉全体を上げ(AU1+2)，それから内側だけ上げるのを試みる．
  - AU1+4の写真のように内側に引っ張る．それからAU1を試す．
  - 指で，写真のようにAU1を作る．それを維持するようにする．

  AU1がつくれたら，AU2ができていないか確かめること．

- AU1の最低必要条件

  1. 眉の内側が少しはあがっていなくてはならない．
  2. ひたいの中心に少しは皺ができなくてはならない．

- 参照 — 他のAUとの組合せ，たとえばAU1+2, AU1+2+4の時は条件が変わる．

#——#

3.2.4 オプティカルフロー抽出アルゴリズム

オプティカルフローとは輝度パターンの見かけの動きのことである．従って，理想的には，動き場(motion field)に対応している[4].

オプティカルフローを計算する方法には，様々な方法がある[Aggarwal and Nandhakumar, 1988]．大別すると輝度の時空間パターンに基づく方法としてのパターンマッチングによる相関法，勾配法，フィルタリング法と，線分や領域などの特徴パターンをトークン(token)とするマッチング法とに分けられる．対象とする表情動画像はショートレンジの動きを仮定できるので，処理量の多い相関法は不要である．フィルタリング法[Heeger, 1988]は数フレームを使って，安定したフローを得ることができる

[4]常に等しいとは限らない．例えばテクスチャのないボールが回転していてもオプティカルフローは検出できない．

が，表情画像は同じ動きが持続しないので使うことができない．トークンマッチングによる方法は一般的にはロングレンジのフローを計算するのに適しているが，安定したトークンを得るために密なフローの計算には適さない．また，顔の場合にはトークンとなる特徴を顔の造作から抽出する処理が必要となり，困難である．

そこで，勾配法の代表的なアルゴリズムである Horn&Schunck の手法[Horn and Schunck, 1981]を用いることにする．Horn らのアルゴリズムはその後いろいろな改良がされているが，ここではオリジナルの 2 フレーム法を適用する．勾配法は時空間画像の輝度の傾きから速度場を求める手法で，特にはっきりとした特徴点がなくても，速度が求まるという点で，頬や額の動きをとらえるのに適していると考えられる．顔は剛体ではないが，動き場は局所的にはスムーズであると考えて差し支えない．勾配法によるオプティカルフロー計算の原理は以下のようになっている．

時刻 t の画像の点 (x,y) における輝度を $I(x,y,t)$ で表すとする．いま点 (x,y) におけるオプティカルフローベクトルを速度成分 $(u(x,y),v(x,y))$ で表すとする（=パターンが速度 (u,v) で平行移動しているとする）と，時刻 $t+\delta t$ における点 $(x+\delta x, y+\delta y)$ における輝度と等しいはずである．なお，$\delta x = u\delta t$，$\delta y = v\delta t$ である．すなわち，

$$I(x+u\,\delta t, y+v\,\delta t, t+\delta t) = I(x,y,t) \tag{3.1}$$

が微少時間について成立する．この式は，輝度の時空間の変化がなめらかであるなら，$(u,v) = (0,0)$ の回りに Taylor 展開でき，1 次の項だけ残すと，

$$I_x u + I_y v + I_t = 0 \tag{3.2}$$

という，オプティカルフローの拘束式と呼ばれる，速度 (u,v) に関する一次方程式が求まる．ここで，I_x, I_y, I_t は I の x, y, t に関する偏微分である．

しかしながら，これは (u,v) の 2 つの未知数をもつので，もう 1 つの拘束が必要である．輝度の時空間の変化がなめらかであるという仮定から，局所的にオプティカルフローのグラディエントが最小になるように制約条件をつけて，以下のようにラグランジェの未定乗数法で解くことによって，各点の速度を求めることが出来る．

$$e = (u_x^2 + u_y^2) + (v_x^2 + v_y^2) + \lambda(I_x u + I_y v + I_t)^2 \rightarrow min. \tag{3.3}$$

これを離散空間で最小化する問題として扱うと，まず e を u と v で微分して次式を導出する．

$$\frac{\partial e}{\partial u} = 2(u-\overline{u}) + 2\lambda(I_x u + I_y v + I_t)I_x \tag{3.4}$$

$$\frac{\partial e}{\partial v} = 2(v-\overline{v}) + 2\lambda(I_x u + I_y v + I_t)I_y \tag{3.5}$$

これらの式がゼロとなるところで e は極値をもつので，式を整理すると，次の繰り返し法による解がえられる．

$$u_{ij}^{n+1} = \overline{u}_{ij}^{n} - \frac{\lambda(I_x\overline{u}_{ij}^{n} + I_y\overline{v}_{ij}^{n} + I_t)}{1+\lambda(I_x^2+I_y^2)} I_x \tag{3.6}$$

$$v_{ij}^{n+1} = \overline{v}_{ij}^{n} - \frac{\lambda(I_x\overline{u}_{ij}^{n} + I_y\overline{v}_{ij}^{n} + I_t)}{1+\lambda(I_x^2+I_y^2)} I_y \tag{3.7}$$

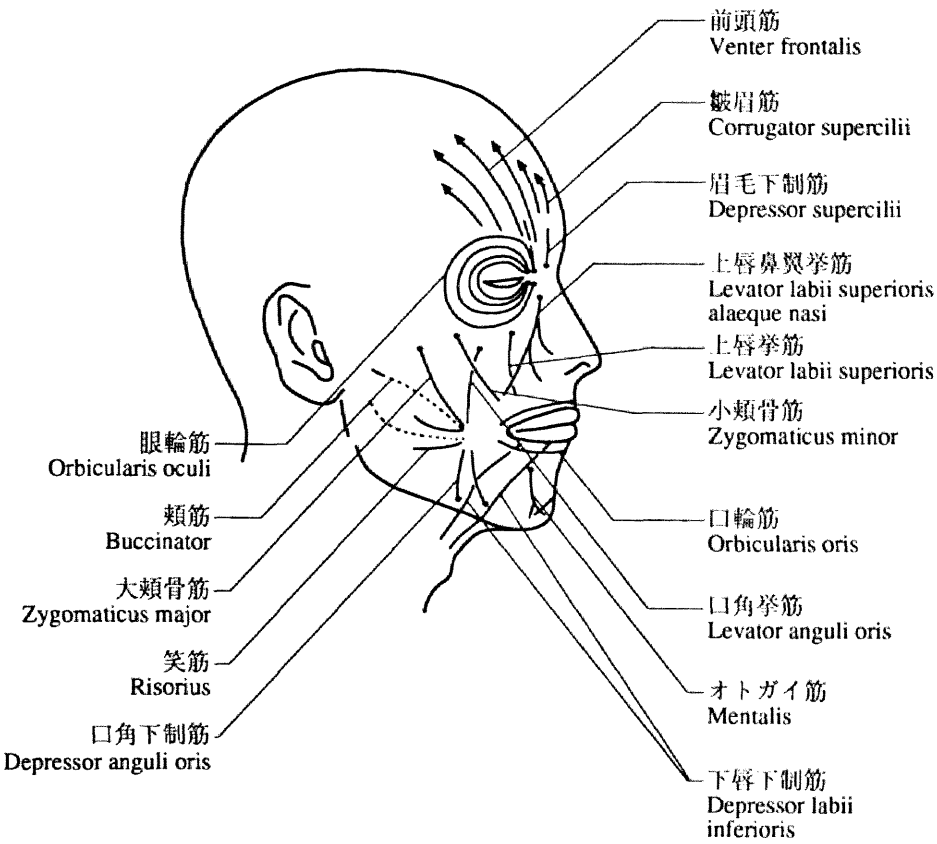


図 3.1: 顔の表情筋

3.2.5 表情筋

顔には大きく分けて2種類の筋肉；表情筋とそしゃく筋があって、いろいろな表情を作る．そしゃく筋は収縮によって下顎骨を頭蓋骨に近付け下顎をさげ、口唇を開閉させる．表情筋は主なものだけで20種類以上あり、それらの収縮・弛緩が複雑にくみあわさって、表皮の変形を形成している．主なものは、前頭筋、眼輪筋、口輪筋、頬筋、頬骨筋、口角筋などである[森 於菟, 1950]．図3.1にこれらの筋肉の位置関係をしめす．

表情筋は多重に結合しているため各々の筋肉の弛緩収縮の量を正確に測定することは困難である．特に頬の周りの筋肉は、例えば、笑うときには頬筋、大頬骨筋などが同時に収縮する[藤田, 1976]とされており，グループ化した筋の動きをとらえるほうが自然である．

3.3 筋肉モデルによる表情の記述

3.3.1 表情筋の動き推定

まず，前処理としてノイズ除去のために時空間ガウスフィルタをかける．次にオプティカルフロー $\mathbf{u}(\mathbf{x}) \equiv (u(x,y), v(x,y))$ を連続する2フレーム間で求める．あらかじめ、筋肉があると思われる場所に図3.2のように窓を設定し、各窓内でフローを平均化して、筋肉の方向成分のフローの大きさを計算

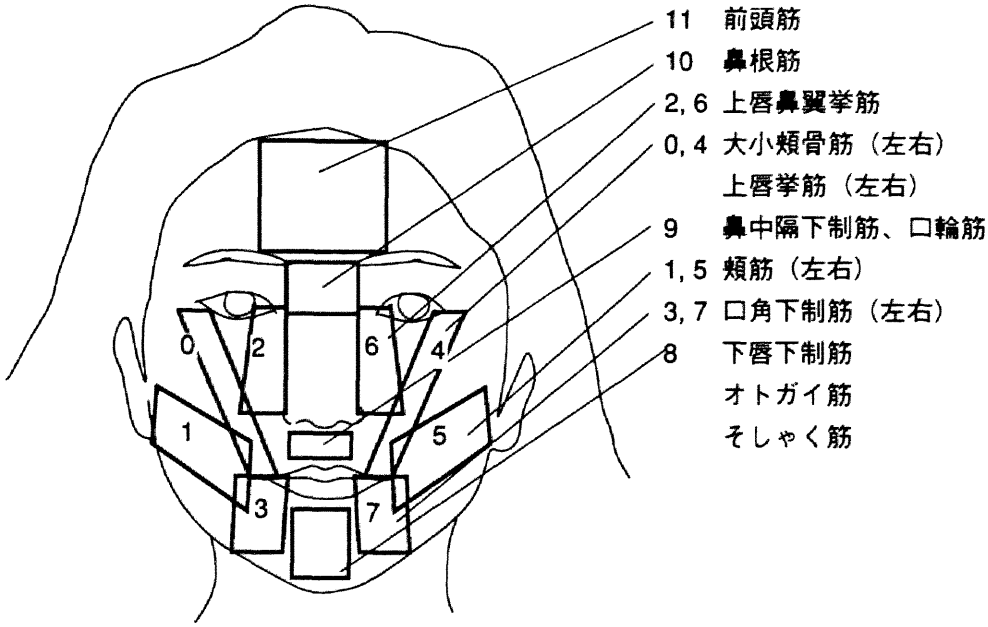


図 3.2: 表情筋の動作検出窓

図 3.3: 表情筋の動作推定例 (幸福)

する．すなわち，

$$m_i = \frac{1}{S_i}\int_{window_i} \mathbf{u}(\mathbf{x}) \cdot \mathbf{n}_i \, d\mathbf{x}. \tag{3.8}$$

ここで $\mathbf{n}_i$ は窓 $i(i = 0, ..., 11)$ の単位方向ベクトルで，各窓に位置する筋肉の主たる収縮方向を示している．m_i が対応する筋肉の動き推定量である．窓の面積 S_i で正規化している．なお図の矢印は筋肉の収縮方向を示しており，$-\mathbf{n}_i$ に相当する．図 3.3はある表情 (幸福，フレーム #10) で筋肉の動きを推定したものである．

これらの窓の位置は目鼻口の位置を参照して，会話的に与えている．将来は自動化することを目指す上からも，これらの窓は実際の筋肉の位置を考慮しながら，設定しやすさに重点をおいて場所を決める必要がある．例えば，図 3.2で窓 0 は頬骨筋の動きをおもにとらえるように考えているが，唇の左端点から左目の左端点に直線を引くように設定している[5]．オプティカルフローで得られるのは，筋肉の弛緩収縮による表皮の変形の推定量であるから，このように注目するところが少しずれても，大きな問題とはならない．

3.3.2 筋肉モデルと FACS の関係

表情を認識するうえで，表情をある符号体系で表現することが必要になる．これまでの議論で，オプティカルフローデータを基礎とする筋肉モデルに基づく記述が一応可能である．一方，すでに述べたように FACS に基づく記述は心理学の分野にとどまらず，コンピュータビジョン，コンピュータグラフィックスなどの分野でも使われ始めている．そこで，ここで示した筋肉モデルによる記述と FACS のアクションユニット (AU) の部分集合との関係づけを行う．

まず，幸福，怒り，驚き，嫌悪，悲しみ，および恐れ，の 6 つの基本的表情を記述するための主なアクションユニット；AU 1, 2, 4, 5, 6, 7, 9, 10, 12, 15, 17, 20, 25, 26 を抽出した ([崔ほか, 1990])．つぎに，FACS のマニュアルと解剖学の文献を参照して対応付けを行った．FACS における AU はそのほとんどが静的な形状の変化の記述であり，1 部のみが動きの記述である．したがって，動きベクトルに基づく筋肉モデルを直接あてはめることのできる AU はごくわずかである．結局，対応付けは動作とそれによる形状への影響を考慮して行った．さらに，現時点での筋肉モデルは顔の全領域をカバーしていないので，全ての AU について割り当てることはできていない．

表 3.2がその対応付けの結果である．この筋肉モデルを使うと，1 つの筋肉の動きは複数の AU と関係づけられる．例えば，AU25(唇を開く) と AU26(顎を下げて唇を開く) は m_8(口の下方) だけでは区別できない．そこで，m_8 が正の時は AU26 を割り当てることにする (AU26$^+$ と記述する)．m_9 は主に口輪筋に関係しているが，ほかの筋の影響も受けながら口が開いたときに観測される．したがって，m_9 が負値のとき AU25 を割り当てることにした (AU25$^-$ と記述する)．

[5] これに対して，Ekman & Friesen[Ekman and Friesen, 1990] は大類骨筋はもっと浅い角度で，骨への起始の実際の位置はもっと耳に近いところであるあるというコメントをしている．このように，実際の筋肉と窓の位置関係は厳密ではない．しかしながら，心理学における応用などで測定の自動化を要しない場合には，専門家の協力を得て，筋肉の位置に近いところに設定する必要があろう．

表 3.2: 筋肉窓とアクションユニットの関係

筋肉窓	関連する筋肉	動作と結果	対応する AU
m_0(左) m_4(右)	大頬骨筋 眼輪筋	唇の端を引っ張り上げる 頬を上げる	AU6$^-$(頬を上げる)
m_1(左) m_5(右)	大頬骨筋 頬筋	唇の端を後ろへ引っ張る	AU12$^-$(唇端を引張り上げる)
m_2(左) m_6(右)	上唇挙筋	上唇を上げる	AU10$^-$(上唇を上げる)
m_3(左) m_7(右)	口角下制筋	唇端を下げる	AU15$^-$(唇端を下げる)
m_8	オトガイ筋 顎の運動(弛緩) 下唇下制筋	顎をあげる 顎を落として唇を開く	AU17$^-$(下顎を上げる) AU26$^+$(顎を下げる), [AU25$^+$(唇を開く)]
m_9	口輪筋 (弛緩)	上唇を弛緩して唇を開く	AU25$^-$(唇を開く)
m_{10}	眉毛下制筋 上唇鼻翼挙筋	鼻根に皺を作る	AU9$^-$(鼻にしわを寄せる)
m_{11}	前頭筋 眉毛下制筋	眉を上げる又はさげる	AU1$^-$(眉の内側を上げる) AU4$^+$(眉を下げる)

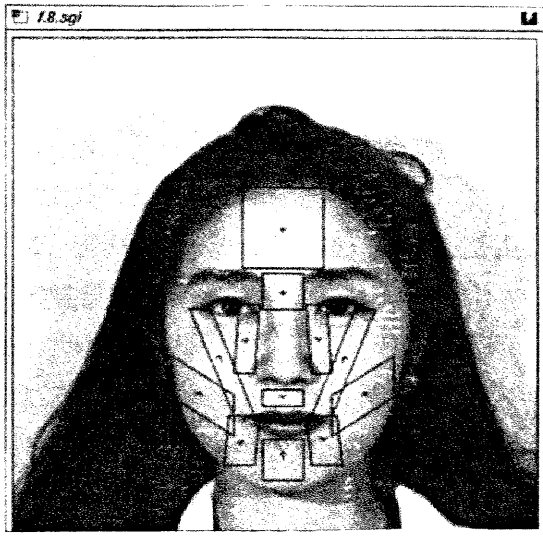
(AU の符号は弛緩と収縮に対応し，$m_i < 0$ のとき AUn^-，$m_i > 0$ のとき AUn^+ とする．)

AU25 は，負の m_9 と正の m_8 が同時に観測されたときに割り当てたほうが適当かもしれない．しかし，ここでは簡単のため一つの筋肉窓の動きに対して複数の AU を割り当てることにして，複数の窓における動きの組合せで AU を推定することはしないことにする．先に抽出した AU のなかの 2,5,7 はこの表のなかにはない．これはいまのところ，目の回り，口唇の端，眉の外側に窓を設定していないことによる．

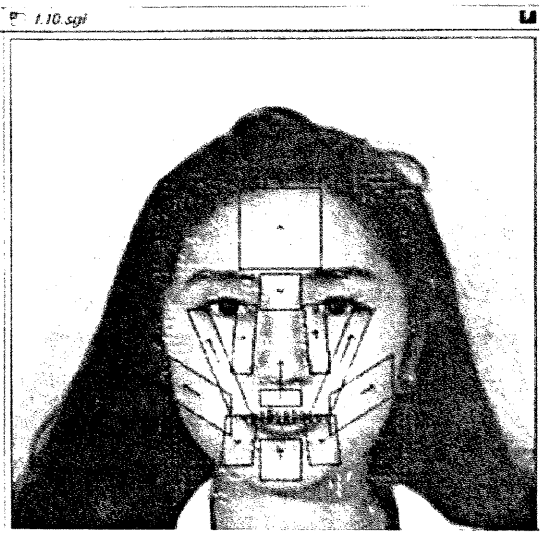
3.4 表情記述の実験結果

3.4.1 筋肉モデルによる記述

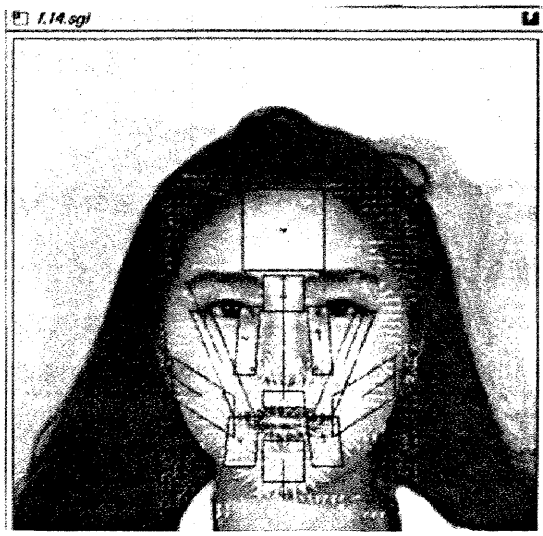
図 3.4は実験に使用した原画像系列に対してオプティカルフローを求めた結果の例である．データは 30 フレーム / 秒（インタレース）で取り込んだディジタル画像を 2:1 でサブサンプリングし，奇数フィールドだけを使った．大きさは 256 × 240 で，1 画素 8bits である．オプティカルフローを求めるのに



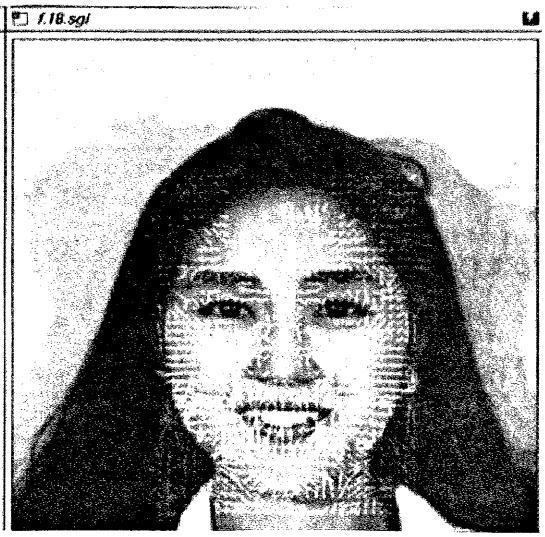
(a) Frame:8 (Neutral)



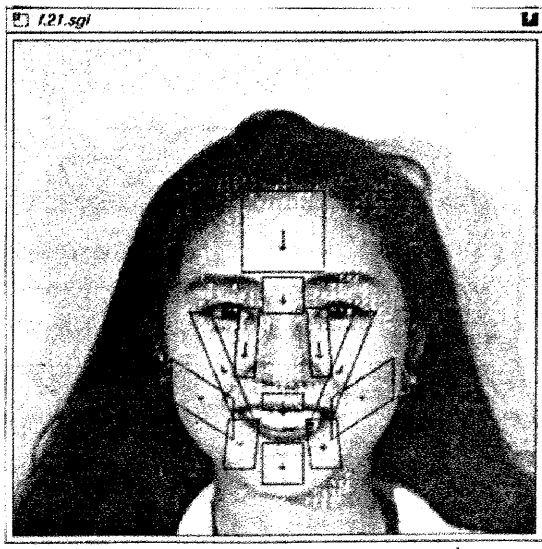
(b) Frame:10 (Start)



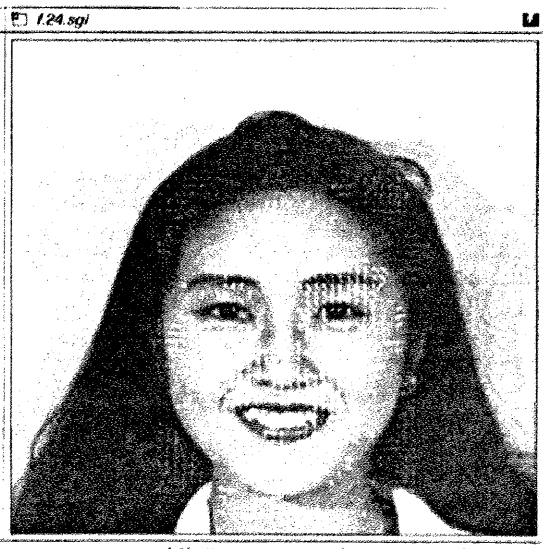
(c) Frame:14 (Maximam Speed)



(d) Frame:18 (Full Motion)



(e) Frame:21 (Maximam Position)



(f) Frame:24 (Relaxing)

図 3.4: オプティカルフロー抽出結果 (女性；幸福の表情)

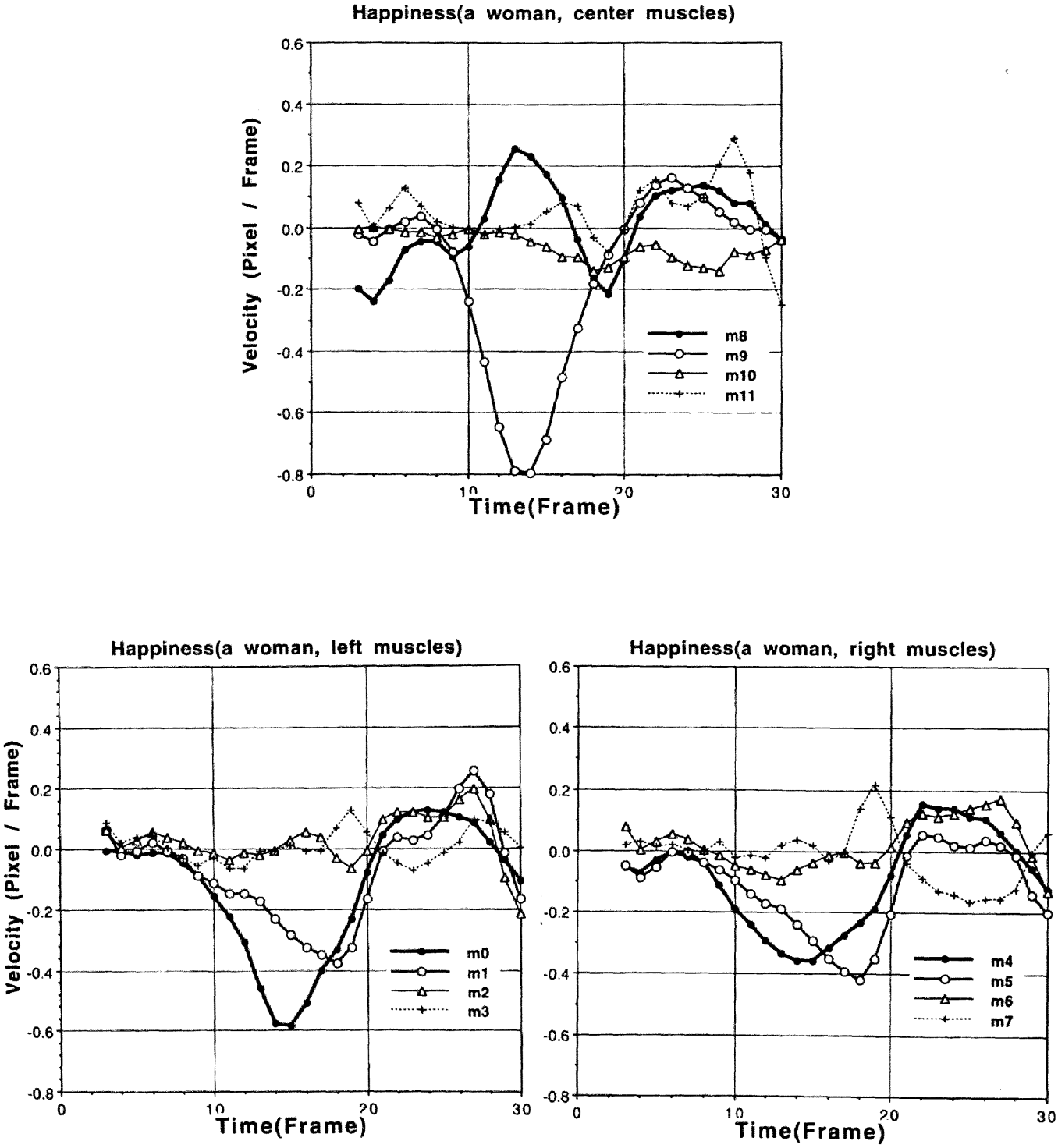


図 3.5: 筋肉の動き推定量の時間変化 (前図の女性)

使った2枚の画像のうちの1枚（変化前）にフローを重ねて表示している．フローは3画素ごとに白矢印で表示している．また，黒矢印は推定した筋肉の速度である．図3.4は「幸福(Happiness)」のカテゴリに入る「微笑む」一連の動作であり，この場合には頬筋，頬骨筋が収縮する．図にみられるように筋肉の動きに対応した場所でフローが観測される．また，一連の表情でも時間と共にフローが変化することがわかる．図中の黒の矢印が(3.8)式で推定された筋肉の単位時間当たりの動き量である．図3.5は筋肉の動き推定量を時間軸に対してプロットしたものである．筋肉窓m_0, m_1, m_4, m_5, m_9で大きな動きが観測され，オプティカルフローデータが表情による筋肉の動きをうまくとらえていることが判る．

図3.7および図3.6は別の人物の表情（男性，幸福）に対して，同じようにオプティカルフローを抽出した例である．このように人物が違っても類似したパターンが得られる．この場合は，口を閉じたまま表情を作っているので，筋肉m_0，m_1，m_9の3つは先の例と似通った速度曲線を示しているが，一方筋肉のm_3とm_8は図3.5とは異なる動きをしていることが判る．なお，m_iの負の値が筋肉の動作として収縮していることに対応し，正の値が伸張に対応していることに再度注意して頂きたい．

さらに図3.8は違う表情（嫌悪）のオプティカルフローであり，明かに筋肉m_3, m_7にフローが集中していることが判る．

3.4.2 FACSに基づく筋肉モデルの記述

図3.5を例にとって，AUを使って，表情の記述を試みる．グラフから平常顔と表情最大の間のフレームで各速度の最大値を求める．そして，それから表3.2を用いて，各AUの有無を判定する．この例は，AU6(0.6)+12(0.35)+17(0.20)+25(0.8)+26(0.28)と，表情を記述できた．ここで，かっこ内の数値は上記の速度の最大値であり，AUの強度とみなすこともできる．このAUの組は，“幸福”の表情のときによく現われるAU6+12+26やAU6+12+17など[Ekman and Friesen, 1978]に近いことがわかる．この記述例は，動画系列からもとめたものであるから，“幸福”に相当するいろいろな表情が順番に現われているとも考えられるので，いくつかのAUの組のOR表現になっていても不思議ではない．

3.4.3 筋肉モデルによる感情の識別

これまでに得られた記述から，感情を識別するにはどうしたらよいか？筋肉モデルに基づく記述から特徴ベクトルを構成して，古典的なパターン認識の手法による識別が可能と思われる．

その際考えられる特徴ベクトルとしては以下のものが考えられる．

1. 各筋肉速度値の最大値の集合
2. 各筋肉速度値の積分（移動量）の集合
3. 各筋肉速度値の時系列パターンの集合あるいはその多項式展開係数
4. 3.4.2節で示したAU記述

これらの筋肉速度に基づいて，トップダウン的に感情を認識するシステムを構成するには，さらに解剖学や心理学の有効に利用した高次のパターン認識手法を検討する必要がある．これは今後の課題である．次節では，ボットムアップ的な感情識別の実験を行なう．

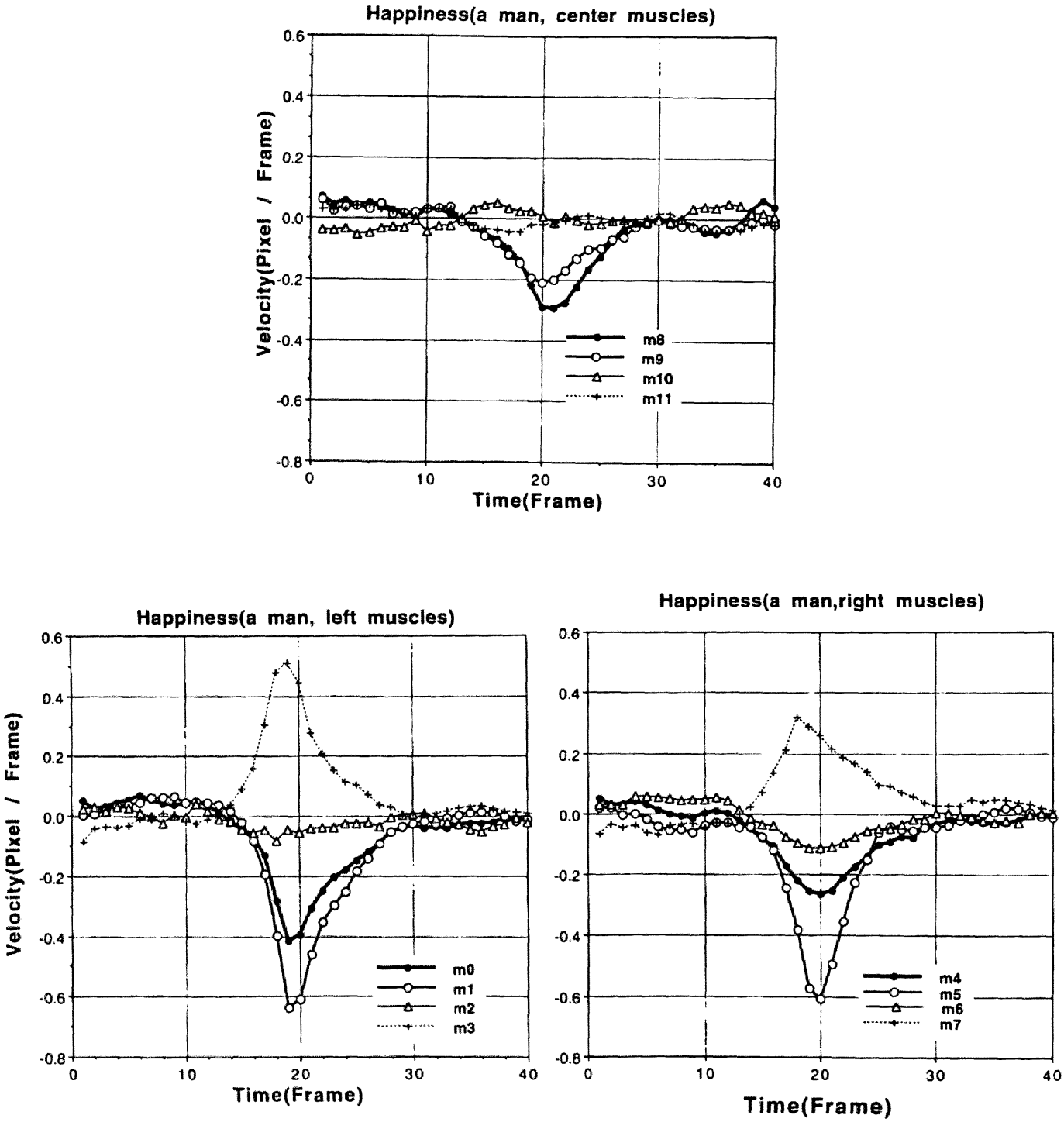


図 3.6: 筋肉の動き推定量の時間変化 (次図の男性)

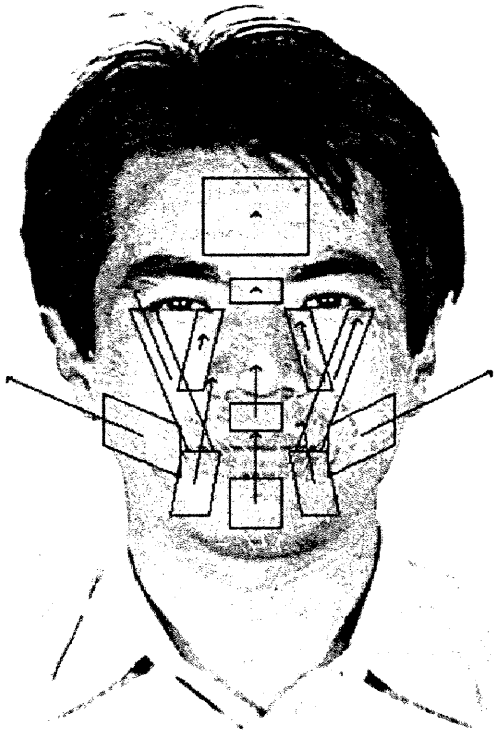


図 3.7: 別の人物のオプティカルフロー抽出結果 (男性、幸福、口は閉じたまま)

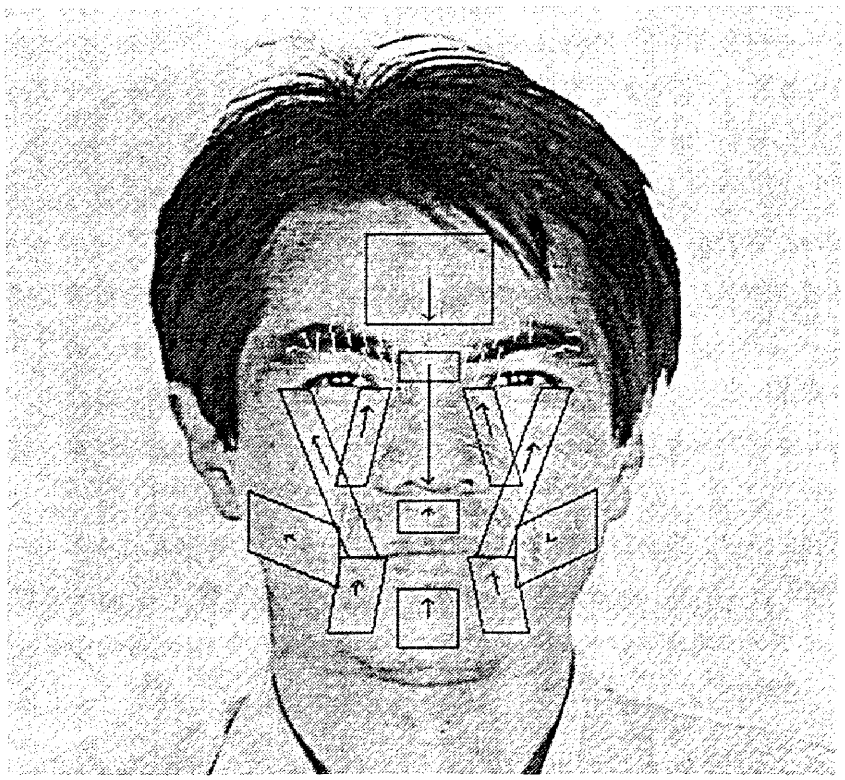


図 3.8: 別の表情（嫌悪）

3.5 感情を表わす表情の識別

3.5.1 顔の表情と感情

Ekman らの表情と感情を結び付ける研究[Ekman and Friesen, 1975]では，“驚き，恐怖，怒り，嫌悪，悲しみ，および幸福”の6つの基本的感情が取り上げられている．Ekman らによると，これらは顔の表情と関連づけられた感情語の語彙として長期間に渡る研究の結果として確認されたものである．さらにこれらの感情と表情のマッピングは万国共通であることが実験的に確認されている．計算機による感情認識システムにとって，これら6つの語彙にあてはまる表情を識別することが第1歩となるだろう．人間による感情の判定に関する Ekman と Friesen の実験([Ekman, 1973, p.206])によれば，幸福，驚き，嫌悪の感情に比べ，他の感情は誤り率が大きい．

また実験用のデータを作るときも幸福の表情に比べて，悲しみと恐怖の表情は作成が困難であることが判った．そこで，“幸福，驚き，嫌悪および怒り”の4つの感情を識別するシステムを作り実験を行った．怒りの表情は悲しみあるいは恐怖の表情より作り易いため含めることにした．

以下では，顔全体のオプティカルフローを原データとして特徴的な場所および統計量を学習サンプルから決定して，表情からの感情認識を行う方法を示す．この手法は筋肉モデルをもたない，パターン認識的な手法である．従って，筋肉窓を設定する必要はないが，顔全体の位置はあわせておく必要がある．

3.5.2 特徴ベクトルの決定

まず，特徴ベクトルの決定方法と特徴ベクトルを用いた識別アルゴリズムについて述べる．すなわちオプティカルフローを計算して得られた時空間の動きデータから統計量を求めK次元の大きなベクトルを作る．さらにその次元数を圧縮して特徴ベクトルとする．その後，特徴ベクトルはk-最近傍法で分類する．k-最近傍法はノンパラメトリックなパターン認識の手法としてよく知られた，準最適な分類手法である[Duda and Hart, 1973].

まず，低次元の特徴ベクトル空間を作る必要がある．オプティカルフローを計算して得られたデータから統計量を求めK次元のベクトルを作る．オプティカルフローは画像の各点，各フレームの組について計算する．動きの境界においてフローが乱れていることによる影響を排除するため，フローを局所的に平均化する．ここでは時間軸方向には独立にフローのデータを扱う．なお，フロー計算の原理から，2フレーム間という時間的にローカルな部分では，時間的な動き情報は保持していることになる．

そこで，ある時間幅について局所空間的な領域を考えフロー成分の1次と2次のモーメント（すなわち平均と分散）を計算する．いま，ある表情サンプルについて大きさ $M \times N$ 画素で T フレームの動画像に対して計算したオプティカルフローデータがあるとする．これを $m \times n$ 個の $r \times r$ 画素の小領域に分割する．時刻 t，位置 (x, y) におけるフローの成分を $u_t(x, y), v_t(x, y)$ とするとき，基本的な特徴量を次式で計算する．

$$\mu_{u,i,j} = \frac{1}{Tr^2} \sum_t \sum_{(x,y)} u_t(x, y) \tag{3.9}$$

$$\mu_{v,i,j} = \frac{1}{Tr^2}\sum_t \sum_{(x,y)} v_t(x,y) \tag{3.10}$$

$$\sigma_{uu,i,j} = \frac{1}{Tr^2}\sum_t \sum_{(x,y)} (u_t(x,y) - \mu_{u,i,j})^2 \tag{3.11}$$

$$\sigma_{uv,i,j} = \frac{1}{Tr^2}\sum_t \sum_{(x,y)} \{(u_t(x,y) - \mu_{u,i,j})(v_t(x,y) - \mu_{v,i,j})\} \tag{3.12}$$

$$\sigma_{vv,i,j} = \frac{1}{Tr^2}\sum_t \sum_{(x,y)} (v_t(x,y) - \mu_{v,i,j})^2 \tag{3.13}$$

ここで，$0 < t \leq T$, $(x,y) \in R(i,j)$ で，$\mu_{u,i,j}$ と $\mu_{v,i,j}$ はそれぞれ，オプティカルフローの水平垂直各方向成分の (i,j) 番目の小領域 $R(i,j)$ における平均である．同じく，$\sigma_{uu,i,j}$, $\sigma_{uv,i,j}$ および $\sigma_{vv,i,j}$ はフローの分散である．これらを使うと $K(=5mn)$ 次元のベクトルができる．いま，そのベクトルを

$$\begin{aligned}\mathbf{F} &= \{\mu_{u,1,1}, \mu_{v,1,1}, \sigma_{uu,1,1}, \sigma_{uv,1,1}, \sigma_{vv,1,1}, \cdots \\ &\quad \mu_{u,i,j}, \mu_{v,i,j}, \sigma_{uu,i,j}, \sigma_{uv,i,j}, \sigma_{vv,i,j}, \cdots \\ &\quad \mu_{u,n,m}, \mu_{v,n,m}, \sigma_{uu,n,m}, \sigma_{uv,n,m}, \sigma_{vv,n,m}\} \\ &= \{f_1, f_2, ..., f_k, ..., f_K\}\end{aligned} \tag{3.14}$$

と表すことにする．

この次元を落とすために，$\mathbf{F}$ の各要素 f_k について Fisher の判別基準[Duda and Hart, 1973, sec. 4.11]にならって，次の評価関数を計算し，評価量が大きくなるものを識別特徴パラメータとする．

$$J(k) = \frac{var_B(k)}{var_W(k)} \tag{3.15}$$

ただし，$var_W(k)$, $var_B(k)$ はそれぞれ k 番目のベクトル要素 f_k のクラス内分散とクラス間分散であり，次式で計算する．

$$var_W(k) = \sum_{i=1}^{c}(\frac{1}{n_i}\sum_{f_k \in \theta_i}(f_k - \bar{f_{k,i}})^2), \tag{3.16}$$

$$var_B(k) = \sum_{i=1}^{c}(\bar{f_{k,i}} - \bar{f_k})^2. \tag{3.17}$$

なお，c は認識しようとするクラスの数であり，たとえば“笑い，悲しみ，怒り，驚き”を認識するときには $c = 4$ とする．θ_i は i 番目のクラスの学習用サンプルから計算したベクトル要素の集合を示す．また，$\bar{f_{k,i}}$ は クラス i 中の学習サンプルに対するベクトル要素 f_k の平均値であり，$\bar{f_k}$ は学習サンプル全体の平均値である．以上計算した $J(k)$ を大きい順に並べ，上位 $\hat{K}$ 個のベクトル要素を選び，特徴ベクトル $\hat{\mathbf{F}}$ を決定する．

$$\hat{\mathbf{F}} = \{f_1, ..., f_{\hat{K}}\} \tag{3.18}$$

なお，ここでは Fisher の判別基準を，判別関数を作るためではなく，特徴空間の次元数を削減するために用いた．

図 3.9: 表情動画像の学習サンプル

3.5.3 k- 最近傍法による特徴識別

パターン分類のために，ここでは簡単な k- 最近傍法を用いる．実験では $k = 3$ として，3 つの最近傍になるラベルから多数決で識別を行った．従って，同時に 3 つの異なるラベルが得られたときはリジェクトする．学習サンプルの特徴ベクトル $\hat{\mathbf{F}}_n^p$ とテストサンプルの特徴ベクトル $\hat{\mathbf{F}}_n^q$ の距離はユークリッド距離で計算する．

$$D(p,q) = \min_{p \neq q} \|\hat{\mathbf{F}}_n^p - \hat{\mathbf{F}}_n^q\|^2 \tag{3.19}$$

ただし $\hat{\mathbf{F}}_n$ は $\hat{\mathbf{F}}$ を正規化した特徴ベクトルで，学習サンプルから得られるすべての $f_{\hat{k}}^p$ が正規分布 $p(f_{\hat{k}}^p) \sim N(0,1)$ になっている．テストサンプルは学習サンプルの正規化パラメータで正規化をしておく．

3.6 感情識別の実験結果

3.6.1 データ収集

Ekman らの表情作成手順[Ekman and Friesen, 1975]に従って著者自身が表情を作り[6]それを撮影した顔表情画像の実験データで実験を行った．学習用のサンプルデータを 20 系列とテストサンプルを 30 系列用意した．サンプルデータは幸福，怒り，嫌悪，驚きの 4 種類の感情についてそれぞれ 5 個の系列を用意した (図 3.9にサンプルデータの一例を示す)．テストサンプルも上記の 4 種類について別に作成した．前節で述べた方法に従って，学習用サンプルから特徴ベクトルを決定し，その後テストサンプルで認識実験を行った．オプティカルフロー場は $m = 16$(列), $n = 15$(行) の 16×16 画素の正方小領域に分割して式 (3.9)-(3.13) に示した統計パラメータを計算した．

表情動画像はディジタルビデオレコーダを使って収集してオフラインでファイル化しておいて，後に処理に使用した．被験者の頭部は固定して，グローバルな動きが表情に関するフロー抽出に影響を及ぼさないようにした．通常の固定照明を顔面に当てて撮影を行った．ビデオ信号はインタレースしているので，半分の大きさの画像にサブサンプリングを行った画像を最終的な処理画像とした．すなわち，それぞれの表情画像について $M = 256$, $N = 240$ の大きさで $T = 30$ フレーム分の長さを取り出した．動画像中の表情開始終了点を指示する作業は手作業で行った（これはフローのゼロ値を頼りに，将来は自動化することも可能である）．

3.6.2 特徴ベクトルの決定（学習）

20 の学習用のサンプルデータから計算した統計量ベクトル $\mathbf{F}$ から，特徴ベクトル $\hat{\mathbf{F}}$ を求めた．ベクトルの次元はサンプル数を越えない数として $\hat{K} = 15$ とした．結果としては，特徴ベクトルの要素となったのは，口のまわりの μ_u, μ_v と，額の μ_v および，1 箇所のみ口のまわりに分散パラメータ σ_{uu} が特徴ベクトルの成分として得られた (図 3.10参照)．

この特徴ベクトルを使って，学習サンプル自身で識別 (Leave-one-out 法) を行った．その結果 20 サンプル中の 19 は正しく識別され，1 サンプルはリジェクトされた．従って，識別面はほぼ良好に形成されていると推定できる．

3.6.3 識別結果

これまでに述べた方法に従って，テストサンプルの識別を行った．その識別率の評価をここでは 2 つの方法で行った．テストサンプルを 2 つの方法で主観評価分類し，識別率を計算した．すなわち，各テストサンプルのうち表情を十分に表していると思われる瞬間のフレームを写真にとり，10 人（男 5, 女 5 ）の主観評価者に表情から読み取れる感情を 2 回にわたって分類してもらった．1 回は「幸福，怒り，驚き，嫌悪」の 4 つの言葉に分類し，あとの 1 回は，それに「わからない（不明）」を加えて 5 つに分類してもらった．6 人以上の評価者の分類が一致したサンプルはそのとおり分類し，評価が分か

[6]前にも述べたように，表情を作るのは簡単ではない．訓練された俳優や FACS を熟知して各 AU を再現できる人に実験に協力してもらうのが最適であろう．今回は著者自身が FACS に関する経験に基づいて表情データを作成した．

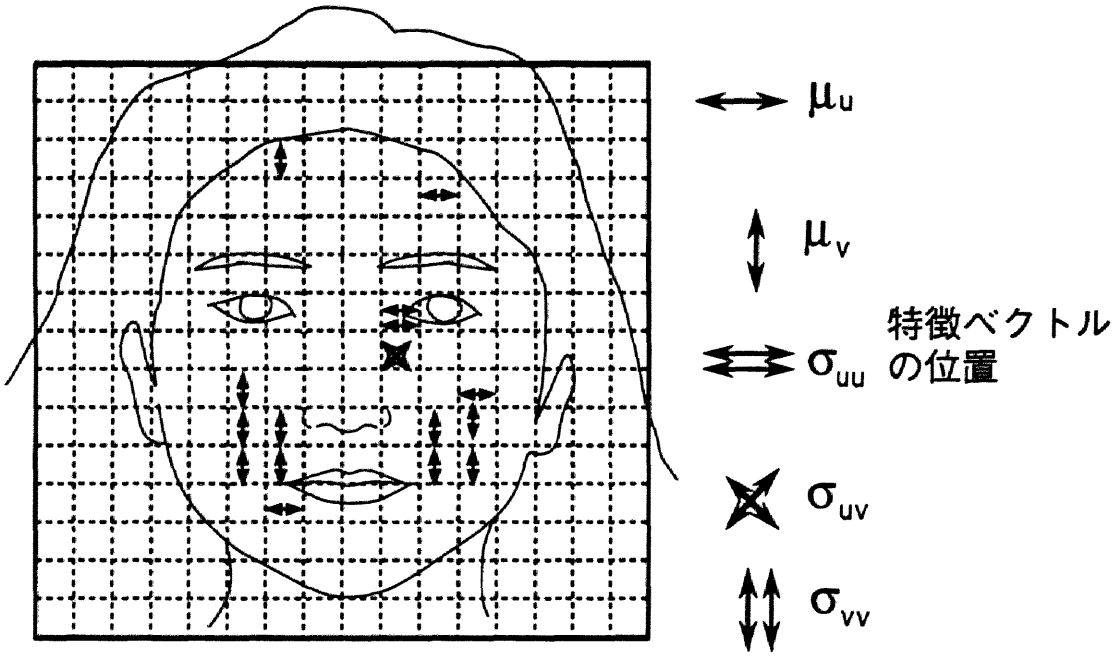


図 3.10: 特徴ベクトルを構成する点

図 3.11: 識別実験に用いた表情動画像の例

れるものはやはり「不明」とした．従って，テストサンプルはクラス間で一様に分布していない．図3.11はテストサンプルの1部の画像のスナップショットである．

表3.3は認識結果を示す．表3.3(a)は4クラスに分類したときの結果であり，表3.3(b)は主観評価が「不明」を含む5クラスの場合の識別結果である．(a)の識別率が80%(24/30)であるのに対し，(b)の場合の識別率86%(19/22)のほうが高い．これは，人間が主観的に判断して識別できる表情ならば，計算機にも識別がしやすいということが言える．

3.7 むすび

表情からのメッセージ抽出の方法としてオプティカルフローを利用した表情認識の手法について検討した．すなわち，フローの抽出による表情筋の動作推定方法と結果および統計的手法による感情識別の実験について報告した．

表情筋の動作推定には，表情を作る主な筋肉の場所に検出窓を置き，その窓内のフローのデータから筋肉の動きを推定する手法を示した．実際の表情画像について実験を行い，その表情を作るための筋肉について動きを検出できることが確かめられた．

今後は，窓の設定場所の検討と，表情認識のアルゴリズムの具体化が主な課題である．ここでは窓を12個設定したが，色々な表情でフローをとって配置場所と個数の最適化をはかる必要がある．たとえば，両目の眉のあたりにも情報があるのでその動きを取る必要がある．また，オプティカルフローのアルゴリズムにも改良の余地がある．

おおきな問題としては，頭全体の揺れの影響がある．全体の動きを補正しながら安定して局所的な動き（＝変形）を検出できるアルゴリズムが必要である．特に，勾配法によるオプティカルフローのアルゴリズムは輝度のグラディエント方向にのみ情報が存在するので，全体の動きを単純に差し引くという処理では解決できない．これは柔物体の運動の検出の問題として，解決しなければならない．

この手法では，1枚の画像からそれがどんな表情をしているかを認識することはできない．そのためには，しわの有無や形状の解析をするアプローチをとる必要がある．逆に非常に微妙な動きもとらえることができるので，複雑な表情，たとえば一瞬驚いたのち笑うといった表情もとらえられる可能性がある．あるいは，偽りの表情に対しても，一瞬みせる正直な表情を捕捉することができるかもしれない．これらは，心理学や精神分析学の分野に貢献すると考えられる．

後半ではオーソドックスなパターン分類の手法を表情画像のオプティカルフローデータに適用して感情の識別を試みた．学習サンプルのオプティカルフローの成分から特徴ベクトルを決定して，表情画像から有意な場所と特徴量を抽出した．1人の人間の表情変化に対する少数データではあるが，識別実験により80%以上の識別率を得ることができ，表情動画像からの感情の自動識別の可能性を示した．人によってどのように特徴パラメータが変化するか，またほかの感情を導入するとどう変化するかなど，今後の課題は多い．

応用面では，表情のイメージャとしてのメッセージを如何にマンマシンインタフェースの領域で利用できるかを考える必要がある．また表情の測定や理解は，マンマシンインタフェースの向上だけでな

く，モデルベース画像符号化，キャラクタアニメーション制作への応用も期待できる．さらに，各種モニタリングシステムにおいて対象に対する興味の度合を測る手段を与えることが可能となる．

表 3.3: 感情認識結果

(a) 4 クラス (幸福, 怒り, 驚き, 嫌悪) の場合

	計算機の識別結果				
主観評価による分類	幸福 happiness	怒り anger	驚き surprize	嫌悪 disgust	合計
幸福	7	—	—	—	7
怒り	—	4	1	1	6
驚き	—	2	9	—	11
嫌悪	1	1	—	4	6
合計	8	7	10	5	30

(b) 5 クラス (幸福, 怒り, 驚き, 嫌悪, 不明) の場合

	計算機の識別結果				
主観評価による分類	幸福 happiness	怒り anger	驚き surprize	嫌悪 disgust	合計
幸福	7	—	—	—	7
怒り	—	4	—	1	5
驚き	—	2	5	—	7
嫌悪	—	—	—	3	3
小計	7	6	5	4	22
(不明)	(1)	(1)	(5)	(1)	(8)
(合計)	(8)	(7)	(10)	(5)	(30)

第 4 章

発話動作の認識 — リップリーダ —

4.1 はじめに

顔の表情は感情の表出結果であるが、発話時の調音動作によっても顔の表情は変化する．すなわち、人間は口腔、鼻腔、舌、口唇などで構成される声道を、筋肉を動かして変形させることによって調音を行っているため、口唇と顎を動かす筋肉によって、顔の表情にも変化をきたす．そこで、その表情の変化をもとに話している言葉をある程度理解することが可能となる．読唇術（Lipreading）と呼ばれる技術がこれである．本章では、前章で述べたオプティカルフローに基づく表情変化の抽出法を、口の動きによる発話単語の認識という問題に特化して適用する．

以下、第 2 節では、計算機による読唇の研究を調べ、解決すべき問題点とここで明らかにしようとしている課題を明確にする．第 3 節は口唇の動きを抽出するための筋肉モデルを検討し、オプティカルフローに基づく特徴ベクトルを決定する．第 4 節では単語認識実験システムを構築する．第 5 節は連続発声した英数字の認識実験を行い結果を示す．第 6 節は考察とまとめを行う．

4.2 計算機による読唇の研究

読唇情報から得られるのはおもに調音位置に関する情報[Fukuda and Hiki, 1982]なので，文脈なしではすべての言葉を読唇によって理解することは難しい．しかしながら，音声信号に基づく言葉の認識（以下，音声認識と呼ぶ）と補いあうところはおおい．例えば鼻音の /n/,/m/,/ng/ は音声認識では識別しにくい音であるが，口唇の形は非常に異なる．また工場内や車内などの雑音環境において音声認識が困難であるときでも，視覚情報には影響がない[1]．このように，人間は知らないうちに聴覚障害のあるなしにかかわらず読唇をして会話をおこなっている．

計算機とのマンマシンインタフェースにおいても，視覚情報と音声情報の両方をうまく結合すると，計算機の言語理解の能力を上げることが出来る．Petajan[Petajan, 1984] はこれに着目して，音声認識ボードの認識候補出力に口形の画像処理結果を組み合わせて，認識率向上をはかった．しかしながら，

[1] 最近，これらが相補的ではなくむしろ融合することがあるという報告もある[McGurk and J.McDonald, 1976; 積山, 東倉, 1989]

彼の実験では口腔の暗い部分と皮膚、歯、舌などを2値化処理により区別し口腔部の形状を解析するため、顔を特殊な照明・撮影条件におかなければならなかった.

そのほかにも松岡ら[松岡ほか, 1986]、栗田ら[栗田ほか, 1988]、Petajan ら[Petajan and Bodoff, 1988]、Finn &Montgomery[Finn and Montgomery, 1988]、田村ら[田村ほか, 1989]により画像処理による読唇の報告がなされている. これらのほとんどは、口形あるいは口唇輪郭を解析し読唇を試みているが、口唇周りの画像は濃度変化がゆるやかで、これらの形状を正確に求めるのは本質的に困難である. そのため、松岡ら[松岡ほか, 1986]は唇に黒い口紅をつけて口唇輪郭の抽出を容易にしている. また、田村ら[田村ほか, 1989]はスプラインモデルのフィッティングにより、より正確な口唇輪郭を求めようとした. このように、読唇では映像信号からいかに必要な情報を取り出すかの工夫が重要な課題となっている. なお発声学の分野ではこれらより以前に、唇の形状が調音とどうかかわるかを調べるために、ストロボスコープなどを使って発声中の連続写真をとった[Fukuda and Hiki, 1982; Fujimura, 1961]り、LED 発光素子を唇の上下左右につけるなどして、手作業で口形を計測、解析する研究が多数おこなわれている.

対象が映像信号になるという違いはあるが、計算機による音声認識で解決すべき問題が、視覚情報をもちいる読唇でも同様にあてはまる. すなわち、

(i) 認識に必要な情報の確実な収集

(ii) 連続発声の単語認識

(iii) 不特定話者の認識

(iv) 大語いの認識

などである. (ii)-(iv) の問題は読唇情報を音声認識の補助として考えた場合は、重要とはならないが、騒音環境のように読唇のみでまたは読唇情報を主として言語認識を行う時にはさけられない. ところがこれまでの報告では、(i) の問題を解決するためにいかに唇の形状を正確に取り出すかに注目しており、(ii)-(iv) の問題を扱ったものはない.

そこで本論文はこれらの問題のうち特に (i)-(iii) を解決するために、口唇の形状ではなく、口唇周りの動きに注目して読唇を試みる[Mase and Pentland, 1989]. 調音のもとになる筋肉の動きが表情を変化させ、その変化をたよりに読唇が行われているのであるから、筋肉の動きを捉えることができれば、その動きのパターンによって、発声している言葉を識別できるはずである. また、筋肉の動きは発声を行うための神経インパルスに直接関係しており、話者が変わっても一定であると推定される. そこで、口唇周りのオプティカルフローを求め、特徴パターンとなる速度を検討し、連続発声の単語（英数字）の識別実験を行った. オプティカルフローを使うことの利点は、

1. 人間の視覚は変化する照明条件のなかでも動きには敏感であり、読唇の場合も動きに注目している. オプティカルフローを直接求めることにより、口唇形状の抽出という問題から逃れられる.
2. 単語の切れ目には一瞬の動作の停止がある. すなわち単語の切れ目は口唇の速度がゼロとなり、速度を特徴量とすることにより、連続発声した単語列を分解できる. 形状解析では、分節は非常に難しい.

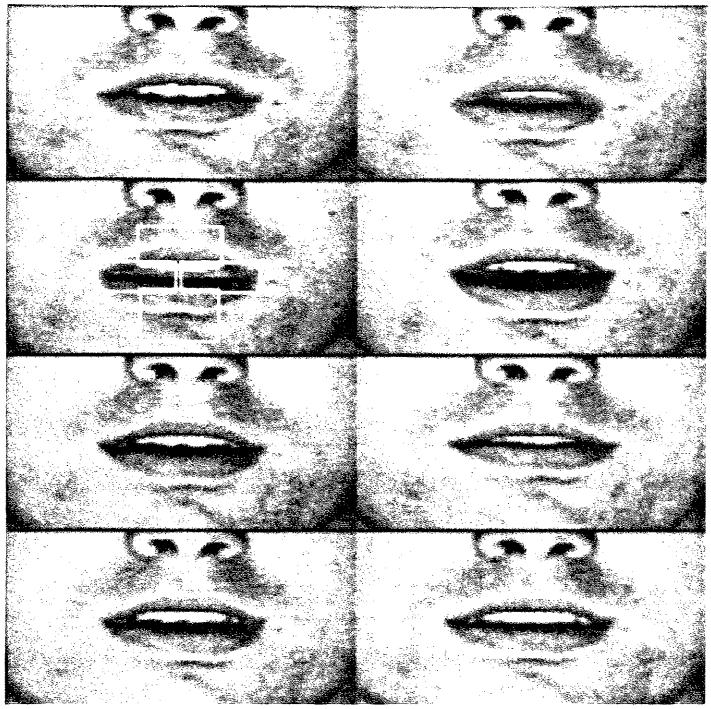


図 4.1: 入力画像と動き検出窓

3. 生理学的にみて、同じ単語を発声するためには、話者が変わっても同じ様に筋肉を動かしていると考えられる．口唇形状の個人差や髭の有無の影響を受けない特徴量を得られる．

などが、考えられる．

以下、本文ではオプティカルフローに基づく特徴を求め単語認識を行うシステムを示す．また、そのシステムで英語の数字列の発声に対して行った予備実験の結果の詳細を示し、上に示した仮説を検証する．

4.3 筋肉モデル

4.3.1 発声に関わる筋肉

FACS のマニュアルには、ある表情をアクションユニットで記述するための指針が書かれているが、その中で、口唇周りの形状を記述する言葉として、8 つの特徴が挙げられている．すなわち、口形の延伸 (elongate), 口形の収縮 (de-elongate), 口唇の細め (narrow), 口唇の太め (widen), 薄い口唇 (flatten), 口唇の突出 (protrude), 口唇の緊張 (tighten), 口唇の引延 (stretch) となっている．この内発声時に大きく変化するのは口形の延伸と収縮 (elongate / de-elongate) である．突き出し (protrude) も発声時の特徴としては重要な情報を含んでいる[Perkell, 1986]が、ここでは正面からの視覚情報だけを考慮するため直接的な特徴量としては使えない．しかし、突き出しは「すぼめ」となって、収縮と似た変形をおこす．

そのほかに、下顎の上下による口唇の開閉も発声時には重要な動作である．従って、口唇の上下方向の開閉、左右方向の伸縮が発声に関わる主な 2 つの動きと考えられる．

4.3.2 口唇画像のオプティカルフロー

頬筋、口輪筋および顎の動きを捉えるために口唇の上下左右に窓を設定し、各窓内のオプティカルフローの平均値を計算し代表的な動きデータを測定した．図 4.1は口唇画像の 1 例で、窓が白枠で表示してある．図 4.2(a) は口を閉じる時点のオプティカルフローを矢印で表示してある．下唇，下顎，口唇両側に大きな動きが見られる．また，同図 (b) は口を横に開く時点であることがオプティカルフローから読みとることができる．

各窓の (x, y)- 軸方向の速度成分 (u, v) を用いて、8 次元の特徴量で各時刻の動きを記述できる．図 4.3は学習サンプル (/one/ ～ /four/) を発声したときの各成分の時間変化をプロットしたものである．即ち時刻 i の特徴ベクトル $\mathbf{x}^i$ は [上、下、左、右] の各窓の平均速度成分 $[(u_a(i), v_a(i)), (u_b(i), v_b(i)), (u_l(i), v_l(i)), (u_r(i), v_r(i))]$ を要素とする：

$$\mathbf{x}^i = \begin{pmatrix} u_a(i) & v_a(i) & u_b(i) & v_b(i) & u_l(i) & v_l(i) & u_r(i) & v_r(i) \end{pmatrix} \tag{4.1}$$

これらの成分は互いに独立ではないので、学習サンプルに対して主成分分析を行い、真の特徴ベクトルを求めた．

即ち、上記の $\mathbf{x}^i$ の分散行列 S を求め、その固有ベクトルを調べた．

$$S = \sum_{i=1}^{N} (\mathbf{x}^i - \overline{\mathbf{x}})^T (\mathbf{x}^i - \overline{\mathbf{x}}) \tag{4.2}$$

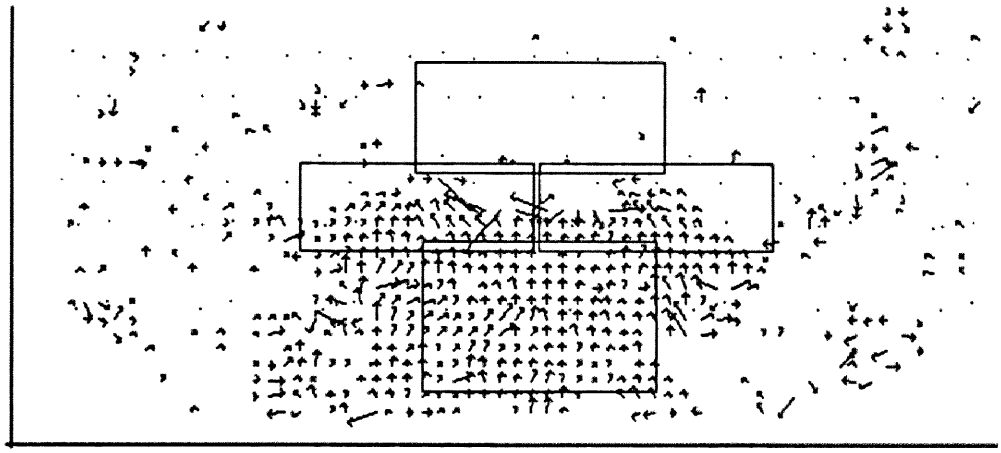
ここで、N は学習サンプル全体のフレーム数 ($N \approx 700$)、$\overline{\mathbf{x}}$ は $\mathbf{x}$ の平均値、$(\bullet)^T$ は転置行列である．

その結果、S の固有ベクトルは、

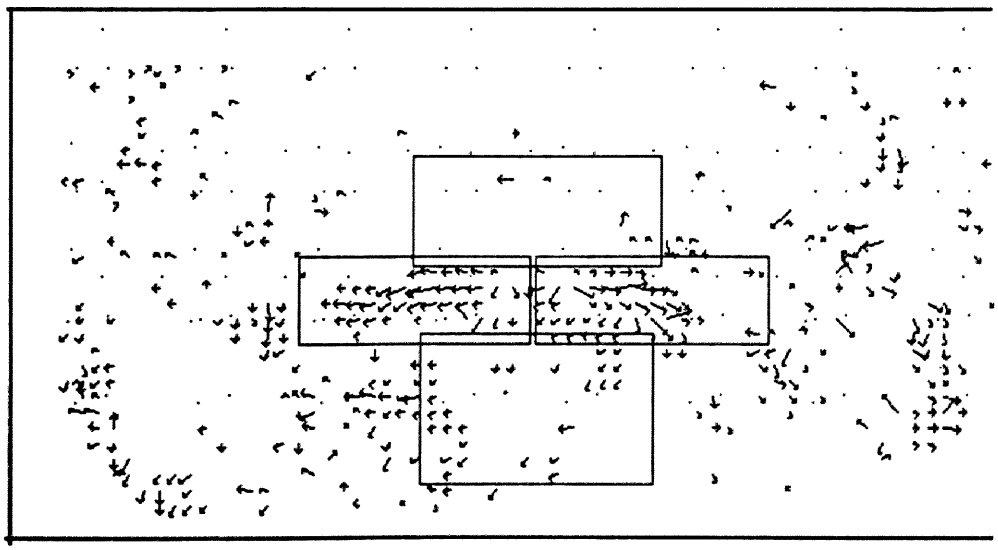
$$\begin{pmatrix} \underline{e_1} \\ \underline{e_2} \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{pmatrix} = \begin{pmatrix} -0.51 & 0.00 & 0.74 & \underline{0.94} & 0.47 & \underline{1.00} & 0.02 & \underline{1.00} \\ 0.20 & \underline{1.00} & 0.08 & 0.13 & \underline{-0.83} & 0.08 & \underline{1.00} & 0.01 \\ 1.00 & 0.02 & 0.60 & -0.48 & 0.41 & -0.35 & 0.17 & -0.35 \\ -0.58 & -0.23 & 1.00 & -0.14 & -0.33 & -0.10 & -0.02 & -0.06 \\ -0.06 & 1.00 & 0.19 & -0.31 & 0.01 & 0.06 & -0.96 & -0.12 \\ -0.61 & 0.44 & 0.05 & -0.04 & 1.00 & -0.02 & 0.51 & -0.15 \\ 0.05 & 0.16 & 0.13 & 1.00 & 0.03 & -0.66 & -0.23 & -0.42 \\ -0.02 & 0.01 & -0.04 & -0.12 & 0.05 & -0.87 & 0.12 & 1.00 \end{pmatrix} \tag{4.3}$$

となり、またその固有値は、

$$\begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 & \lambda_5 & \lambda_6 & \lambda_7 & \lambda_8 \end{pmatrix}^T = \begin{pmatrix} 3.57 & 2.41 & 1.23 & 0.45 & 0.19 & 0.14 & 0.03 & 0.01 \end{pmatrix}^T$$



(a) Closing lips



(b) Elongating lips

図 4.2: 口唇まわりのオプティカルフロー

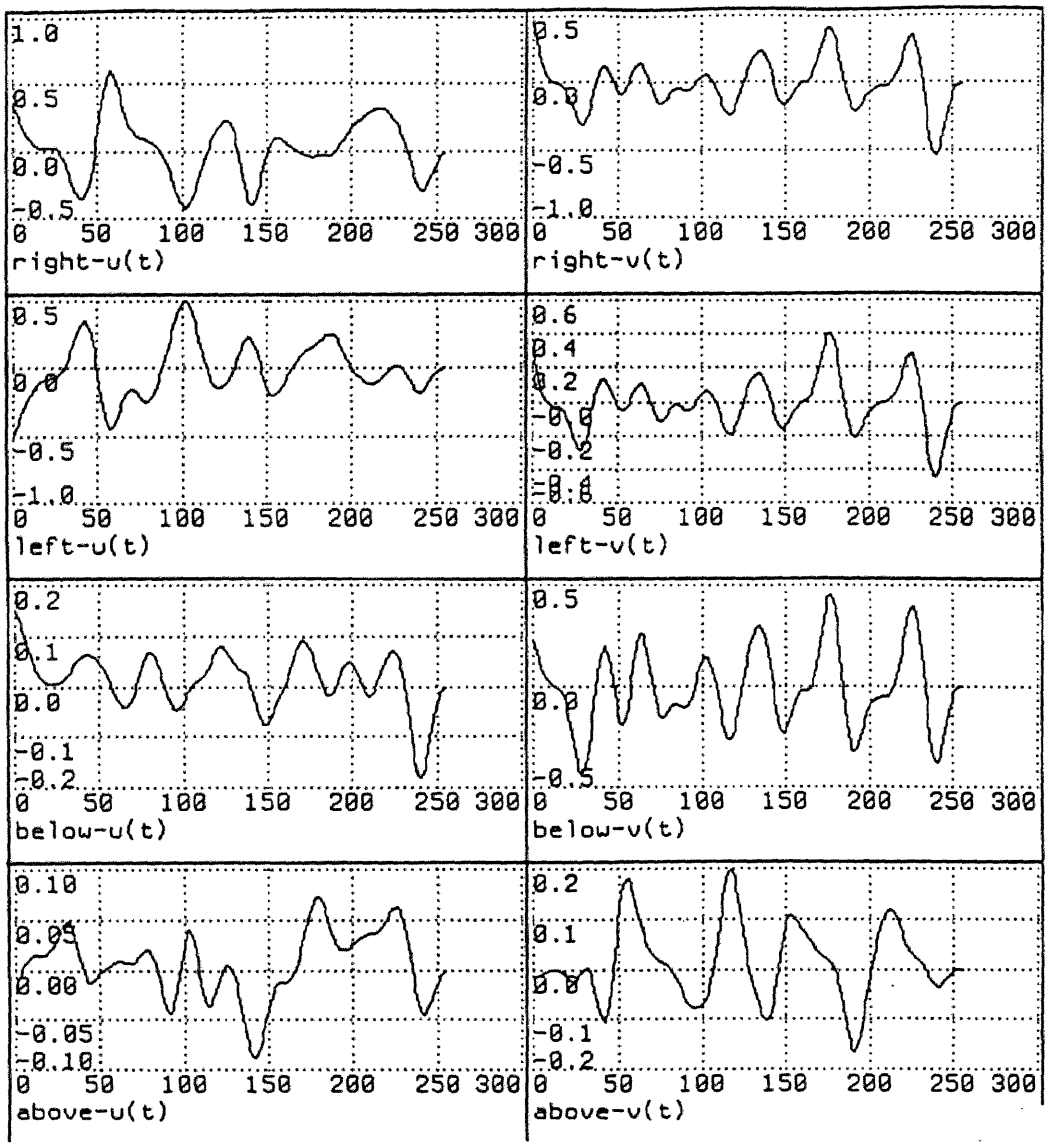
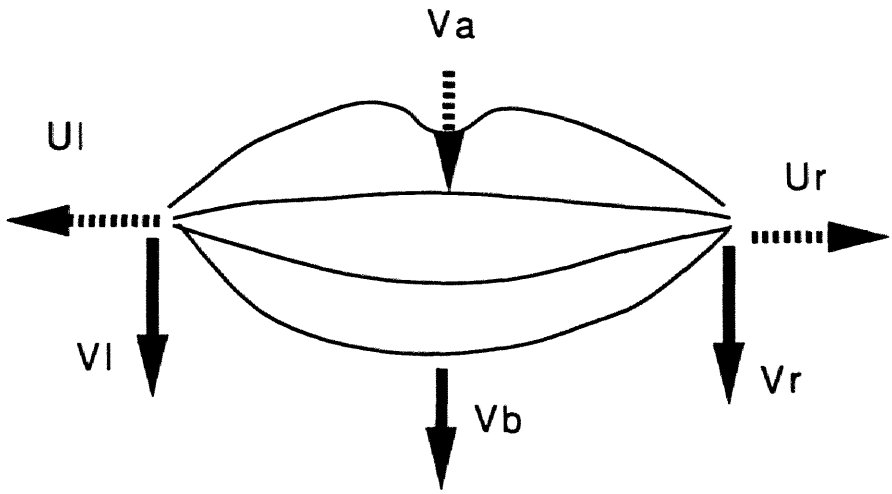


図 4.3: 窓内の速度成分の平均値による動きデータ



$$O(t) = v_b(t) + v_l(t) + v_r(t)$$
$$E(t) = \kappa v_a(t) - u_l(t) + u_r(t)$$

図 4.4: 特徴量 $O(t)$ と $E(t)$

となった．第 1 の固有値が全体の約 45%、第 2 が約 30% をしめ、これら 2 つを主成分ベクトルとして見なすことが出来る．そこで、第 1、2 固有ベクトル $\mathbf{e_1}, \mathbf{e_2}$ の各成分のうち、絶対値が 1.0 に近いものだけ (式 4.3の下線部) を残すように量子化すると、2 つの特徴量をえられる：

$$O(t) = v_b + v_l + v_r \tag{4.4}$$

$$E(t) = \kappa v_a - u_l + u_r \tag{4.5}$$

ここで、κ は速度成分のアスペクト比である．これを図示すると、図 4.4のようになり、$O(t)$ は顎の上下による口の開閉、$E(t)$ は口唇の伸縮に対応することが判る．これは、オプティカルフローによる特徴抽出が筋肉の動作による変形を捉えるのに適していることを示す．

4.4 英数字認識システム

4.4.1 画像入力と前処理

英数字読唇システムの構成図を図 4.5に示す．動画像データは CCD カメラを使って画像処理装置 (Detacube) でディジタイズする．本装置は最大 256 枚の 128 × 64 画素 (8bit/ 画素) の大きさの画像を 1 度 (60 枚 / 秒) に入力することが出来る．入力画像に対し、ノイズ除去のため時空間ガウスフィルタをかける．60Hz で入力することによって、フレーム間の口唇の動きは充分小さくなり、簡単なオプティカルフロー抽出アルゴリズムが適用できる．前述の通り、Horn-Schunck[Horn and Schunck, 1981]の方法で各点の動きベクトルを推定する．窓を設定して、マクロな動きをとるので、推定のための繰り返し計算は 3 回で打ち切ることにした．結論からいって、繰り返しの回数はこの程度で充分である．文献[Horn and Schunck, 1981]には、時刻 t の結果を使って時刻 $t+1$ のオプティカルフローを計算する手法も示されているが、舌や歯の出現などによる動きの不連続な場合も想定して、オプティカル

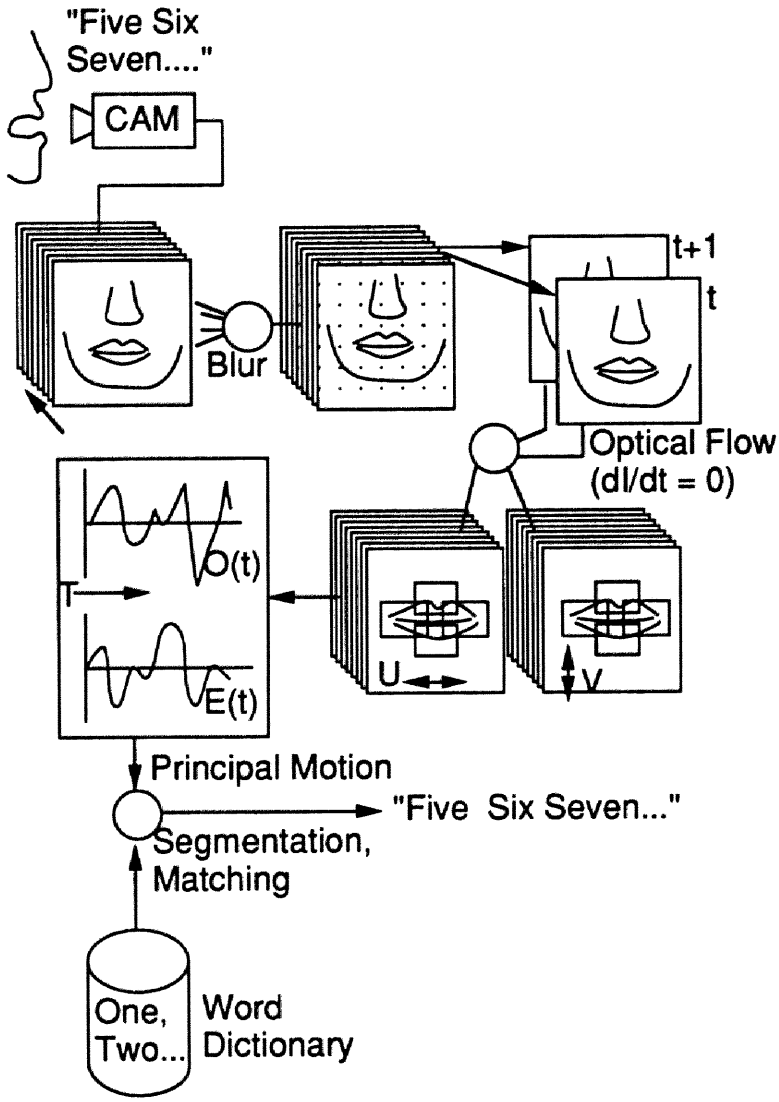


図 4.5: オプティカルフローを用いた読唇システム

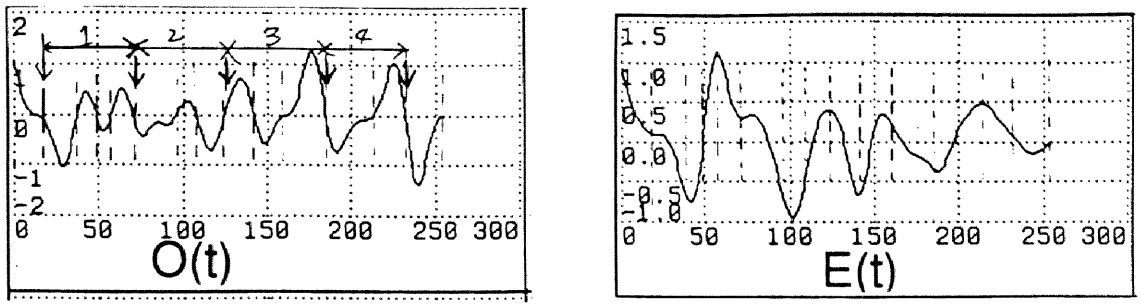


図 4.6: /one/ から /four/ までの数字の連続発声の時の $O(t)$ と $E(t)$
$O(t)$ のゼロクロッシングが破線で示してある.
正確な単語の境界は矢印で示す. これから、テンプレートを作る.
(図 4.3に対応する)

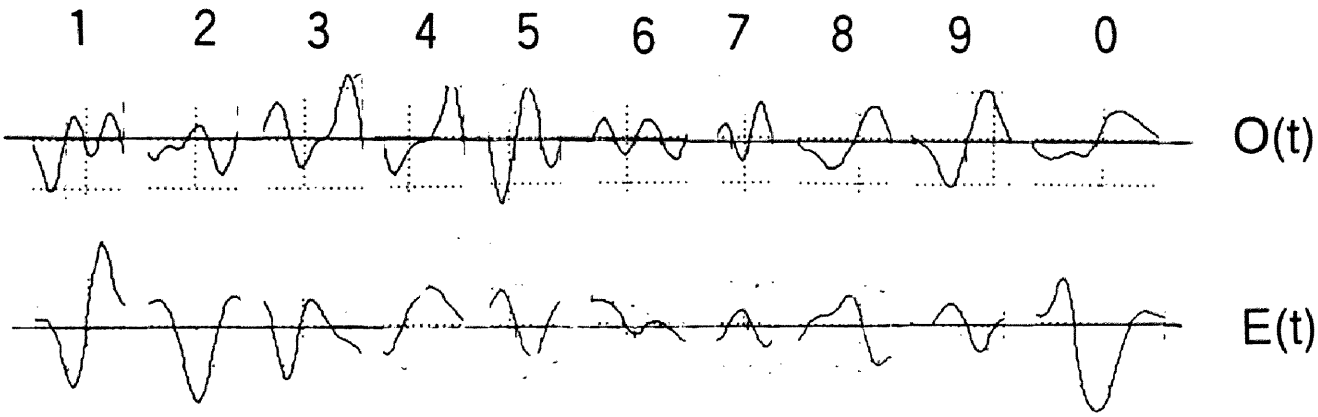


図 4.7: 認識辞書として用いたテンプレートパターン

フローの計算は隣接する 2 フレームのみ使い、時間に対して独立におこなった. 特徴量 $O(t)$, $E(t)$ を (4.4),(4.5) 式に基づいて計算し単語のマッチングを行う. 図 4.6は、学習データの /one/ から /four/ の発声の際の特徴量をプロットしたものである.

4.4.2 単語検出とマッチング

図 4.6に示す特徴量と原画像を見比べると、確かに単語の境界で $O(t)$ がゼロになっている. $O(t) = 0$ となるのは、図 4.6に示すように単語の境界以外にもあり、これは口唇の開閉が逆方向に動いたり、止まる時が単語境界以外にもあることを示す. これを使うと、音声信号でいう音素 (phoneme) に対応する視覚情報のプリミティブ (1 種の Visime であるが、文献[Massaro, 1987, p.36]の定義とは異なる) を隣接する $O(t) = 0$ で決まる区間で定義でき、1 つの単語は複数のプリミティブから構成されるセグメントであると考えられる. 従って、学習データから単語セグメントの辞書を作成しておいて、対象とする実験データから 1 つ以上のプリミティブ列でセグメントを作って辞書とのマッチングを順に行うと、連続した単語識別が可能となる. ここでは、$O(t)$ と $E(t)$ の波形を単語別に 16 点で標本化して、時間伸縮して辞書を作成した. 図 4.7は時間伸縮前の /zero/ ~ /nine/ のテンプレートパターンで

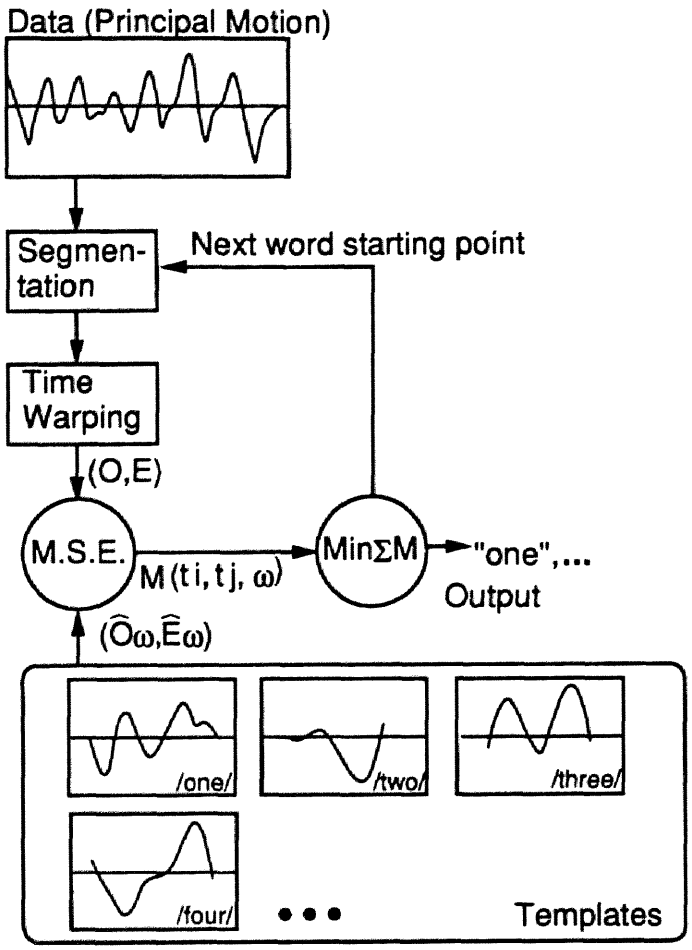


図 4.8: 単語のマッチング手順

ある.

マッチングは、実験データから得られる特徴量から、単語セグメントとなる区間を仮定し、区間内ををそれぞれ 16 点で標本化して時間軸の伸縮を行い、マッチング尺度 M を計算する. 時刻t_i から t_j の区間を単語 w と対応づけたときのマッチング尺度 $M(t_i, t_j, w)$ は辞書（単語 w ）の特徴パターン $\hat{O}_w(\tau)$, $\hat{E}_w(\tau)$ と実験データとの重み付き二乗誤差の和として計算される. すなわち、

$$M(t_i, t_j, w) = \sum_{\tau=0}^{\tau=15} (O(\tau') - \hat{O}_w(\tau))^2 + \lambda(E(\tau') - \hat{E}_w(\tau))^2 \tag{4.6}$$

ここで λ は固有値 λ_1, λ_2 の比、$\tau' = t_i + \frac{\tau}{16}(t_j - t_i), (t_i < t_j)$, $O(\tau = 0) = 0$ である.

図 4.8に単語のマッチング手順を示す. マッチングは発話の先頭から順に単語の分離と識別を同時に行う. 単語辞書とのマッチングで充分小さい $M(t_i, t_{i+1}, w)$ となるセグメント (区間 $[t_1, t_{i+1}]$) と、単語 w を捜し、そのセグメントの終点 t_{i+1} を次の単語セグメント候補の始点として、繰り返し行う. 最終的に一連の発話にたいして、マッチング尺度の総和が最小となる単語列を認識結果とする :

$$\min_{(t_0, t_1, \ldots, t_n), (w_1, w_2, \ldots, w_n)} \sum_i M(t_i, t_{i+1}, w_k) \tag{4.7}$$

4.5 実験結果

3人の被験者を対象に初期実験を行った．1人(被験者1)からは学習データと実験データの両方をとり、学習データから辞書を作成した．この辞書をすべての被験者の実験データとのマッチングに用いた(図4.7参照)．被験者のの1人(被験者3)には口髭があった．

3から5個の数字（英語）をふくむ発話を複数回収集し認識実験を行った．被験者は通常の照明で約2メートルの距離から撮影し、被験者には通常の速度で発声するよう指導した．全体では6回の発話（のべ単語数=21）による実験となった．また、頭部を特に固定する装置は使用しなかったが、全体のゆれを軽減するために、後頭部を壁にもたれかけるように指導した．

実験データの詳細と認識結果を表4.1に示す．図4.9は実験に用いた全データから得られた特徴量をプロットしたものである．これから認識率を集計すると、[被験者1]—73%(11/15), [被験者2]—100%(2/2), [被験者3]—50%(2/4)となった．

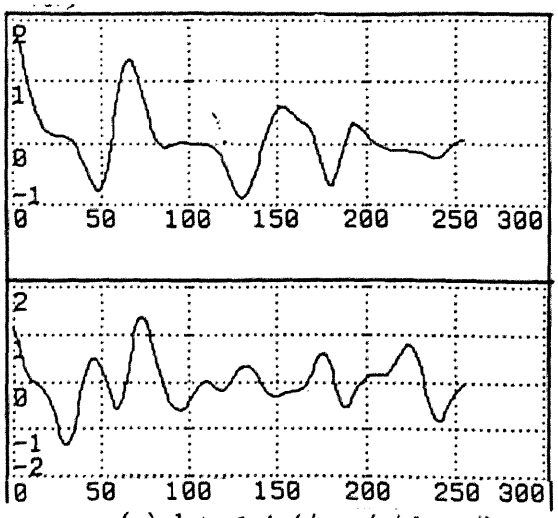
認識結果を解析すると，ほとんどの失敗は各発声の第1単語で起こっている．表からは読み取れないが、特に第1単語の開始点の識別が困難となっている．すべての区間でマッチング尺度を試算してみると正しい単語の開始点からは，実際に発話された単語の辞書パタンに対してマッチング尺度が最小となることから，音声情報などにより，単語の開始点に関する情報が提供されれば，単語の認識率をあげることができる．試算によると，被験者1では93%(14/15)，被検者3では75%(3/4)になる．

4.6 むすび

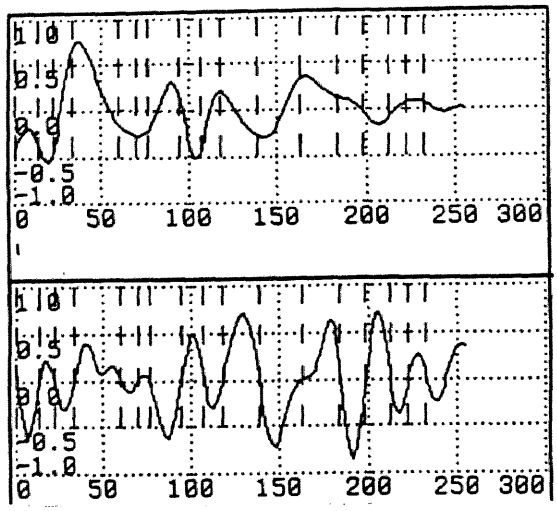
読唇の仕組みを発声学に基づいて検討し、コンピュータによる読唇の1手法を提案した．口唇の周りのオプティカルフローを解析してシステムを構築し，不特定話者，連続発声の単語認識が通常の撮影条件で可能であることを示した．文脈の無い言葉に対する人間の読唇の能力は50から70%前後であるとも言われ、それと比較して、かなり高い認識率を得られることが判った．

本手法の特に注目すべき点は，オプティカルフローを原情報とする事により連続発声した単語の境界の候補となる時点を簡単に得られることである．また、オプティカルフローを計算することにより、これまで煩わしかった口唇の形状を正確に抽出する必要が無くなった．特定の単語を発声させるための筋肉の動作の話者独立性については、発声生理学での検討が必要となるが、今後実験データを増やすことにより実験的に検証できると思われる．被験者2,3についてはごく僅かの実験データによる結果のみを示したが、被験者1の辞書パターンに対して約70%の認識率を得たことは十分可能性があるといえる．なお田村ら[田村ほか, 1989]らも、読唇のためにオプティカルフローを利用することを本研究とは独立におこなっているが、その目的はスプラインモデルのフィッティングのための補助的情報として使っており、本文で示したようなオプティカルフローの利点が生かされていない．

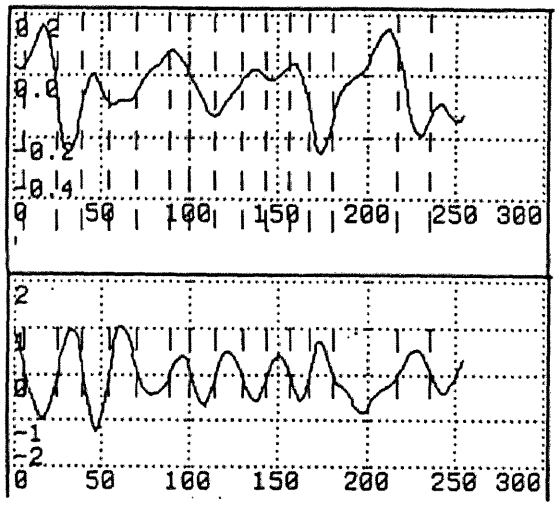
一方、本文で示した実験結果は少量のデータに対して成功しており、手法の一般性の評価については今後の実験データの追加によるところが大きい．オプティカルフローの平均化のために設定した窓は現在、会話的に各発話に対して1回、設定しているが、鼻孔の検出ができれば自動位置合わせも可能である．



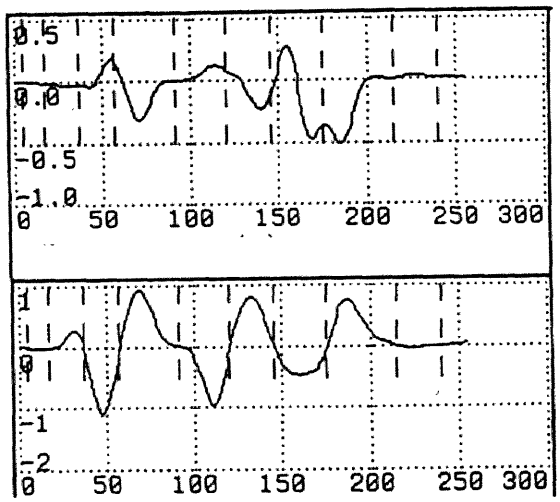
(a) data 1-A (/one/-/three/)



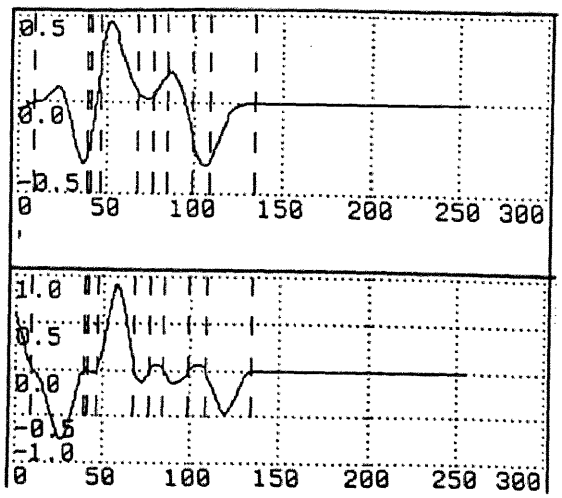
(b) data 1-B (/one/-/five/)



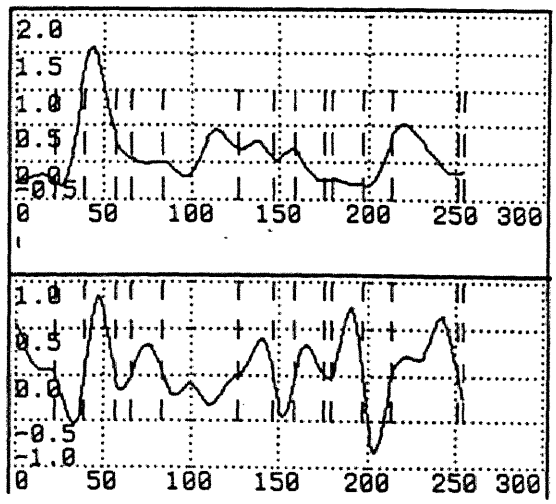
(c) data 1-C (/five/-/eight/)



(d) data 1-D (/eight/-/zero/)



(e) data 2 (/one/-/two/)



(f) data 3 (/one/-/four/)

図 4.9: 実験データの特徴量（上段：E(t)、下段：O(t)）

表 4.1: 単語認識結果：被験者 1 の辞書とのマッチング (被験者 1-[A,B,C,D],2,3)

n/f: not found, 被験者 3: 髭あり

実験データ	区間		認識結果 (w)	マッチング尺度 (M)	正解	備考
	開始点 (t_i)	終了点 (t_j)				
1-A	11	84	one	5.43	one	
	84	166	two	5.65	two	
	166	232	three	4.91	three	
1-B	33	94	six	7.61	one/two	error
	94	138	three	4.43	three	
	138	184	four	3.53	four	
	184	222	five	2.97	five	
1-C	22	88	six	8.19	five	error
	88	143	six	5.93	six	
	143	180	seven	5.24	seven	
	180	236	eight	4.30	eight	
1-D	17	91	—	—	eight	n/f
	91	145	nine	3.72	nine	
	145	215	zero	6.18	zero	
2	9	68	one	5.64	one	
	68	134	two	5.89	two	
3	—	—	—	—	one	n/f
	—	—	—	—	two	n/f
	158	197	three	8.11	three	
	197	251	four	6.19	four	

本手法で使ったマッチング処理は単純なもので、処理は高速であるが現在最適とはいえない．調音結合や発音の伸縮等についての考慮を含めることにより認識率を向上できると思われる．また辞書として用いるパターンは現在，認識データと同様に /one/ から /zero/ までを連続発声して得られたパターンを使っている．このため調音結合のある単語の認識能力については不明である．また、語いを増やすためには、窓を増やすなりして特徴空間の次元を上げる必要があると思われる．

英語を母国語としない被験者のデータも採取し、被験者 1 の辞書に対して同様の実験を行ったところ被験者 3 と同程度の結果を得た．これは、以下のことを示唆する．(1) 地域や原言語発声体系の違いにより生ずるアクセントや発声方法の違いに対しては、話者依存性がある．(2) 一方で、ある程度の類似性もある．(3) 標準的な発声方法との違いを生理学的、視覚的に示すことができ、発声法の治療、他言語の発声法の学習等に利用できる．

辞書とのマッチング方法、文脈に対する知識の応用など、自然言語理解へ向けた検討は、従来音声認識で行なわれてきた手法を取り入れることが必要となろう．実用的には、本システムを音声認識システムに組み込んで、視覚情報と聴覚情報を組み合わせる[青野, 石川, 1991]ことによって、より強固な言語認識システムを構築することが可能となろう．

第 5 章

物体フローの推定と歩行者の計数 — ピープルリーダ —

5.1 はじめに

これまでの章が 1 人の人間のパーツの動作を抽出対象としていたのに対し，この章は人間全体の動きとして流れの抽出を行なう．流れの計測には集団としての動き方，個々の行動パターンの追跡，流れの定量的な時間あたりの流量がある．ここでは，時間あたりの通過量を計算するための方法を検討する．

本章では，まず時空間画像解析に基づき，物体のフローを推定する方法を提案する[1]．ここでは物体は人物に限らず，一般の物体に適用できる方法を示す．フローの推定法を応用して人物の計数を行なう．本章の構成は以下のとおりである．

まず，2 節で時空間画像解析に関する従来の研究を概観するとともに，歩行者計数に関する動向を調べる．

次に 3 節では時空間中の非エピポーラ面画像の性質をしらべ，直交 2 断面を使った物体フローの推定法を示す．

4 節では，グラフィックスで発生した物体でフロー推定シミュレーションを行い，本手法の有効性を確認し，さらに実際の歩行者画像を使って推定ができることを示す．

5 節は，直交 2 断面法を簡単化した直交 1 断面法で，物体フローの移動方向を判別できることを示し，これが歩行者の方向別計数に応用できることをしめす．

6 節は，直交 1 断面法を使った，歩行者の計数実験の結果を示し本手法が有効であることをしめす．

5.2 時空間画像の解析と歩行者計数

5.2.1 時空間画像解析によるフロー抽出

動画像を時空間表現する方法は，固定したシーンを移動するカメラで撮影してシーンの depth map を作るモーションステレオのための手法[Bolles *et al.*, 1987]として使われたり，オプティカルフロー計算の画像モデルとして使われたりする．とくに前者の場合には，カメラパス (path) に平行な平面でスラ

[1]ここで物体フローとは，実世界における物体速度を画像面上へ投影した見かけの速度をさす．オプティカルフローが画像中の特徴点等の流れに注目しているのに対し，物体フローは個々の物体を明確に意識している

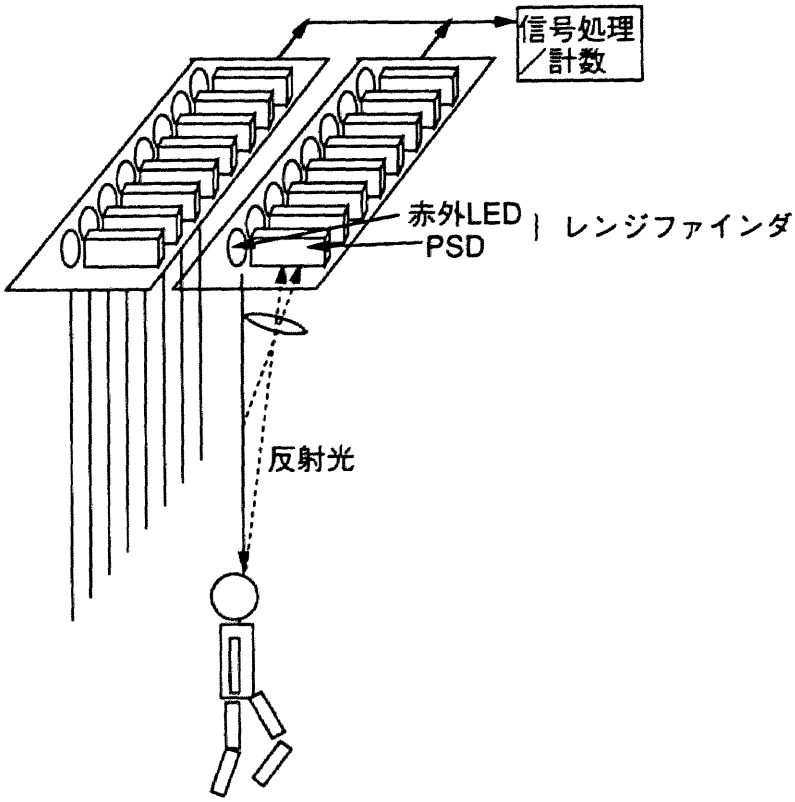


図 5.1: レンジファインダによる歩行者計数システム (市販されているシステムの構成図 (推定))

イスしたものをエピポーラ面画像とよび，その画像中の各点の軌跡の傾きからフローを計算し，透視投影の仮定に基づいて各点の深さを決定している．これは相対的にカメラを固定した際には，移動する物体のフローと平行な面でスライスしていることに相当する．本論文で提案するのは，この断面を物体のフローと角度をもたせて設定し，これを非エピポーラ面画像と呼び，その画像上の物体の幾何情報から物体フローを推定する方法である．

本報告で提案する非エピポーラ面画像による物体フローの推定法は，Zheng & Tsuji[Zheng and Tsuji, 1990]による Dynamic Projection の 1 形態と見なすことができる．Dynamic Projection とは動画像中の 1 本の直線（または曲線，サンプリングスリットと呼ぶ）上でサンプリングした 1 次元画像情報を積み重ねて得られる，時間と空間の軸をもった 2 次元画像である．このサンプリングスリットをカメラパスに対して角度を持つように設定すると，得られる像が速度と形状の情報を備えるというものである．Zheng らはガウス球を用いて，サンプリングスリットとカメラパスの一般的な場合についてシーンがどのように投影されるかを調べた．しかしながら，カメラを固定した場合の動画像の Dynamic Projection と物体の速度の関係については簡単に記述されているのみであった．本文では，動画像における Dynamic Projection を特化すると，動物体のフローが，投影像の幾何情報を使って推定できることを示す．

5.2.2 歩行者計数技術の動向

歩行人物の自動計数は，各種の催し物の入場者数や，商店街やデパートなどの交通量の計測[黄，高羽, 1983][笹間, 1988]のためにニーズが高い．現在のところ，人手によって数え上げるのがほとんどであり，自動化が望まれている．特定の会場への入退場の場合には，機械式や光電管を用いたものが使われることがある．しかしながら歩行路に何等かの制限を加えたり，設置場所が限られることや，精度が低いなどの問題があった．近年では赤外光を使った PSD(Position Sensitive Device) [2]の入手が容易になり，これを 2 次元的に配列したレンジセンサを使った歩行者計数装置が市販されるようになっている[技研,]．これは，通路の上にセンサを設置して，肩及び頭より高い領域を抽出しその動きを単純な回路で追跡し方向別に計数するというものである (図 5.1)．これらは，通過する物体の計数を効率的に行なう手法としては優れているが，物体の属性についてはなんの情報も提供できないという欠点がある．

5.2.3 画像処理による歩行者の計数

そこで TV カメラの画像から画像処理によってこれらの計数が精度よくできれば，簡便で，自然な状況での計測が可能となる．また，通過したのが人間かほかの物体かなどの判定も将来必要になり，画像による計測を行っていれば形状処理による情報が使えるであろう．

画像処理で歩行者の計数を行うためには，一般に，動画像から，(i) 人物を背景から抽出し，(ii) 個々の人物を分離し，(iii) 各フレームで人物を同定して追跡する必要がある．とくに (iii) の物体追跡は，変形している人物像を完全に同定する[浅田ほか, 1979]か，あるいは，動きの滑らかさなどをつかって速度を計算[Horn and Schunck, 1981]し追跡する必要があり，いろいろな手法が開発されているが決定的なものはない．本文では，カメラ位置を固定して撮影した動画像の時空間画像表現[Bolles *et al.*, 1987]においては，物体が連続した 1 つの領域を構成することに注目して，同定・追跡が必要ない手法を提案する．

移動物体の計数を目的とする場合に，x-t 時空間画像を用いると，通常計数する場合に前提となる物体の追跡[浅田ほか, 1979]という問題を避けることができる[山本, 1981]．すなわち，計数問題は動物体の追跡という問題から，スリット上の通過物体の計数という制約条件のもとで，x-t 時空間画像中の領域計数という問題に変換される[黄, 高羽, 1983]．これに，本文で提案するようなフロー情報を付加することによって，方向別の計数をすることも可能となる[間瀬, 1990]．このような，移動物体の計数は歩行者に限らず各方面で必要であり[笹間, 1988]，応用範囲は広い．

5.3 直交断面像による物体フロー推定

時空間画像表現は原動画像の構成によってその性質が異なる．いま目的は，シーンの 3 次元記述を得ることではなく，静止したカメラで撮影した動物体の解析である．このような動画像を時空間表現すると，図 5.2のように，背景で満たされた 3 次元の立体空間中に動物体が体積をもった一般化円筒状の領

[2]三角測量の原理で 1 点における物体までの距離を測定する

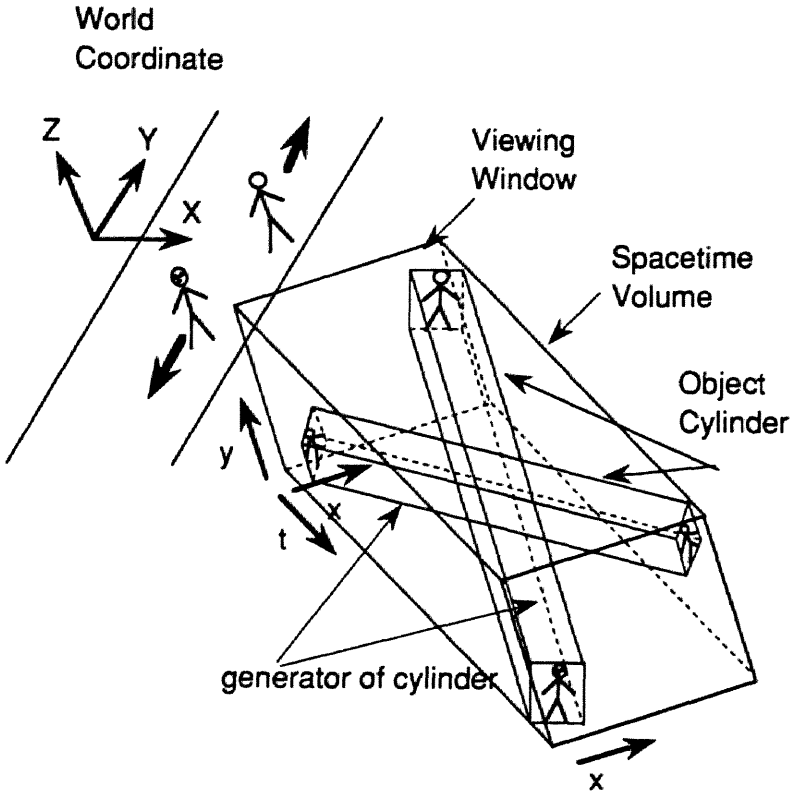


図 5.2: 動画像の時空間表現

域を形成する．この3次元立体を適当な平面でスライスした断面には様々な性質がある．ここでは，時空間立体を2枚の互いに直交する（非エピポーラ面）平面でスライスした断面から，物体フローを直接推定する方法を示す．これを**直交断面法 (Ortho-sectioning Method)** とよぶ．典型的な2断面の組合せは，1面が通常の画像（x-y 平面）で，もう1面が x-t 時空間画像の場合である．

5.3.1 動物体の非エピポーラ面画像

まず，断面が時間軸と動物体のフローを含む（フローと平行な）場合には，モーションステレオ画像解析等で定義されるエピポーラ面画像 (Epipolar Plane Image(EPI)) の類推で，**動物体のエピポーラ面画像** (EPI of Moving Object, EPImo) と呼ぶものが得られる．EPImo は時空間画像中の動物体領域を円筒の母線に沿ってスライスしているので，物体のフローに関する情報を，直線群で表された特徴点の流れ画像として保持している．したがって，その直線の傾きから各点の速度（速度の絶対値）を知ることができる[Bolles *et al.*, 1987][山本, 1981]．しかしながら，与えられた問題は，動物体のフローの方向と大きさの両方が未知の場合であり，したがってフローと平行な断面を決めることができない．また，フローの方向が物体ごとに異なれば，物体の数だけ断面を解析しなければならなくなり，非現実的である．さらに EPImo は常に同じ点を表示しているので，原形状に関する情報を与えない．

一方，この断面が物体のフローと角度づけされているときはその断面は動物体領域を斜めに横切ることになり，金太郎アメを斜めに切ったような像が現れる．動物体が剛体ならば，平行投影および平行移動の条件のもとでは，原画像における物体像を斜交軸変換した像となる．非剛体の場合でも，

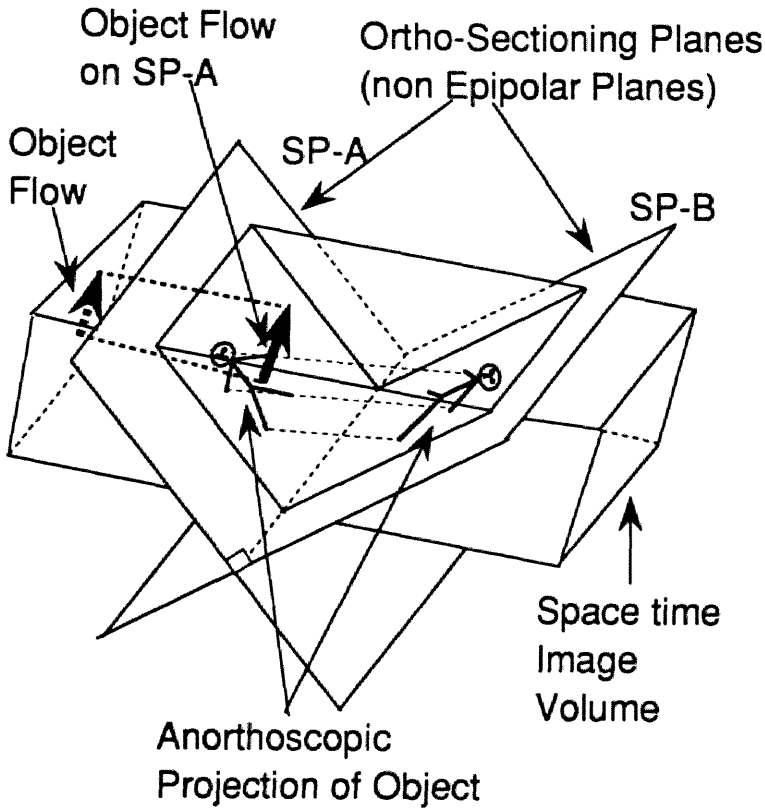


図 5.3: 直交 2 断面と物体の斜交投影像

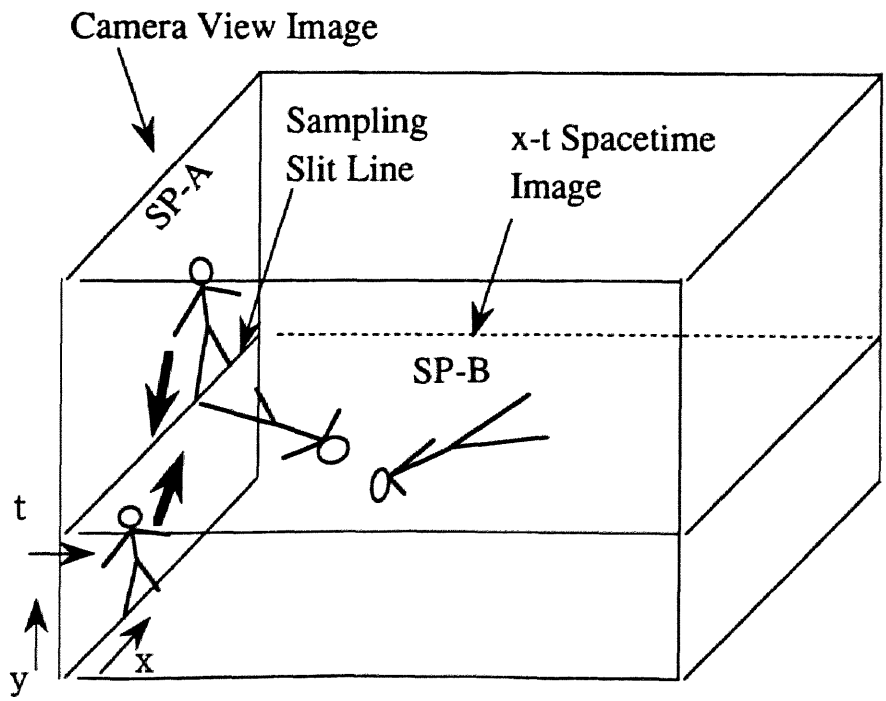


図 5.4: 典型的な直交 2 断面の配置

歩行者のような例では，かなり原物体の形状を反映した像がえられる（これを便宜的に**斜交投影像 (Anorthoscpic Projection)** とよぶ[3]）．この様な断面画像を本文では**動物体の非エピポーラ面画像**（以下，単に非エピポーラ面画像 (non-EPI)）と呼ぶ．non-EPI が時間軸の関数になっていれば，斜交投影像は速度と形状の両方の情報を備えている．そこで，斜交投影像を使って，動物体のフローを推定することができる．

5.3.2 直交 2 断面法によるフロー推定

いま，1 つの斜交投影像だけでは，速度と形状についてあいまいさが残るので，2 つの斜交投影像を使ってフローを推定する．そこで，2 枚の互いに直交する非エピポーラ面から得られる斜交投影像を用いる．これを直交 2 断面法と呼ぶ．図 5.3は，2 断面の一般的な配置の例である．このとき，物体フローは，まず 1 つの断面（例えば SP-A）上のフローとして求め，次にそれを x-y 画像平面に投影して求められる．しかしながら，現実問題としては，このような一般的な配置の 2 断面を用いることはあまり意味がない．このような断面画像をつくるには，2 本のサンプリングスリットを時間と共に移動することが必要となり，計算機構が複雑になってしまう．そこで，2 断面の特別な配置として，1 枚を x-y 画像平面，もう 1 枚を x-y 画像平面に垂直で t 軸を含む面とする．さらに一般性を失う事なく，光軸周りにカメラを回転することによって，サンプリングスリットを x 軸に平行にとって，x-t 時空間画像[間瀬, 1990]とすることもできる．こうして決められた 2 断面の配置は図 5.4のようになる．

図 5.5は，動物体の向きと，フローの方向および，サンプリングスリットの位置関係による x-t 時空間画像上の物体の斜交投影像の現れ方の例である[間瀬, 1990]．図 5.5 (b) は，フローとサンプリングスリットが平行の場合で，x-t 時空間画像は EPI になっている．そのほかの例では，すべて原形状を反映していることがわかる．同図 (c)(d) の配置では，シリンダの上面が左右どちらにあるかを調べないと速度の方向は決定できない[Zheng and Tsuji, 1990]が，(e)-(g) のような一般的な配置では，斜交投影像の傾きを調べるだけで決定できる．

ここでは，直交する 2 断面として，先に述べた画像平面と x-t 時空間画像を用いたときに物体のフローを推定する式を示す．すなわちフローの推定には，動物体の画像平面上の物体像と x-t 時空間画像上の斜交投影像の幾何情報を用いる．図 5.6は動物体がフロー $\mathbf{V}$ で平行移動しているときの 2 断面上の像とフローの関係を示している．ここで，物体は主軸をもっていて，その軸はサンプリングスリットに対して θ の角度があり，軸の長さ（物体の高さ）が $H = |\mathbf{H}|$ であると仮定する．さらに物体は時刻 t_0 から時刻 t_1 の間にスリットラインを通過するとする．すなわち通過時間を $\tau = t_1 - t_0$ とする．物体の移動によって，斜交投影像の主軸は幅 h，高さ τ の大きさを持つ．ここで，h は物体がサンプリングスリットを通過し始めたところの交差 x 座標 x_0 と通過し終った時点の交差 x 座標 x_1 で，$h = x_1 - x_0$ とする．このとき物体のフローは次式で定義される．

$$\mathbf{V} \equiv (V_x, V_y) = \frac{1}{\tau}(h - H\cos\theta, -H\sin\theta) \tag{5.1}$$

[3] Anorthoscopic Perception[Rock, 1981]の類推によるが，これに対する適当な日本語が見あたらないので斜交投影像と呼ぶことにする．

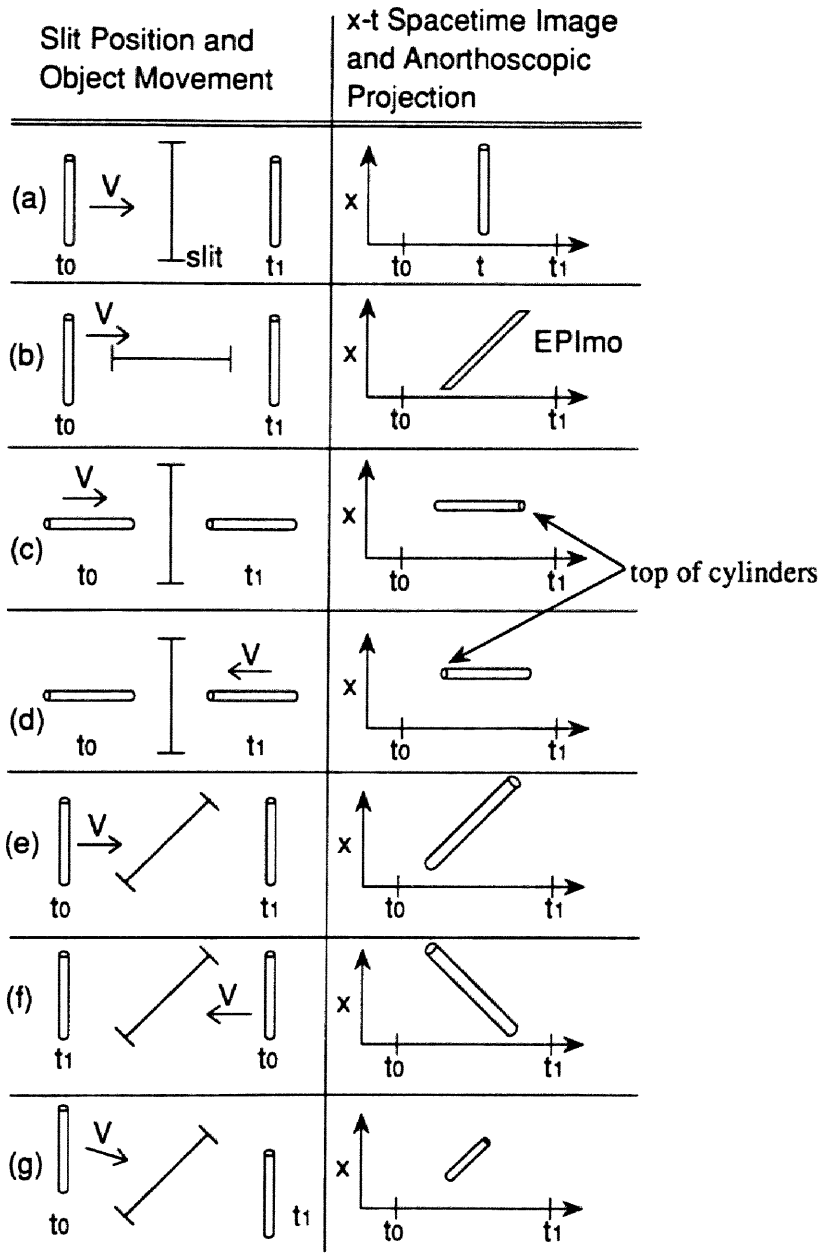


図 5.5: スリットの配置と斜交投影像の関係

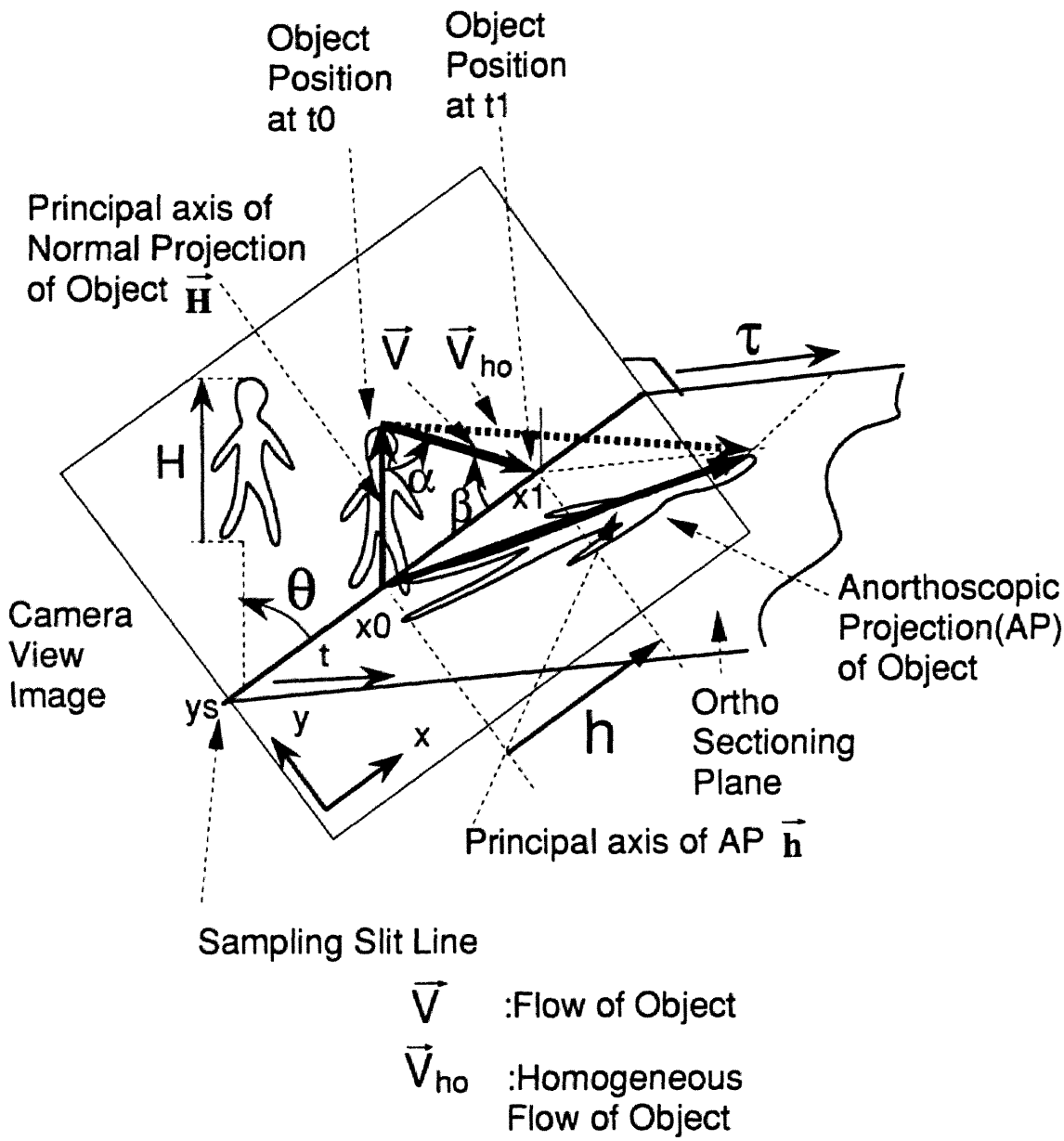


図 5.6: 直交断面法による物体フロー

あるいは，(x, y, t) の同次座標系で，

$$\mathbf{V_{ho}} \equiv (u, v, w) \equiv \mathbf{h} - \mathbf{H} = (h - H\cos\theta, -H\sin\theta, \tau) \tag{5.2}$$

となる．ここで，$\mathbf{H} \equiv (H\cos\theta, H\sin\theta, 0)$ および $\mathbf{h} \equiv (h, 0, \tau)$ は，それぞれ，物体の物体像の主軸と斜交投影像の主軸である．である．フロー (V_x, V_y) は同次座標形 (u, v, w) を用いると，$V_x = u/w, V_y = v/w$ と表される．

すなわち，物体のフローは物体像の主軸 $\mathbf{H}$ から斜交投影像の主軸 $\mathbf{h}$ への変位ベクトルに相当する．実際，物体が時空間中を t_0 から t_1 の間に移動したときに，$\mathbf{V_{ho}}$ は物体頂点の軌跡に相当することがわかる．したがってフローの推定精度はこれらの主軸の一致度に依存する．また，斜交投影像の主軸の大きさには次のような関係があり上記のフロー計算式の制約条件となる．

$$h = |\mathbf{H}|\frac{\sin\alpha}{\sin\beta} \tag{5.3}$$

$$\tau = \frac{|\mathbf{H}|\sin\theta}{|\mathbf{V}|\sin\beta} \tag{5.4}$$

ここで $\alpha(-\pi < \alpha < \pi)$ はフローベクトル $\mathbf{V}$ と，物体の主軸 $\mathbf{H}$ が画像平面上でなす角，$\beta(0 \leq \beta < \pi)$ はフローベクトルとサンプリングスリットがなす角で，$\alpha + \beta + \theta = \pi(0 < \alpha + \theta < \pi$ のとき) または 2π(その他) である．これらの角度とフロー計算の関係を調べると，まず $\sin\beta = 0$ のときは，h と τ は無限大となり，断面が EPImo になっていることに相当する．このとき h/τ，すなわち斜交投影像の傾きがフローの x 成分を与える．もし，$\sin\theta = 0$ ならば τ はゼロになり，斜交投影像は時間に関する情報をもたないのでフローを計算することができない．

すなわち，フローを計算するにはサンプリングスリットを物体の主軸と（平行でも直角でもない）適当な角度におくことが必要である．動物体の姿勢がまったくランダムに現れる場合でも，この条件により計算不能になる特殊な配置になることはまれである．複数のサンプリングスリットを使ってそのような状況に対処することも可能であろう．

また，歩行者を斜め上から撮影した場合などは主軸の向きはほぼ一定であるというヒューリスティックを使うことができる．さらに歩行者の実際の速度は，通路に対するカメラの設置条件から，フローをつかって計算する．

5.4 物体フローの推定実験

上記の方法に基づいて，動物体のフローを推定する 2 つの実験を行った．まず，グラフィックスで矩形の動物体を描画してフロー推定のシミュレーションを行った．次に，実際の画像を対象にフローの推定を試みた．

それぞれの実験で，画像平面上の物体像は，物体通過完了時 (時刻 t_1) の画像中のものを使用した．主軸の抽出は物体領域を抽出した後，領域中の最上点と最下点を求め，これらを主軸の端点とした．すなわち，x-y 画像平面では y_{max} および y_{min} となる 2 点，x-t 時空間画像の斜交投影像に対しては t_{max} および t_{min} となる 2 点を用いた．実験的に，変形する物体では，2 次モーメントによる主軸の計算に対して，本手法のほうが主軸の一致度は高いようである．

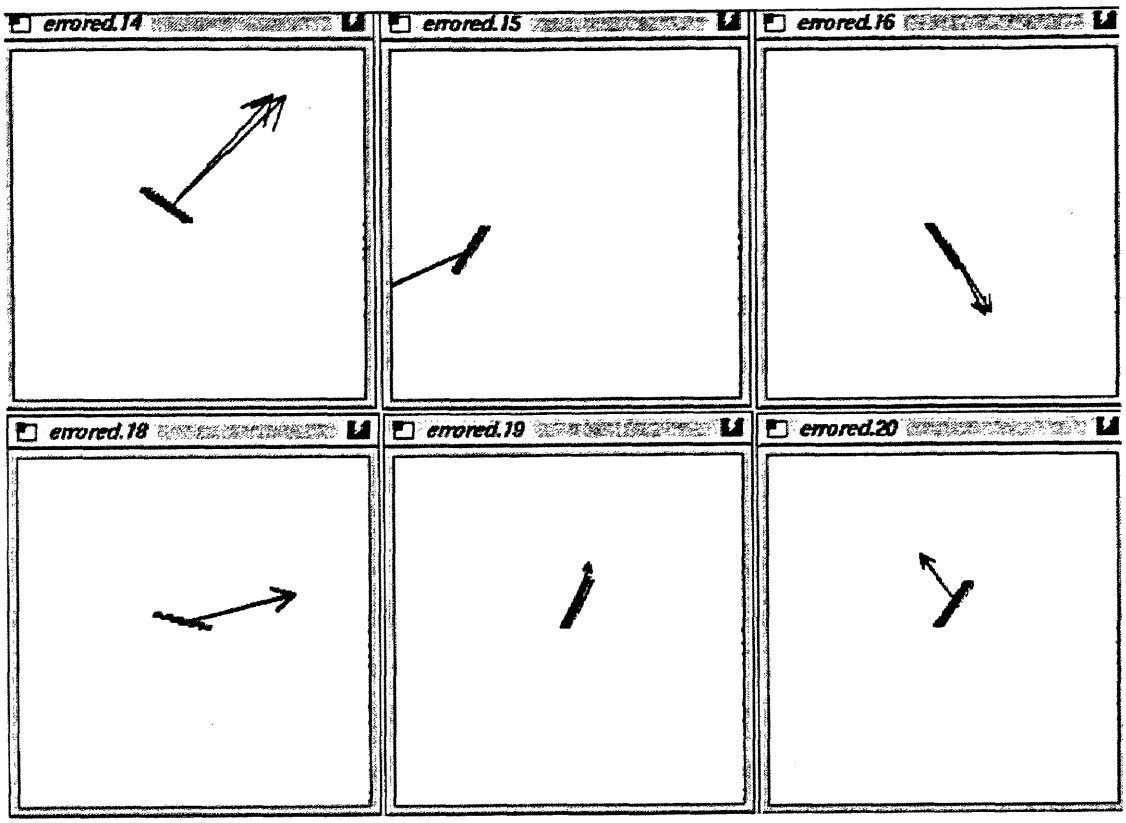


図 5.7: 物体フロー推定シミュレーション結果の例

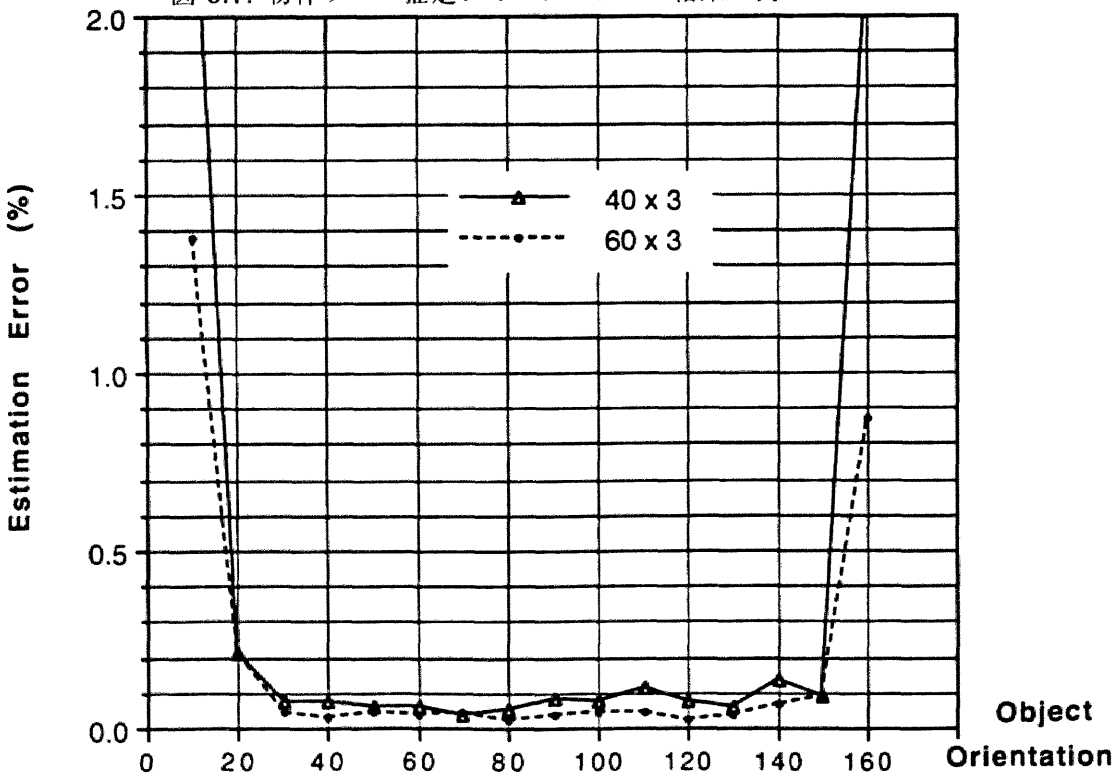


図 5.8: シミュレーションにおける推定誤差

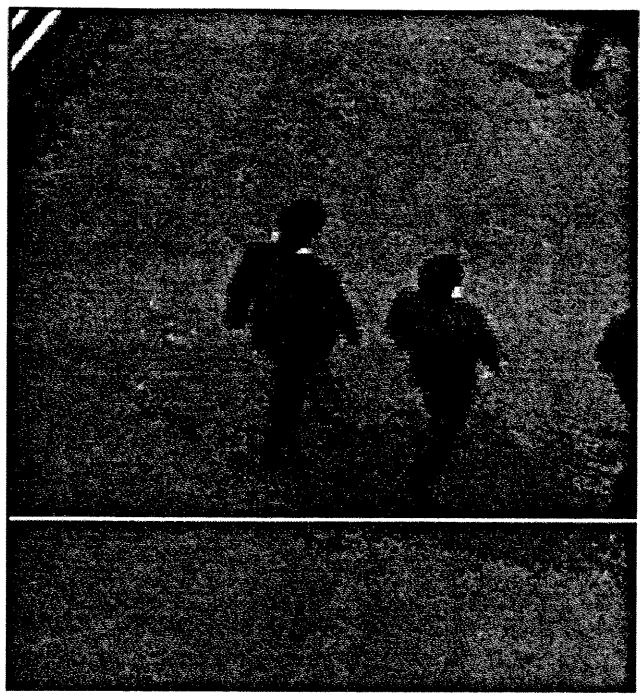


図 5.9: フロー推定実験にもちいた歩行者画像の一部

5.4.1 シミュレーション実験

256 × 256 の画像中にいろいろな大きさの矩形図形を発生させ，重心が画面の中央を通過するようにしながら，図形の傾きと速度をランダムに設定して，実際のフローと推定フローの比較を行った．図5.7はフローの推定結果の例である．実際のフローと推定結果を重畳して表示してある．実際に近いフローが推定できている．図中右上は物体主軸とフローの角度が平行に近かったので，誤差が大きくなっている．

また，図 5.8には，2 種類の矩形図形の大きさについて，物体の主軸とスリットの角度とフローの推定精度との関係をグラフで示した．誤差の計算式は次のようである．

$$e = \frac{|\mathbf{V}_{estimated} - \mathbf{V}_{true}|^2}{|\mathbf{V}_{true}|^2} \times 100.0(\%) \tag{5.5}$$

図中の平均誤差は 500 回の図形発生における推定の平均値（10 度区間について 20-40 サンプル）である．これから主軸の角度が 20 ～ 160 度の範囲では誤差が 0.5% 以下（平均）で推定できることが判る．また，主軸が水平に近くなると主軸の検出精度が速度推定に大きく影響を及ぼしていることが確かめられる．なお，速度に対して，物体がおおきくなれば主軸検出精度が相対的に高くなるので，図形の大きさ 40×3 に対して 60×3 の方が誤差が小さくなっている．

5.4.2 歩行者画像による実験

高所に設置したカメラで歩行者を斜め上方から撮影した画像を用いて実験を行った．焦点距離の長いレンズを使って，平行投影に近い条件で撮影した．こうすることによって，主軸は身長の方向になり，

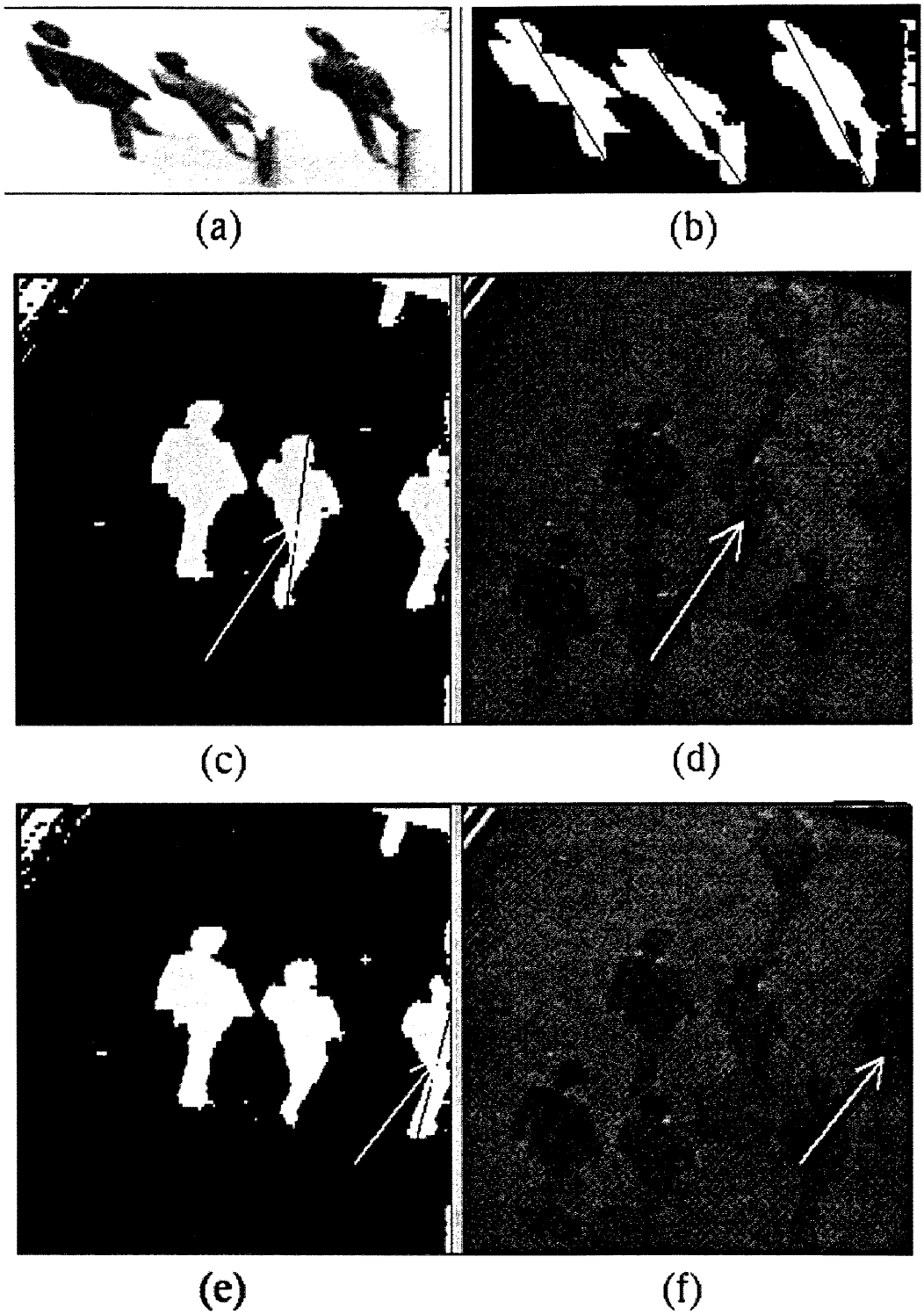


図 5.10: 歩行者画像のフロー推定結果

(a) x-t 時空間画像 (斜交投影画像), (b) 抽出された物体の斜交投影像と主軸,
(c) 抽出された物体像と主軸 (例 1), (d) 物体フローの推定結果 (例 1),
(e) 抽出された物体像と主軸 (例 2), (f) 物体フローの推定結果 (例 2)

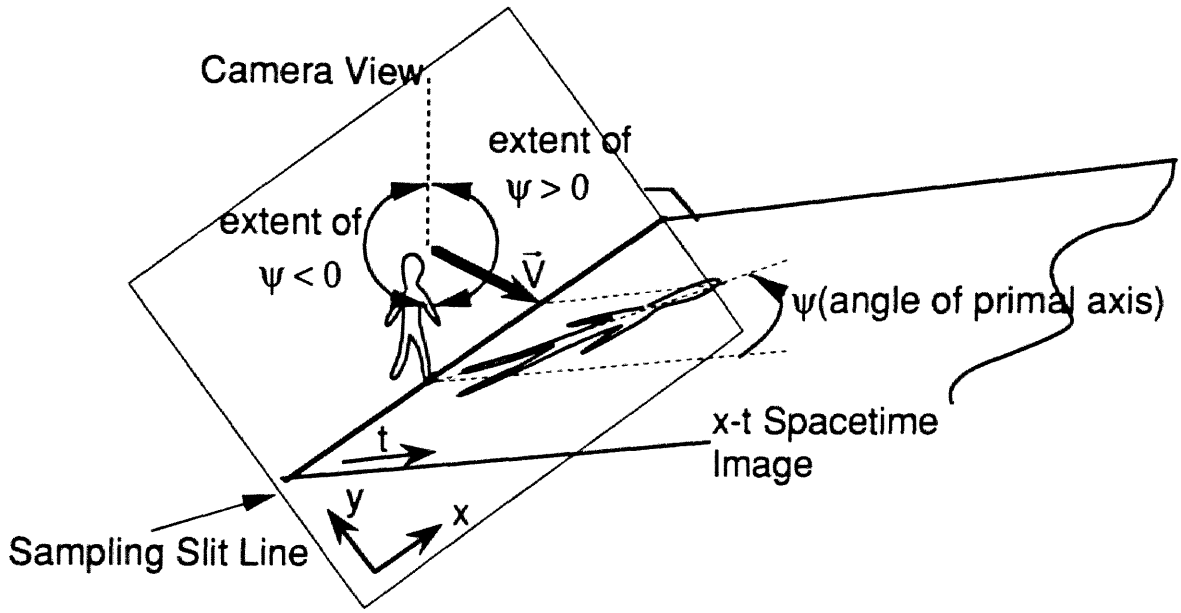


図 5.11: 斜交投影像の傾きと方向の正負

(5.3,5.4) 式の条件を満たすことになる．物体像の抽出には背景色が一定であるとの仮定により色空間での分離を行い，形状の平滑化を行った．図 5.9は，実験で用いた動画像の 1 コマで，横線がサンプリングスリットである．図 5.10(a)(b) はその x-t 時空間画像と，物体の斜交変換像の抽出結果である．抽出した主軸を直線で示してある．図 5.10(c) は画像面での物体抽出と主軸抽出結果である．図 5.10(d)(e) は 2 つの物体についてそれぞれ，1 秒 (30 フレーム) だけ離れたフレームを合成した画像に，推定したフローベクトルを重畳した．これから速度がほぼ正しく推定できていることがわかる．なお，これから実際の歩行速度を計算するには，歩道の平面に対して，このフローベクトルを射影することで求まる．

5.5 1 断面法による歩行者計数

前節で検討した直交 2 断面法を利用すると歩行者などの動物体の計数システムに応用できる．2 断面法では物体のフローを推定できたが，そのうち，x-t 時空間画像の 1 断面画像だけを使っても，移動方向のおおよその判定ができることを示す．したがって，歩行者がどちらに向かって歩いているかという方向別の計数が可能となる．また 2 断面法では物体がスリットを通過するたびに，画像平面上の物体像を調べる必要があるが，1 断面だけを使うことで，実時間処理が容易となる．この節では，移動方向の判定のアルゴリズムを示し，さらに歩行者計数システムのインプリメンテーションについて述べる．次節で，実験結果を報告する．

5.5.1 斜交投影像からのフローの方向判別

図 5.5(e),(f) で示したように，斜交投影像の傾き方から，物体の移動方向の正負を簡単に判別できる．投影像の傾きと移動方向の関係を図 5.11に示す．画像平面への物体投影像の主軸に対する物体フローの角度を α，斜交投影像の主軸の角度を ψ とすると，図からわかるように，両者には次の関係がある．

$$sign(\sin\alpha) = sign(\psi) \tag{5.6}$$

これは，式 (5.3) を変形した次式からも確認することができる．

$$\sin\alpha = \frac{h}{|\mathbf{H}|}\sin\beta, \tag{5.7}$$

ただし，$h(= x_1 - x_0)$ は $\psi > 0$ のとき正，$\psi < 0$ のとき負であり，$\sin\beta \geq 0$ である．

この判定においては，物体のどちら（上か下か）が先にスリットラインを通過したかにかかわらず方向を判定することができる．すなわち，物体の形状を認識する必要がなく，単に傾きを計算すればよいことになる．したがって，例えば，斜交投影画像の物体領域の 2 次モーメントを計算して，その主軸の時間軸に対する角度で傾きを決定できる．

これは，フローの方向の正負を決める手法であるが，フローのとりうる方向を物体像の主軸によって 2 つの領域に分けている．いいかえると，スリットラインに対してフローの方向の正負を決めているのではない．従って，任意のスリットラインを横切る物体の数をその方向によって完全に区別することは不可能である．しかしながら，歩道などの移動の方向に制約がついている場所では，カメラの位置に注意して，フローの正負判定領域が移動方向の制約条件の領域をオーバーラップするように設定することが可能である．歩行の方向が 360 度任意ではなく，適当に制限されているときには，十分判定することができる．

5.5.2 歩行者計数システム

上記で説明した $x - t$ 時空間画像から人物の領域を抽出し，進行方向を判定しながら計数するシステムについて述べる．$x - t$ 時空間画像を用いるもっとも大きな利点は，動画像のフレーム間で人物同士の同定・追跡をする必要がないことである．すなわち，1 つの動物体は $x - t$ 時空間画像上では，もとの形状をある程度反映した 1 つの領域になる．すなわち，原理的には $x - t$ 時空間画像中の領域数が，その時点までのスリット位置（走査線）を通過した物体数（＝人数）となる．また，処理対象の次元を 1 つ減らせるので，実時間処理も可能となる．

システム化するには，なお以下の問題を解決しなければならない．

1. 背景からの切り出し
2. 人物相互の分離
3. 領域数の数え上げ

背景からの切り出しには，背景画像を記憶しておいてそれとの差分で動領域を抽出した (2 章を参照)．ただし，背景は照明条件によって常に変化しているので，人が通過していない時の背景を使って，記

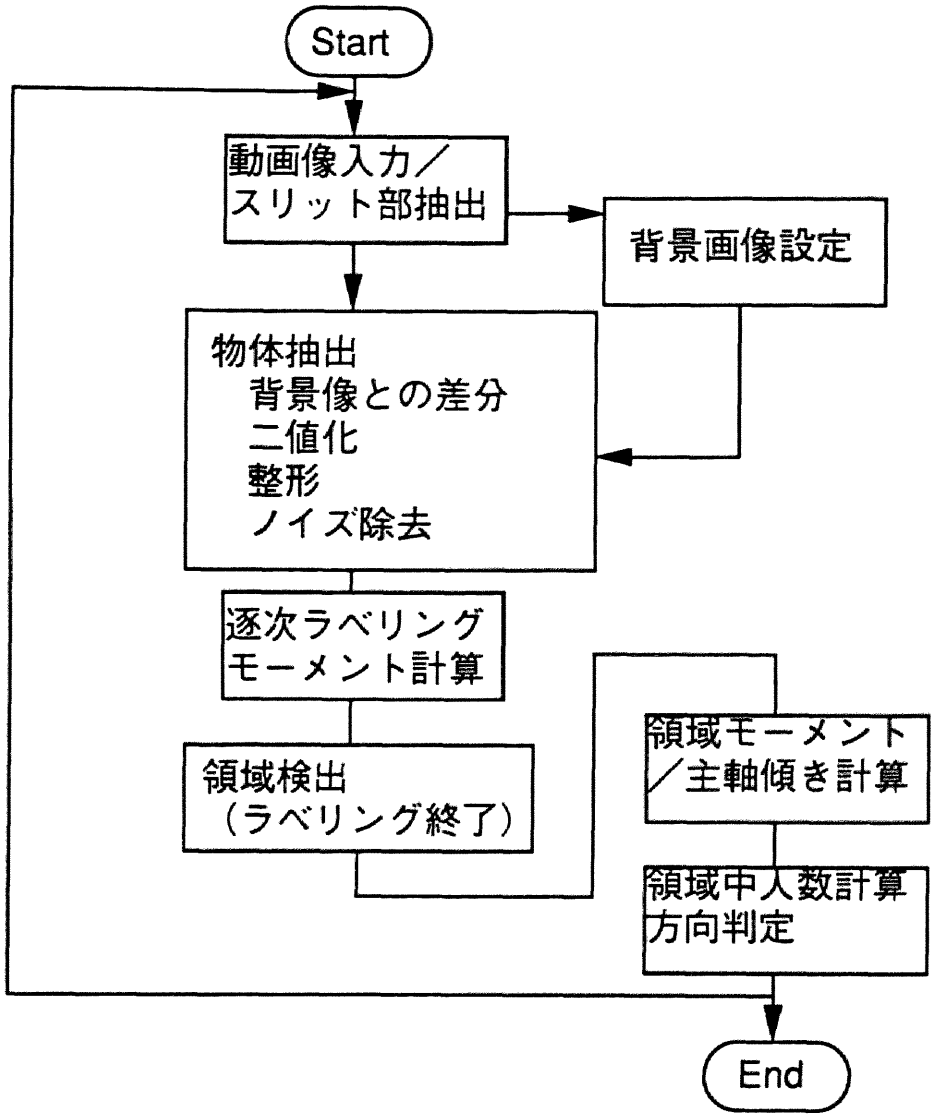


図 5.12: 歩行者計数システム基本処理ブロック図

憶した背景画像を更新するようにした．人物相互の分離の問題は，カメラが真上にあって，鉛直方向に撮影していれば，かなり簡略化される．一方，1断面法の性質から，カメラを歩行者に対して少し角度づけして撮影すると，歩行者の像の主軸とフローの方向に角度がついて，歩行の方向を判別することができる．主軸は真上からとった場合には肩の幅の方向にあるし，少し横からとれば身長の方向に現われる．ただし，同時に，並んで進む歩行者は領域が重なってしまうので分離の問題を解決する必要が生ずる．本文では，一般的なカメラ位置を想定し，速度も検出するようなシステムについて述べる．重なった領域については分離を行わず，その領域の総面積から，何人が重なっているかを推定する手法をとった．[4]

図 5.12は本システムの基本的な処理ブロック図である．すなわち，処理手順は

[4]速度がかわると，当然面積も変わるから，これはいつもうまく行くとは限らない．また，歩行者の密度が大きいと，方向の異なる歩行者が重なってしまうこともあるので，分離の問題は計数の精度をあげる上で重要な要素である

1. 入力動画像から，ある走査線（スリット）における $x-t$ 時空間画像を作成
2. 自動登録した背景画像との差分をつかって，人物領域を抽出
3. 局所マスクをつかって領域を整形，孤立点を除去
4. 領域のラベリング
5. 各領域の 0,1,2 次モーメントの計算
6. 領域の面積と傾きを求め，面積から各領域の人数を計算し，領域の傾きから方向（ここでは正負だけ）を判定
7. 通路をいきかう人の計数

となる．

実際には $x-t$ 時空間画像は時間軸方向に無限の長さがあるため，各スリット像の入力毎に上記の処理を繰り返して，逐次，各時刻での計数を報告するシステムが必要である．以下，逐次処理に適したラベリング手法と，逐次的に各モーメントを計算する方法，さらに速度の検出法と面積からの人数の推定法について述べる．

5.5.3 領域数の逐次数え上げを目的としたラベリング法

通常の画像のラベリング法は 1 枚の静止画像を対象として，画像を何回か走査したのち，各領域にラベルを振り，全領域数を報告するものである．この処理においては，ラベルのテーブルを用意して，走査中に接続が判明した領域同士は，それらのラベルをつなぐポインタをはる処理が行われ，走査終了後その連結関係に基づいてラベルの付け替えを行う[鳥脇, 1988][後藤ほか, 1989]．ところが，さきに述べたように $x-t$ 時空間画像は無限大の大きさであり，そのためのラベルテーブルも無限大の大きさが必要になる．

幸運なことに，現在必要としている特徴量は，領域の 0, 1, 2 次の各モーメントと領域数だけであり，これらは逐次的に計算できる．そこで，時刻 t で領域が途切れたときに，その領域のために必要だったラベルを解放して再利用するタイプのラベリング手法を開発した．

$x-t$ 時空間画像の連結成分の性質

横軸 (x) 方向の画像の幅を N，縦軸方向に時刻 t をとる．このとき画像中の連結成分には次の性質がある[5]．ただし，ここで連結成分のラベルは各時刻で判明している連結関係でラベルの振り直しがおこなわれ，1 つの連結成分は 1 つのラベルで表現されているとする．

[性質 1]（最大連結成分数）

各横軸方向の 1 画素高さのスライスにおいて，存在可能な連結成分の数（あるいは黒ランの数）はたかだか $N/2$ である．したがって，任意の時刻 T において，最低必要なラベルの数は最大 $N/2$ である．（証明略）

[5] この性質は一般の 2 次元画像なら常に成立する．

[性質 2] (因果関係)

時刻 $t < T$ に存在する連結成分が，時刻 $t = T$ の少なくとも 1 つの成分と連結していなければ，$t > T$ の成分と連結することはない．(証明略)

[補題] (連結成分の終結)

時刻 $t < T$ に存在する連結成分が，時刻 $t = T$ のどの成分とも連結しなかったら，その連結成分はその時点で終結する．このときその連結成分のラベルは解放できる．(時刻 $T-1$ で使われたラベル L が，$t = T$ で使われなかったら，$t > T$ において，使われることはない．) (証明略)

これらの性質を使って，有限長のラベルテーブルを使った，ラベリングアルゴリズムを考えることができる．

テンポラル－ラベリングアルゴリズム

ここで提案するのは 4 連結のラベリング法である．アルゴリズムのゴールは 2 値画像 (値 1 が対象領域) を走査して，その各走査において，連結している領域を見つけ出し，連結領域のラベルのトリーを構成するとともに，連結性が途切れたときにそれを検知することである．

[初期化]

まず，1 行分のラベル画像バッファ $L(x)$ を用意する．ラベリングはこのバッファ上で行われる．また，ラベルのトリーを作るためのテーブルを用意する．ラベルテーブルのエントリには，親ノードのラベル (l→parent)，走査中連結成分表示フラグ (l→active)，統計量テーブル (l→statistics) を用意する．ラベル画像バッファも NIL に初期化しておく．l→active は各ラインの走査に先だって初期化 (NIL) する．

[走査]

以下，$x-t$ 画像において，画素 $f(x,t)$ を走査するときに次の処理を行う．

```
when f(x,t) = 1
 if( L(x) = A and L(x-1) = B )
   then   C := union(A,B),  L(x) := C
 if( ( L(x) = A and  L(x-1) = NIL ) or
     ( L(x) = NIL and  L(x-1) = A ) )
   then   C := union(A,1),  L(x) := C
 if( L(x) = NIL and L(x-1) = NIL)
   then   C := create_label(), L(x) := C
when f(x,t) = 0
         L(x) := NIL
```

ここで，union(A,B) はノード A,B の各親ノードを調べて，小さい集合の親ノードを新たな親ノード C として，さらに各ノードの統計量をマージして返す．このとき，走査画素のデータを統計量に追加し，ノード C の l→active をセットする．

create_label() は新しい親ノードを作成し，走査画素のデータを統計量にセットする．

1ラインの走査終了時には，ラベルテーブルをサーチして，親ノードで，かつ，その走査で未利用のラベル (l→active = NIL) を捜し，そのようなラベルを走査終了の連結成分として報告する．また，そのラベルのすべての子ラベルを解放する．

5.5.4 モーメント計算

領域のモーメントは形状を表す統計的な特徴量として使われるが，ここでは2次までのモーメントを計算する．領域 S の 0,1,2 次の各モーメントは，

$$n = M_0 = \sum_{(x,t)\in S} 1 \tag{5.8}$$

$$\mathbf{M_1} = (\bar{x},\bar{t})^T = \frac{1}{n}\sum_{(x,t)\in S}(x,t)^T \tag{5.9}$$

$$\begin{aligned}\mathbf{M_2} &= \begin{pmatrix} M_{xx} & M_{xt} \\ M_{xt} & M_{tt} \end{pmatrix} = \frac{1}{n}\sum_{(x,t)\in S}\begin{pmatrix} (x-\bar{x})^2 & (x-\bar{x})(t-\bar{t}) \\ (x-\bar{x})(t-\bar{t}) & (t-\bar{t})^2 \end{pmatrix} \\ &= \frac{1}{n}\mathbf{V} - \mathbf{M_1}\mathbf{M_1}^T \end{aligned} \tag{5.10}$$

で，計算される．ここで，

$$\mathbf{V} = \sum_{(x,t)\in S}\begin{pmatrix} x^2 & xt \\ xt & t^2 \end{pmatrix} \tag{5.11}$$

である．

そこで，ラベルの統計量統合の際に，$M_0, \mathbf{M_1}, \mathbf{M_2}$ の計算の基礎となるデータ x, t, x^2, xt, t^2 の総和をとっていけばよい．2つの領域 (A,B) の統合の場合にもそれぞれの領域のデータの和を計算する．領域の走査が終った段階で，上式によってモーメントが計算される．

5.5.5 速度の推定方法

2次モーメントから，領域の主軸を計算して領域の傾きをもとめる．主軸 $x = t\tan\psi$ の傾き ψ は，

$$\tan^2\psi + \frac{M_{xx} - M_{tt}}{M_{xt}}\tan\psi - 1 = 0 \tag{5.12}$$

の解として求まる[南, 中村, 1989]．このうち，慣性モーメントを最小にする傾き（固有値の大きいほうの固有ベクトル）を物体の主軸の傾きとする．

5.5.6 人数の計数法

各領域の0次モーメントから領域中の人数を計算する．人数の計算式は，

$$number = \begin{cases} int(M_0/SS) & \text{if } M_0 > T_s \\ 1 & \text{if } T_n < M_0 \le T_s \\ 0 & \text{otherwize} \end{cases} \tag{5.13}$$

表 5.1: 歩行者計数結果 (10 分間における計数結果, in/out)

時刻 (分)	1	2	3	4	5	6	7	8	9	10	計 (6,7 分を除く計)
計測結果	9/11	13/9	9/3	7/0	8/7	(9/21)	(19/6)	13/18	6/6	4/4	97/85(79/58)
実数	9/11	13/9	6/4	6/1	11/10	(18/18)	(30/2)	17/14	5/9	5/3	120/81(78/61)

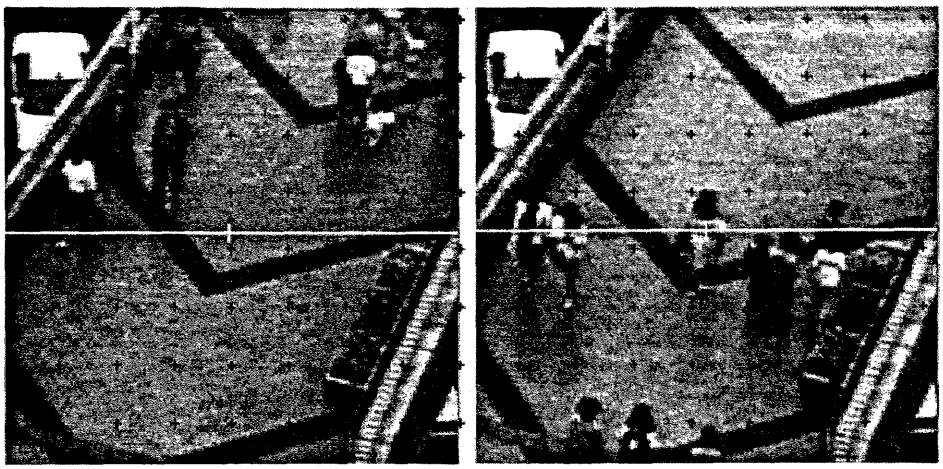


図 5.13: 計数実験対象画像の例

のようになり，面積 M_0 があるしきい値 T_s より大きいときは，標準の面積 (SS) で割って人数を計算し，T_s 以下で T_n （雑音領域のしきい値）より大きいときは 1 人，そのほかはノイズとして 0 とする．ここで関数 $int()$ は整数化を行なう．

5.6 実験システムの構成と実験結果

上記システムをフレームグラバー付きのグラフィックスワークステーション上に作成し，ビデオデッキの出力を処理対象動画像として，実時間処理するシステムを構築した．

計数対象は斜め上方から通路を撮影して，えられた画像である (図 5.13(a),(b)). 原動画像サイズ 256 × 256 画素に対して幅 256 画素のスリットを設定し，$x-t$ 時空間画像を構成するようにして，処理を行った．

この画像を実時間で走査して，歩行者を計数した．図 5.14は領域抽出，領域計数を行った結果である．処理はディジタイズから領域抽出，ディスプレイへの表示を含めて平均 38msec/frame で行なわれ，実時間処理が可能であることがわかった．ただし，領域のラベリングはプログラミング上，大きめのバッファを設け，背景の更新が行われるときに，リセットするようにしたため，連続して人が通過すると，ラベルの使用量が増え，ラベルバッファのチェックに時間がかかり，すこし遅くなる．

表 5.1は方向別の判定と計数方法の評価のために，10 分間の計測結果を実測値と比較したものである．背景からの人物の抽出は，服装が背景に近いと非常に困難である．そこで，この実験では，領域の抽出に失敗したものは実際の通行量には含んでいない．この抽出の精度を上げる工夫は，環境の変化に

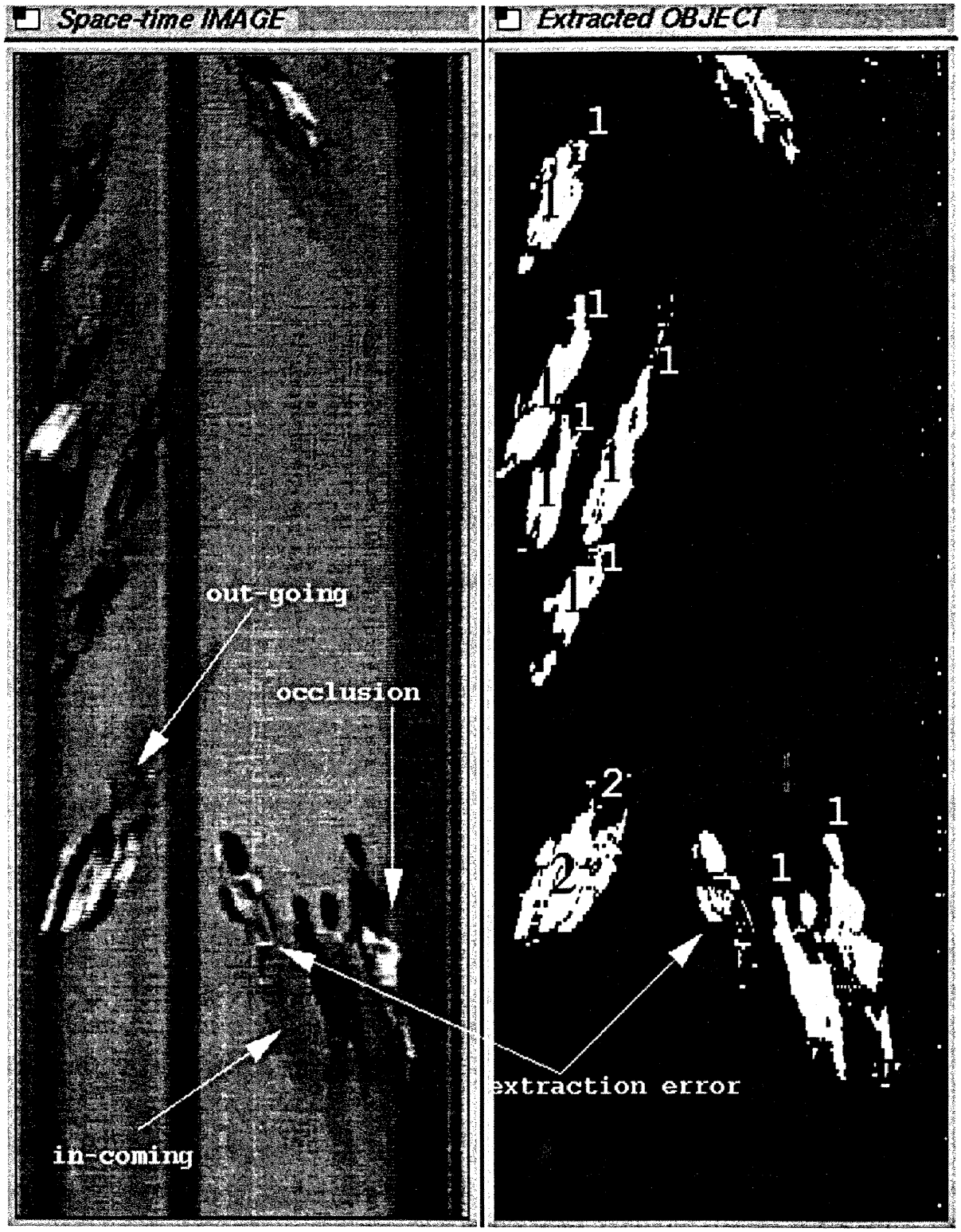


図 5.14: 領域抽出・計数結果

対する追従と併せて解決する問題である[前田ほか, 1991]. 全体としては, 入／出が実際 120/81 人に対して, 計数結果 97/85 人とあまりよくない. 特に 6 ～ 7 分のところで, 精度が悪くなっている. そこで, その部分を省くと, 実際 78/61 人に対して, 計数結果 79/58 人と 90% 以上の精度で計数できている. この原因は次のように考えられ, 今後, この計数方法の改良するべき点を示唆している. まず, 事実としてこの時間帯は通行量が多かった. そのとき, 次の 3 つの問題がある.

1. 方向判定のアルゴリズムは個々の物体が分離できているときに成立する. したがって, 通行量が多く密度が上がるとオクルージョンがおきやすくなり, 時空間画像上で斜交投影像が重なると方向判定に失敗する.
2. 方向判定のアルゴリズムはスリットに対して直角な方向で区別しない. したがって, 歩行者が通路を斜めに通ると方向判別が失敗する.
3. 密度があがると, ラベリングの為に処理速度が遅くなり, 単位面積をつかった計数に影響をおよぼし, 少な目に計数する.

上記の 6 ～ 7 分の区間は以上のことが起きていると考えられる. これらは, 時空間画像上の形状を処理して頭部の抽出や領域のセグメンテーションにより回避する必要がある. これらは今後の課題である.

5.7 むすび

動画像の時空間画像立体表現した中の動物体領域の性質から, 時空間画像の直交 2 断面を使った物体フローの推定法を提案した. 断面が動物体の非エピポーラ面となっているときの断面上の投影像の幾何情報に基づくフローの計算式を示した. 本文では, 2 断面として, 画像平面と x-t 平面を用いて説明した. 一般的にはエピポーラ面とならない任意の直交 2 断面で同様な計算が可能であるが, プログラミングなどの点で, 上記の組合せが最適で, 十分である. さらに, シミュレーション画像と実際の動画像を使って, フローの推定実験をおこない, 本推定法の動作を確認した. 実験によれば, 物体とスリットの位置関係によって, 投影像の主軸の抽出が不安定になりフロー計算の誤りとなる場合がまれに確認された. これを排除するには物体の位置や向きに関するヒューリスティックを積極的に活用するか, 複数のスリットを角度を変えて設定し, 複数の断面の組合せで, 計算精度をあげることが考えられる.

なおここで推定しようとするフローは, マクロな量である. すなわち, サンプリングスリットを物体が通過し始めてから終るまでの間の, スリット位置における物体全体の平均的な見かけの速度である. したがって, 物体の大きさや速度にも依存するが, 通常はフロー推定に数十フレームを必要とする. いいかえると, 瞬間の速度推定には向いていない. 一方, 歩行者のように形状の変化する対象でも追跡をする必要[浅田ほか, 1979]なく処理できるという利点がある.

このような物体の速度推定には相関マッチングによる追跡に基づく方法があり, 単純な比較は困難である. あえて計算量を比較すると, $O(nS^2\delta\tau)$ かかるのに対し, 本手法では $O(nS + nS_x\tau)$ ですむ（ S は物体面積, S_x は物体の x 方向長, n は物体数）ことになる.

本手法を有効に利用するには, フローを求めたい個々の動物体が分離されて抽出できる必要がある. したがって, 背景から動物体を正確に抽出するための手法の検討が重要である.

さらに，$x-t$時空間画像中の動物体の物体像の性質から動物体の移動方向の正負を判別する方法を示し，これを利用して歩行者を計数するシステムを構成した．システムは，まず，物体像領域の傾きの正負だけを計算して，速度の正負の方向だけをもとめて，通路を出入りする人を区別して計数した．実時間処理可能なシステム構築のために，テンポラル－ラベリングアルゴリズムをインプリメントした．

カメラ位置を斜めにすることによって，速度が求まる一方で，歩行者の領域が1部重なってしまう．ここで示したシステムでは領域面積から，歩行者数を推定したが，速度が変わると，この面積も変化することから，将来は個々の歩行者の速度を使って面積を補正する必要がある．また計数精度向上のためには，領域形状を使った歩行者の分離・計数も考えられる．

第 6 章

結論

本論文は，動画像処理による人物動作の認識理解に基づくマンマシンインタフェースの研究について論じた．

人間にやさしく快適なインタフェースは，非接触で情報を入手できる視覚機能を持ち，人間が日常的な動作の中で発するメッセージを取り込むことが必要である．本研究は，身体の各部位ごとに発せられるメッセージを抽出・認識するためにヘッドリーダ，フェイスリーダ，リップリーダ，ピープルリーダの各サブシステムを構築するための検討を行った．

第 1 章では，視覚をもちいたインタフェースのシステムとしてヒューマンイメージリーダの概念を提案した．また，インタフェースをユーザである人間とコンピュータとのあいだのメッセージの伝達機関ととらえ，人間がおくるメッセージを従来のグラフィック入力デバイスとの対比で整理し，人間の画像情報からのメッセージをいかに抽出するかの問題を提起した．さらに従来の関連研究を概観するとともに，本論文の位置づけを明確にし，研究の基本方針及び本論文の構成を述べた．

第 2 章では，ヘッドリーダを構築した．頭部の動作からおおまかな視方向を抽出して，ロケータとしてのメッセージを抽出するとともに，「はい・いいえ」などのセレクタメッセージの抽出を可能にした．そのために，まず頭部の 3 次元動作を，目鼻などの造作の抽出を必要としないで抽出する方法を導出した．遠方からの顔の向きの判別に我々が用いていると思われる，髪と顔の見え方のモデルを使った手法を導出した．さらにその効果をシミューレーションと実際の画像で実験し確認した．このモデルは大まかな動作の抽出には適しているが，詳細な動きを推定することは得意ではない．したがって，今後はこのモデルを使って大まかな動きを抽出したあと，顔の部品のモデルを使って詳細な動きを再構成する coarse-to-fine の手法を開発する必要がある．一方，この手法はワークステーション程度の処理能力で実時間処理が可能であり，このようなメッセージ伝達機能をもったインタフェースの実験システムを構築することが可能となった．実際には知的符号化方式のプロトタイプシステム，頭部の動きを使ったメニューの選択，「はい・いいえ」のメッセージによる電子秘書システムとの対話環境などを試作し，視覚によるインタフェースの実体験を可能にした．その結果，実際のインタフェースとしてシステムに組み込んで使うには，まだ，速度の問題やほかのメディアとの統合などの解決をはかる必要があることがわかったが，実現の可能性はあると考えられる．

第 3 章では，フェイスリーダを構築した．顔が伝えるメッセージのなかで，個人情報についで重要

な，表情による感情などの感性やイメージに関わるメッセージの抽出を可能にする手法を導出した．顔の表情をとらえるには，画像処理による困難な問題の1つである目鼻口などの造作の詳細な抽出を必要としない，オプティカルフローによる表皮の動きの観測を利用した．これにより，造作抽出を行なわずに微妙な変化を含む動きを観測でき，さらに顔の筋肉モデルを使って表情筋の動きを推定することが可能となった．表情が伝える原メッセージは，(1) 筋肉の動き，(2) コード化された体系による表情の記述，(3) 表情変化から推定できる基本感情，などがあり，3.3節，3.4節，3.5節でそれぞれのメッセージの抽出方法を導出した．コード化された体系にはFACSを適用した．3.6節では感情識別の実験をおこない，4種類の基本感情 (幸福，怒り，驚き，嫌悪) の識別実験を行い，約85% の識別結果を得られることを確かめた．オプティカルフローによる表情メッセージの抽出は動きを直接的に計測できるという利点があるが，頭部全体のゆれに影響をうけるという問題がある．すなわち，表情の動き推定は非剛体の動き抽出の問題の典型的な例であり，今後は頭部全体のグローバルな動きと表情のローカルな動きを同時に抽出する問題を解決していく必要がある．また，1章のグラフィック入力デバイスとの対応づけにおいて，本論文で初めて導入したイメージャデバイスの概念は実存するデバイスがないため，そのメッセージをインタフェースがどの様な場面でどう使うかを検討することも今後の課題である．その際，表情を読みとるフェイスリーダはイメージャデバイスの典型的具体例として用いることができよう．

第4章は，リップリーダを構築した．発話動作が顔の表情の1部であることに注目して，3章で導出した手法を口唇まわりに適用して，発話時のことばのメッセージを抽出するサブシステムを作成した．表情と同じくオプティカルフローを原データとして，発話時の口唇まわりの動きをとらえることに成功した．特に，オプティカルフローデータを主成分分析して，発話時の主となる動き成分を抽出した結果，主観的にも重要な動きと考えられる，顎の上下による口の開閉と口の伸張の2つの動きを得ることができた．これらの動きを特徴ベクトルとした英数字認識システムを構築し，実験で効果を確認した．システムの特徴として，オプティカルフローによる動きデータを特徴ベクトルとしたことによって，連続発声時でも口唇の停止や反対方向への動きを，速度ゼロの点として抽出することが可能になり，連続発声の単語認識システムとなっている．実験で作成した0から9までの認識システムで実験を行なったところ約75% の認識率を得た．これは，通常の音声認識システムに比べれば，低い認識率ということもできるが，人間でも文脈無しでは読唇は非常に困難であることからも，十分高い認識率であるということができる．また，実際にはリップリーダ単独で使うよりも，音声認識システムと組み合わせて，精度の高い認識を行なううえで利用することが現実的であろう．

第5章は，ピープルリーダの1つであるp- カウンタを構築した．人がある場所を何人通過したかというメッセージは，ビルや街を管理するシステムにとっては重要である．部屋の中に何人がいるかという情報は1人1人とのやりとりでも，事前知識として使える場面がある．まず，p- カウンタを構築する準備として，動画像からある計数ラインを通過する物体の速度を計測する手法を導出した．ここでは，時空間立体表現した画像を2つの直交する平面でスライスした断面上の物体像の幾何学的特徴を使う直交断面法でフローの推定を行なった．そしてシミュレーションと実画像による実験で効果を確認した．さらに，スライス平面を1枚とする1断面法により，歩行者計数で必要な方向の識別が容易にでき

ることを示し，実験システムを構築した．ワークステーションで実時間で動作する実験システムで歩行者計数の実験を行い，調整により計数誤差 10% 以内という精度を得た．現在の実験システムでは物体像の抽出に ad hoc な画像処理を使っているため，照明条件の変化に追随することは困難であり，これを改良することが今後の課題といえよう．また，直交断面法は時空間画像の新しいとらえ方として，フローの解析手法に新しい道を開く可能性がある．

快適なインタフェースを構築するにはあらゆる部分の動きの解釈を総合する必要がある．その意味でこの研究は緒についたばかりといえる．しかしながら，以上述べたように，本論文の研究成果は動画像処理を用いた視覚機能を有するインタフェースの構築のための，技術レベルの向上，今後の研究テーマの発掘に寄与するところが大きい．この研究成果をもとに，将来快適でやさしいインタフェースを持ったコンピュータや家電製品が登場して，生活を豊かにしてくれることを確信する．

謝辞

本研究をまとめるにあたり，名古屋大学工学部鳥脇純一郎教授，杉江 昇教授ならびに横井茂樹助教授には丁寧な指導と励ましを賜りました．私の在学中の御教授も併せて深く感謝するとともに厚くお礼申し上げます．

本研究は，NTT ヒューマンインタフェース研究所 (旧複合通信研究所) と米国マサチューセッツ工科大学 (MIT) メディア研究所において，多数の方のご指導，ご協力により行なわれました．特に本研究の開始から今日にいたるまで親しくご指導頂いた，ヒューマンインタフェース研究所マルチメディア処理研究部 末永康仁博士には心から感謝いたします．また，MIT メディア研究所滞在中に同所 Alex Pentland 準教授による指導と encouragements により，本研究のうちのリップリーダの仕事が行なわれ，本研究を大きく推進させることができたことに心よりの感謝の意を表します．また，ヒューマンインタフェース研究所元視覚情報研究部においてご指導ご鞭撻頂くとともに，MIT への研修の機会を作って頂いた，NTT インテリジェントテクノロジー社 小森和昭氏 (元視覚情報研究部長) に感謝致します．また，遅々として進まない本研究の進展を暖かく見守り，ご指導ご鞭撻いただいたヒューマンインタフェース研究所 研究所長 釜江尚彦博士，画像メディア研究部長 安田 浩博士，音声処理研究部長 小林幸雄博士 (元視覚情報研究部長)，マルチメディア処理研究部長 遠藤隆也氏，酒井高志氏，石井健一郎博士，および新規事業開発部 玉邑嘉章氏，MIT メディア研究所 Ted Adelson 準教授に深く感謝致します．

名古屋大学福村晃夫名誉教授（現中京大学教授）の御助言により大学院に進み，研究者への道が開けました．パターン認識の基礎理論等，在学中の数々の御教授と併せて心から御礼申し上げます．

本研究の遂行にあたって，マルチメディア処理研究部 渡部保日児氏，秋本高明氏（現総合企画本部），佐藤 敦氏，福本雅朗氏，企業通信本部開発部 新田正樹氏，電気通信協会 安達亜紀子嬢ならびに吉田桂子嬢には数々の協力を頂きましたことを感謝いたします．

また日頃から御討論頂くマルチメディア処理研究部 赤松 茂氏，塩 昭夫氏，上田修功氏（現コミュニケーション科学研究所），志沢雅彦氏（現ＡＴＲ），前田栄作氏，知能ロボット研究部 尺長 健博士，を始めとする NTT ヒューマンインタフェース研究所の皆様にお礼申し上げます．マルチメディア処理研究部において数々の計算機環境を整備されて，実験を可能にしていただいた森本正志氏，高橋裕子氏ほかの方々に感謝致します．電気通信協会 橋本伸江夫人には Macintosh で何枚も印象的な図面を作成して頂き，大変助かりました．

両親と，妻 正美と 2 人の娘の愛情と協力に心から感謝します．

参考文献

[Aggarwal and Nandhakumar, 1988] J. K. Aggarwal and N Nandhakumar. On the computation of motion from sequences of images - a review. *Proc. IEEE*, Vol. 76, No. 8, pp. 917–935, 1988.

[Akita, 1984] Kiochiro Akita. Image sequence analysis of real world human motion. *Pattern Recognition*, Vol. 17, No. 1, pp. 73–83, 1984.

[Birdwhistell, 1970] Ray L. Birdwhistell. *Kinesics and Context: Essays on Body Motion Communication.* Univ. of Pennsylvania Press, Philadelphia, 1970.

[Bolles *et al.*, 1987] R.C. Bolles, H. Baker, and D.H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *IJCV*, Vol. 1, No. 1, pp. 7–55, June 1987.

[Bolt, 1980] Richard A. Bolt. Put-that-there: Voice and gesture at the graphics interface. *ACM Computer Graphics*, Vol. 14, No. 3, pp. 262–270, Apr. 1980.

[Bolt, 1987] Richard A. Bolt. The integrated multi-modal interface. 信学論, Vol. J70-D, No. 11, pp. 2017–2025, Nov. 1987.

[Bruce, 1988] Vicki Bruce. *Recognising Faces.* Lawrence Erlbaum Assoc., London, 1988.

[Darwin, 1872] Charles Darwin. *The Expression of the Emotions in Man and Animals.* John Murray, London, 1872. Reprinted Chicago: Univ. of Chicago Press, 1965.

[Duda and Hart, 1973] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, 1973.

[Ekman and Friesen, 1975] Paul Ekman and Wallace F. Friesen. *Unmasking the Face.* Consulting Psychologist Press, Inc., CA, 1975.

[Ekman and Friesen, 1978] P. Ekman and W. V. Friesen. *The Facial Action Coding System.* Consulting Psychologists Press, Inc., San Francisco, CA, 1978.

[Ekman and Friesen, 1990] Paul Ekman and Wallace F. Friesen. in personal communication with P. Ekman and W. F. Friesen, Feb. 1990.

[Ekman, 1973] Paul Ekman, editor. *Darwin and Facial Expression.* Academic Press, inc., New York, NY, 1973.

[Ekman, 1985] Paul Ekman. *Telling Lies.* Berkley Book, 1985.

[Finn and Montgomery, 1988] E. K. Finn and A. A Montgomery. Automatic optically-based recognition of speech. *Pattern Recognition Letters*, Vol. 8, No. 3, pp. 159–164, 1988.

[Foley and Van Dam, 1982] J. D. Foley and A. Van Dam. *Fundamentals of Interactive Computer Graphics.* Addison-Wesley, Reading, Mass., 1982.

[Fujimura, 1961] Osamu Fujimura. Bilabial stop and nasal consonants: a motion picture study and its acoustical implications. *J. of Speech and Hearing Research*, Vol. 4, No. 3, pp. 233–247, 1961.

[Fukuda and Hiki, 1982] Yumiko Fukuda and Shizuo Hiki. Characteristics of the mouth shape in the production of japanese – stroboscopic observation. *J. of Acoust. Soc. Jpn.*, Vol. (E)3, No. 2, pp. 75–91, 1982.

[Goldstein *et al.*, 1972] A. J. Goldstein, L. D. Harmon, and A. B. Lesk. Man-machine interaction in human-face identification. *Bell Sys. Tech. Journal*, Vol. 51, No. 2, pp. 399–427, Feb. 1972.

[Heeger, 1988] David J. Heeger. Optical FLow Using Spatiotemporal Filters. *International Journal on Computer Vision*, Vol. 1, pp. 279–302, 1988.

[Horn and Schunck, 1981] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, Vol. 17, pp. 185–203, 1981.

[Izard, 1979] C. E. Izard. *The maximally discrimitive facial movement coding system(Max).* Instructional Resouros Center, Univ. of Delaware, Neward, DE, 1979.

[Kass *et al.*, 1987] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *Proc. ICCV-87*, pages 259–268, June 1987.

[Kaya and Kobayashi, 1971] Y. Kaya and K. Kobayashi. A basic study on human face recognition. In *Int. Conf. Frontiers of Pattern Recognition*, pages 265–289, HI, 1971.

[Kruger, 1983] Myron W. Kruger. *Artificial Reality.* Addison-Wesley, 1983.

[Mase and Pentland, 1989] Kenji Mase and Alex P. Pentland. Automatic lipreading by computer. 信学会, 画像理解の高度化と高速化シンポジウム予稿, pages 65–70, Apr. 1989.

[Mase *et al.*, 1987] Kenji Mase, Yasuhito Suenaga, and Takaaki Akimoto. Head reader: A head motion understanding system for better man-machine interaction. *IEEE proc. SMC*, pages 970–974, Nov. 1987.

[Mase *et al.*, 1990] Kenji Mase, Yasuhiko Watanabe, and Yasuhito Suenaga. A real time head motion detection system. *proc. SPIE 1260*, pages 262–269, Feb. 1990.

[Mase, 1990] Kenji Mase. An application of optical flow –extraction of facial expression–. *IAPR Workshop on Machine Vision and Applications*, pages 195–198, Dec. 1990.

[Mase, 1991] Kenji Mase. Recognition of facial expression from optical flow. *Trans. IEICE (E-74)*, 10, pages 3474–3483, 1991.

[Massaro, 1987] D. W. Massaro. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Lawrence Erlbaum Assoc. Publ., New Jersey, USA, 1987.

[McGurk and J.McDonald, 1976] H. McGurk and J.McDonald. Hearing lips and seeing voices. *Nature*, Vol. 264, pp. 746–748, 1976.

[Minsky, 1985] Marvin Minsky. *The Society of Mind.* Simon and Schuster, New York, 1985.

[Norman, 1988] Donald A. Norman. *The Psychology of Everyday Things.* Basic Books Inc, New York, 1988. (野島久雄訳「誰のためのデザイン？」, 新曜社).

[Otsu, 1979] N. Otsu. A threshold selection method from gray-level histograms. *IEEE trans. SMC*, Vol. SMC-9, pp. 62–66, 1979.

[Parke, 1975] Frederic I. Parke. A model for human faces that allows speech synchronized animation. *Computer & Graphics*, Vol. 1, pp. 3–4, 1975.

[Perkell, 1986] J. S Perkell. Coarticulation strategies : preliminary implications of a detailed analysis of lower lip protrusion movements. *Speech Communications*, Vol. 5, pp. 47–68, 1986.

[Petajan and Bodoff, 1988] B. Petajan, E. D. Bischoff and D. Bodoff. An improved automatic lipreading system to enhance speech recognition. In *ACM SIGCHI'88*, pages 19–25, 1988.

[Petajan, 1984] E. D. Petajan. *Automatic Lipreading to Enhance Speech Recognition.* PhD thesis, U.of Illinois, 1984.

[Petajan, 1985] E. D. Petajan. Automatic lipreading to enhance speech recognition. In *Proc. CVPR'85*, pages 40–47, 1985.

[Platt and Badler, 1981] S.M. Platt and N.I. Badler. Animating facial expressions. *Computer Graphics*, Vol. 15, No. 3, pp. 245–252, Aug. 1981.

[Rock, 1981] I. Rock. Anorthoscopic perception. *Scientific American*, pages 103–111, Mar. 1981.

[Tamura and Kawasaki, 1986] Shinichi Tamura and Shingo Kawasaki. Recognition system for sign language motion image. *IPS of Japan, tech. report*, CV44-1, Sep. 1986.

[Terzopoulos and Waters, 1990] Demetri Terzopoulos and Keith Waters. Analysis of facial images using physical and anatomical models. *proc. ICCV'90*, pages 727–732, Dec. 1990.

[Turk, 1991] Matthew Turk. *Interactive-Time Vision: Face Recognition as a Visual Behavior*. PhD thesis, MIT, Cambridge, MA, 1991.

[Waite and Welsh, 1990] Jon. B. Waite and William J. Welsh. Head boundary location using snakes. *British Telecom Technol. J.*, Vol. 8, No. 3, pp. 127–136, Jul 1990.

[Watanabe and Suenaga, 1989] Yasuhiko Watanabe and Yasuhito Suenaga. Drawing human hair using wisp model. *proc. CGI'89*, pages 691–700, 1989.

[Waters, 1987] K. Waters. A muscle model for animating three-dimensional facial expression. *Computer Graphics*, Vol. 21, No. 4, pp. 17–24, 1987.

[Yuille *et al.*, 1989] Alan L. Yuille, David S. Cohen, and Peter W. Hallinan. Feature extraction from faces using deformable templates. *IEEE proc. CVPR'89*, pages 104–109, Jun. 1989.

[Zheng and Tsuji, 1990] Jiang Yu Zheng and Saburo Tsuji. From Anorthoscopic Perception to Dynamic Vision. *1990 IEEE Int. Conf. on Robotics and Automation*, pages 1154–1160, 1990.

[相澤ほか, 1989] 相澤清晴, 原島博, 斉藤隆弘. 構造モデルを用いた画像の分析合成符号化方式. 信学論, Vol. J72-B-I, No. 3, pp. 200–207, Mar. 1989.

[青野, 石川, 1991] 青野俊宏, 石川正俊. 確率過程を用いたセンサフュージョン－多系列隠れマルコフモデルを用いた視聴覚融合－. 第2回自律分散システムシンポジウム, pp. 115–118, Jan. 1991.

[青山, 河越, 1988] 青山宏, 河越正弘. 顔画像計測に基づく視線感知法. 計測自動制御学会 5回パターン計測部会研究会, 1988.

[秋本ほか, 1990] 秋本高明, リチャードワレス, 末永康仁. 正面・側面像を用いた頭部形状の自動推定. *90* 春季信学全大, 1990.

[浅田ほか, 1979] 浅田稔, 谷内田正彦, 辻三郎. 運動物体の検出と追跡. 信学論 *(D)*, Vol. J63-D, No. 6, pp. 395–402, 1979.

[阿部ほか, 1990] 阿部享, 阿曽弘具, 木村正行. 3次元表面形状による人間の顔の自動識別 — B-スプライン曲 面の制御点を利用して —. 信学論, Vol. J73-D-II, No. 9, pp. 1477–1484, Sep. 1990.

[石井, 岩田, 1984] 石井威望, 岩田洋夫. 濃淡画像による顔の表情の自動認識. *29* 回情処全大, Vol. 4M-12, 1984.

[上田ほか, 1991] 上田修功, 間瀬健二, 末永康仁. エネルギー最小化に基づく輪郭追跡. 情処研資, CV73-5, Jul. 1991.

[大西ほか, 1990] 大西剛, 竹村治雄, 伴野明, 岸野文郎. 手振りを用いたユーザインタフェースに関する一検討. 信学技報, Vol. HC90-23, 1990.

[大村ほか, 1989] 大村和典, 伴野明, 小林幸雄. 単眼視画像による顔の向き検出法の指示入力への応用. 信学論, Vol. J72-D-II, No. 9, pp. 1441–1447, Sep. 1989.

[小野, 黒川, 1985] 小野眞, 黒川隆夫. 構動素による身振りの記述法とこれを用いた動画上の身体 動作の解析. 情処研資, CV35-2, 1985.

[金子ほか, 1988] 金子正秀, 羽鳥好律, 小池淳. 形状変化の検出と3次元形状モデルに基づく顔動画像の符号化. 信学論, Vol. J71-B, No. 12, pp. 1554–1563, Dec. 1988.

[技研,] 技研トレーディング. Passer dual counter. 製品.

[栗田ほか, 1988] 栗田知好, 本多清志, 垣田有紀. 口唇画像情報を併用する音声の分析. 信学技報, SP88-94, pp. 41–48, 1988.

[黒川, 1988] 黒川隆夫. ヒューマン・インタフェースとしての動作言語. 計測と制御, Vol. 27, No. 1, pp. 49–55, Jan. 1988.

[黄, 高羽, 1983] 黄乗元, 高羽禎雄. ITV 画像による人の流れの実時間計測. 信学論, Vol. J66-D, No. 8, pp. 917–924, Aug. 1983.

[後藤ほか, 1989] 後藤敏行, 太田善之, 吉田真澄, 白井良明. 連結領域の高速ラベル付けアルゴリズム. 信学論, Vol. J72-D-II, No. 2, pp. 247–255, Feb. 1989.

[崔ほか, 1990] 崔昌石, 原島博, 武部幹. 顔の 3 次元モデルに基づく表情の記述と合成. 信学論, Vol. J73-A, No. 7, pp. 1270–1280, July 1990.

[崔ほか, 1991] 崔昌石, 原島博, 武部幹. 顔の 3 次元モデルを用いた顔面表情の分析. 信学論, Vol. J74-D-II, No. 6, pp. 766–777, June 1991.

[坂井ほか, 1973] 坂井利之, 長尾真, 金出武雄. 計算機による顔写真の解析. 信学論, Vol. 56-D, No. 4, pp. 226–233, 1973.

[佐々木ほか, 1991a] 佐々木努, 赤松茂, 深町映夫, 末永康仁. 正面顔画像の自動識別法の検討. 信学技報, IE91, Sep. 1991.

[佐々木ほか, 1991b] 佐々木努, 赤松茂, 末永康仁. 顔画像認識のための色情報を用いた顔の位置合わせ法. 信学技報, IE91–2, Apr. 1991.

[笹間, 1988] 笹間宏. 鉄道における画像処理の応用. 計測と制御, Vol. 27, No. 12, pp. 51–58, 1988.

[末永, 渡部, 1990] 末永康仁, 渡部保日児. A Synchronized Cylindrical Range and Color Data Scanner and its Application to 3D Face Data Acquisition. 情処研資, CV67–5, Jul. 1990.

[末永, 1991] 末永康仁. 知的インタフェースのための人物像の認識と合成. *91* 信学春季全大, GD–5, Mar. 1991.

[末永ほか, 1992] 末永康仁, 間瀬健二, 福本雅朗, 渡部保日児. Human reader: 人物像と音声による知的インタフェース. 信学論, Vol. J75-D-II, no. 2, pp.190–202, Feb. 1992.

[角, 太田, 1989] 角 保志, 太田友一. 並行トップダウン処理方式による顔画像の解析. 情処研資, CV63–1, 1989.

[積山, 東倉, 1989] 積山薫, 東倉洋一. 読唇情報が音声知覚に果たす役割. テレビ学技報, Vol. 13, No. 44, pp. 31–36, Sep. 1989.

[高橋, 岸野, 1990] 高橋友一, 岸野文郎. 手振り認識方法とその応用. 信学論, Vol. J73–D–II, No. 12, pp. 1985–1992, Dec. 1990.

[田村ほか, 1989] 田村進一, 梶見直樹, 岡崎耕三, 光本浩士, 河合秀夫, 副井裕. エネルギー関数とオプティカルフローを用いた口形輪郭の抽出・補完と追跡. 信学技報, PRU89-20, June 1989.

[寺田, 1991] 寺田康和. B–ISDN の展望. *NTT R&D*, Vol. 40, No. 1, pp. 1–8, Jan. 1991.

[伴野ほか, 1989] 伴野明, 石井淳一, 岸野文郎. 視線検出のための瞳孔抽出光学系の設計法. 信学技報, OS89–37, 1989.

[鳥脇, 1988] 鳥脇純一郎. 画像理解のためのディジタル画像処理 *(II)*. 昭晃堂, 1988.

[畠山, 1989] 畠山卓朗. 重度肢体障害者のための自立生活支援機器 — 電子機器を 中心に —. 計測と制御, Vol. 28, No. 7, pp. 621–624, Jul. 1989.

[服部, 1991] 服部桂. 人工現実感の世界. 日刊工業出版社, 1991.

[福本ほか, 1991] 福本雅朗, 間瀬健二, 末永康仁. 動作と音声を統合したマルチメディアインタフェース. *91* 秋季信学全大, Sep. 1991.

[藤田, 1976] 藤田恒太郎. 生体観察. 南山堂, 17th edition, 1976.

[本名 信行, 1981] 本名 信行ほか訳. ノンバーバル・コミュニケーション *(W. Raffler-Engel* 著 *)*. 大修館書店, Japan, 1981.

[前田ほか, 1991] 前田 英作ほか. 多次元輝度空間上での分布を利用した物体抽出法. 信学技報, PRU91–60, Sep. 1991.

[間瀬, ペントランド, 1990] 間瀬健二, アレックス・ペントランド. オプティカルフローを用いた読唇. 信学論, Vol. J73–D–II, No. 6, pp. 796–803, Jun. 1990.

[間瀬, 末永, 1985] 間瀬健二, 末永康仁. 顔画像の動き検出の一手法. 情処 *30* 全大, *7M–5*, Mar. 1985.

[間瀬, 1990] 間瀬健二. x-t 時空間画像を用いた歩行者の計数. 信学技報, IE90–43, Sep. 1990.

[間瀬ほか, 1988] 間瀬健二, 末永康仁, 秋本高明, 玉邑嘉章. 顔の動き検出・生成システム – パソコンによる顔方向の検 出 –. *63* 情処全大,*1V-7*, Mar. 1988.

[間瀬ほか, 1989] 間瀬健二, 渡部保日児, 末永康仁. リアルタイム頭部動作認識合成システム. *PCSJ'89*, pages 127–128, Oct. 1989.

[間瀬ほか, 1990] 間瀬健二, 渡部保日児, 末永康仁. 人物を対象とする表現技術の研究動向. 情処研資, CG46-3, Aug. 1990.

[間瀬ほか, 1991a] 間瀬健二, 前田英作, 末永康仁. 表情動画像からの感情の認識の 1 手法. 信学技報, PRU91-24, May 1991.

[間瀬ほか, 1991b] 間瀬健二, 渡部保日児, 末永康仁. ヘッドリーダ : 画像による頭部動作の実時間検出. 信学論, Vol. J74-D-II, No. 3, pp. 398–406, Mar. 1991.

[松岡ほか, 1986] 松岡清利, 古谷忠義, 黒須顕二. 画像処理による読唇の試み. 計測自動制御学会論文集, Vol. 22, No. 2, pp. 191–198, Feb. 1986.

[南, 中村, 1989] 南敏, 中村納. 画像工学. コロナ社, 1989.

[南, 1991] 南敏. 顔画像による個人識別の技術. システム / 制御 / 情報, Vol. 35, No. 7, pp. 415–422, July 1991.

[森 於菟, 1950] 森 於菟ほか. 解剖学. 金原出版, 1950.

[安田, 1988] 安田浩. テレビ電話はいま. スペクトラム, Vol. 1, No. 5, pp. 88–102, May 1988.

[山本, 1981] 山本正信. 画像化された運動軌跡による動画像処理. 情処論文誌, Vol. 22, No. 5, pp. 442–449, Sep. 1981.

研究発表一覧

《論文等》

[1] 間瀬健二, 鳥脇純一郎, 福村晃夫："拡張されたディジタルボロノイ線図とその画像処理への応用", 信学論, **J64–D**, 11, pp. 1029–1036(1981).

[2] 間瀬健二, 末永康仁, 玉邑嘉章："ルックアップテーブル動画用複合画像の合成法", 信学論, **J67–D**, 9, pp. 1060–1061(1985).

[3] 間瀬健二, 末永康仁, 玉邑嘉章："ルックアップテーブル動画の合成法", 信学論, **J68–D**, 4, pp. 749–756(1985).

[4] 間瀬健二, アレックス・ペントランド："オプティカルフローを用いた読唇", 信学論, **J73–D–II**, 6, pp. 796–803(1990).

[5] 間瀬健二, 渡部保日児, 末永康仁："ヘッドリーダ：画像による頭部動作の実時間検出", 信学論, **J74-D-II**, 3, pp. 398–406(1991).

[6] K. Mase: "Recognition of facial expression from optical flow", Trans. IEICE , E–74, 10, pp.3474–3483(1991).

《国際会議等》

[7] K. Mase, Y. Suenaga and T. Akimoto: "Head reader: A head motion understanding system for better man-machine interaction", IEEE proc. SMC, pp. 970–974(1987).

[8] K. Mase, Y. Watanabe and Y. Suenaga: "A real time head motion detection system", proc. SPIE 1260, pp. 262–269(1990).

[9] K. Mase and A. P. Pentland: "Lipreading: Automatic visual recognition of spoken words", OSA Image Understanding and Machine Vision, 1989 Technical Digest Series, **14**, pp. 124–127(1989).

[10] K. Mase: "An application of optical flow –extraction of facial expression–", IAPR Workshop on Machine Vision and Applications, pp. 195–198(1990).

《共著による論文・国際会議》

[11] 秋本高明, 間瀬健二："画素選択型光線追跡法", 信学論, **J–69D**, 12, pp. 1943–1952(1986).

[12] T. Akimoto, K. Mase, A. Hashimoto and Y. Suenaga: "Pixel selected ray tracing", EUROGRAPHICS'89, pp. 39–50(1989).

[13] 秋本高明, 間瀬健二, 末永康仁："改良画素選択型光線追跡法", 信学論, **J73-DII**, 1, pp. 62–71(1990).

[14] T. Akimoto, K. Mase, and Y. Suenaga: "Pixel-Selected Ray Tracing", IEEE CG& A, vol.11, no.4, pp.14–22 (1991).

[15] A. Hashimoto, K. Mase and Y. Suenaga: "Vista ray-tracing: High speed ray tracing using perspective projection image", proc. CGI'89, pp. 549–561(1989).

[16] M. Shizawa and K. Mase: "Simultaneous Multiple Optical Flow Estimation", ICPR'90, **1**, pp. 274–278(1990).

[17] M. Shizawa and K. Mase: "A unified computational theory for motion transparency and motion boundaries based on eigenenergy analysis", CVPR'91, pp. 289–295(1991).

[18] M. Shizawa and K. Mase: "Principle of superposition: a common computational framework for analysis of multiple motion", Proc. IEEE Workshop on Visual Motion(1991).

[19] M. Shizawa and K. Mase: "Principle of superposition: for multiple motion and motion transparency", IJCAI91 Workshop on Dynamic Motion Understanding(1991).

[20] 上田修功, 間瀬健二, 末永康仁: "弾性輪郭モデルとエネルギー最小化原理による輪郭追跡手法", 信学論, **J75-D-II**, 1, pp.111–120, (1992).

[21] N. Ueda and K. Mase: "Tracking moving contours using energy minimizing elastic contour models", proc. ECCV-92, Italy, (1992). (to appear)

[22] 末永康仁, 間瀬健二, 福本雅朗, 渡部保日児: "Human reader: 人物像と音声による知的インタフェース", 信学論, **J75-D-II**, 10, pp. 190–202, (1992-02).

《その他の発表 (研究会、全国大会、雑誌、講演) 》

[23] 間瀬健二, 鳥脇純一郎, 福村晃夫, 横井茂樹: "一般化距離変換の逐次型アルゴリズムとその性質", 信学技報, **PRL79–40**, (1979).

[24] 間瀬健二, 鳥脇純一郎, 福村晃夫, 横井茂樹: "一般化距離変換におけるパスの定義とその性質について", 昭55 信学全大 (1980).

[25] 間瀬健二, 鳥脇純一郎, 福村晃夫: "2 値画像の外部スケルトンとその応用", 信学技報, **PRL80–38**, (1980).

[26] 間瀬健二, 鳥脇純一郎, 福村晃夫: "外部スケルトンの抽出アルゴリズムとその応用", 第 22 回情処全大, pp. 615–616(1981).

[27] 間瀬健二: "ディジタル画像の距離変換に関する研究", Master's thesis, 名古屋大学、情報工学専攻 (1981).

[28] 間瀬健二, 名倉正計: "図形コマンドで記述されたカラー画像の実時間重畳表示法", 信学技報, **PRL81–102**, (1982).

[29] 間瀬健二, 玉邑嘉章, 池沢秀樹: "動画を対象とした図形表示装置の構成", 昭 58 信学全大, **1220**, (1983).

[30] 間瀬健二, 玉邑嘉章: "動画像データの階層化表現法に関する検討", 昭 58 信学情報システム部門全大, **136**, (1983).

[31] 間瀬健二, 玉邑嘉章: "LUT 書替による動画表示のためのデータ生成法", 信学技報, **IE83–57**, (1983).

[32] 間瀬健二: "カラー LUT 書替動画データの生成法", 第 14 回画像工学コンファレンス, **4–4**, (1984).

[33] 間瀬健二, 三ツ矢英司, 末永康仁: "LUT 動画複合画像生成法の濃淡画像への適用", 昭 59 信学通信部門全大, **825**, (1984).

[34] 間瀬健二, 河久保秀二, 玉邑嘉章, 石橋聡: "静止画像応答装置の情報作成編集制御法", 昭 59 信学全大, **1402**, (1984).

[35] 間瀬健二, 三ツ矢英司, 末永康仁: "ディジタルビデオディスクを有する画像処理実験システム とその応用", 第 15 回画像工学コンファレンス, **8–3**, (1984).

[36] 間瀬健二, 秋本高明, 末永康仁: "3 次元格子点上の直線発生法", 昭 60 信学全大, **1175**, (1985).

[37] 間瀬健二: "顔画像処理による動作識別法の検討", 昭 60 信学情報システム部門全大, **105**, (1985).

[38] 間瀬健二, 末永康仁: "顔画像の動き検出の一手法", 情処 30 全大, 7M–5(1985).

[39] 間瀬健二, 末永康仁: "逆透視変換を利用した自動景観モンタージュ", 昭 62 信学全大, **1625**, (1987).

[40] 間瀬健二: "顔画像処理による計算機との対話", 昭 62 信学情報システム部門全大, **100**, (1987).

[41] 間瀬健二, 末永康仁: "顔画像処理による頭の動作識別法", 情処研資, CV47–1(1987).

[42] 間瀬健二, 末永康仁, 秋本高明, 玉邑嘉章: "顔の動き検出・生成システム – パソコンによる顔方向の検 出 –", 63 情処全大,1V-7(1988).

[43] K. Mase and A. P. Pentland: "Automatic lipreading by computer", 信学会 – 画像理解の高度化と高速化シンポジウム予稿, pp. 65–70(1989).

[44] 間瀬健二, ペントランドアレックス：“オプティカルフローを用いた読唇”, TV 学技報, **VAI89-8**, pp. 7–12(1989).

[45] 間瀬健二, 渡部保日児, 末永康仁：“リアルタイム頭部動作認識合成システム”, PCSJ'89, pp. 127–128(1989).

[46] 間瀬健二, 赤松茂, 末永康仁：“人物像の認識と生成の研究”, 1990 情処全大, **1P-9**, (1990).

[47] 間瀬健二, 渡部保日児, 末永康仁：“重心による顔方向検出の精度測定”, 1990 春季信学全大, **D–541**, (1990).

[48] 間瀬健二：“オプティカルフローを用いた表情認識の検討”, 1990 春季信学全大, **D–544**, (1990).

[49] 間瀬健二：“オプティカルフロー抽出による表情筋の動作検出”, 信学技報, **PRU89–128**, (1990).

[50] 間瀬健二, 渡部保日児, 末永康仁：“人物を対象とする表現技術の研究動向”, 情処研資, **CG46-3**, (1990).

[51] 間瀬健二：“x-t 時空間画像を用いた歩行者の計数”, 信学技報, **IE90–43**, (1990).

[52] 間瀬健二, 末永康仁：“頭部の動作によるウィンドウ選択法”, 41 情処全大, **6G–6**, (1990).

[53] 間瀬健二：“逐次数え上げラベリング法とその歩行者計数への応用”, 1990 信学秋季全大, **D-300**, (1990).

[54] K. Mase and A. Pentland: "Lipreading by optical flow", Systems & Computers in Japan, 22, 6, pp.67–76(1991). (to appear, english version of mase-pentland90a).

[55] 間瀬健二：“ヒューマンインタフェースと画像処理”, 画像ラボ, 日本工業出版, pp. 34–37(1991).

[56] 間瀬健二：“時空間の直交 2 断面を使った物体フローの推定”, 42 情処全大, **4D-9**, (1991).

[57] 間瀬健二：“非エピポーラ面画像による物体速度の推定”, 情処研資, **CV71-2**, (1991).

[58] 間瀬健二, 前田英作, 末永康仁：“表情動画像からの感情の認識の 1 手法”, 信学技報, **PRU91-24**, (1991).

[59] 間瀬健二, 末永康仁：“フレーム間の動き情報による表情識別の検討”, 43 情処全大 (1991).

[60] 間瀬健二, 赤松茂, 末永康仁：“顔によるインタフェース”, ワークショップ「顔」 (1991).

[61] 間瀬健二, 福本雅朗, 末永康仁：“知的インタフェースにおけるメッセージの 分析と課題”, グラフィクスと CAD シンポジウム, pp.77-84, (1991).

《その他の発表 (共著分) 》

[62] 安田浩, 間瀬健二, 末永康仁：“プログラム管理システム PAGE1 の運用方式”, 30 情処全大, **5T-10**, (1985).

[63] 末永康仁, 間瀬健二：“手の動きの映像処理による情報入力法の検討”, 情処 30 全大, 7M–6(1985).

[64] 玉邑嘉章, 間瀬健二, 小杉信：“地理情報による景観画像合成法の基本検討”, 昭 61 信学全大, **1646**, (1986).

[65] 玉邑嘉章, 河久保秀二, 小倉健司, 間瀬健二：“64kb/s 静止画像提供のための情報作成編集法”, 研究実用化報告, **34**, 10, pp. 1431–1439(1985).

[66] 加藤洋一, 間瀬健二：“階層的計算機ネットワークを利用した動画像処理実験シス テムの構成”, 昭 60 信学情報システム部門全大, **195**, (1985).

[67] 秋本高明, 間瀬健二, 末永康仁：“隣接画素の類似性にもとづく光線追跡法の高速化”, 30 情処全大, **4J–4**, (1985).

[68] 秋本高明, 間瀬健二, 三ツ矢英司：“3 次元メモリ空間による光線追跡法の高速化手法”, 昭 60 信学全大, **1180**, (1985).

[69] 秋本高明, 間瀬健二：“画素選択形光線追跡法による高速画像生成”, TV 学技報, **VVI72–6**, (1985).

[70] 秋本高明, 間瀬健二：“画素選択形光線追跡法における類似性の判定方法”, 昭 60 信学情報システム部門全大, **89**, (1985).

[71] 秋本高明, 間瀬健二：“動きを表現するための顔画像生成方法の基礎検討”, 昭 61 信学全大, **1647**, (1986).

[72] 秋本高明, 間瀬健二, 末永康仁：“形状の自動変形による表情を持つ顔画像の生成方法の検討”, 情処研資, **CG 28–14**, (1987).

[73] 秋本高明, 玉邑嘉章, 間瀬健二, 末永康仁: "顔の動き検出・生成システム - MAGIC による顔画像の高速生成 -", 63 情処全大,1V-8(1988).

[74] T. Akimoto, K. Mase and Y. Suenaga: "Improved pixel selected ray-tracing by using ray-object intersection tree", 1989 信学全大,SD-3-27(1989).

[75] 橋本秋彦, 間瀬健二, 秋本高明: "透視画像を利用した画素選択法", 第 33 情処全大, pp. 2105-2106(1986).

[76] 橋本秋彦, 間瀬健二, 秋本高明: "ボーダー・レイトレーシング法とその評価", 昭 62 信学全大, **1621**, (1987).

[77] 橋本秋彦, 間瀬健二, 秋本高明: "ボーダー・レイトレーシング法", 信学技報, **PRU86-110**, (1987).

[78] 河久保秀二, 小倉健司, 玉邑嘉章, 間瀬健二: "静止画像応答装置の情報作成編集装置構成法", 昭 59 信学全大, **1399**, (1984).

[79] 河久保秀二, 玉邑嘉章, 間瀬健二: "静止画像提供のための情報作成編集システムの構成", 第 15 回画像工学コンファレンス, **13-13**, (1984).

[80] 境野英明, 秋本高明, 間瀬健二, 末永康仁: "映像による頭の動作の検出に関する検討", 信学技報, **PRU88-141**, (1989).

[81] A. P. Pentland and K. Mase: "Lipreading: Automatic visual recognition of spoken words", Technical Report 117, MIT Media Lab Vision Science, Cambridge, MA, USA(1989).

[82] 志沢雅彦, 間瀬健二: "時空間フィルタを用いた多重オプティカルフローの抽出法", 情処研資, **CV62-2**, (1989).

[83] 志沢雅彦, 間瀬健二: "時空間フィルタ法による多重オプティカルフローの抽出", 1989 秋信学全大, **D-161**, (1989).

[84] 志沢雅彦, 間瀬健二: "多重オプティカルフロー抽出における多重度の判定", 信学技報 (1990).

[85] 赤松茂, 間瀬健二, 末永康仁: "人物像を用いる優れたインタフェースの研究", 1990 春季信学全大, **SA-7-1**, (1990).

[86] 尺長健, 間瀬健二, 外村佳伸: "NTT ヒューマンインタフェース研究所におけるコンピュータ ビジョンの研究", 情処研資, **CV-62**, (1989).

[87] 渡部保日児, 間瀬健二, 末永康仁: "CG 表情画像を使ったオプティカルフローによる表情検出の 評価", 1990 春季信学全大, **SD-11-4**, (1990).

[88] 佐藤嘉伸, 間瀬健二: "Gabor ピラミッドの高速計算", 1990 春季信学全大 (1990).

[89] 中嶋, 間瀬ほか: "「グラフィクスと CAD」文献データベース :1989", 情処研資, **CG45-6**, (1990).

[90] 中嶋, 間瀬ほか: "「グラフィクスと CAD」文献データベース :1990", 情処研資, **CG51-6**, (1991).

[91] 柿本, 間瀬ほか: "標準立体データベースについて", 情処研資, **CG53-6**, (1991).

[92] 上田修功, 間瀬健二: "動的計画法による active contour エネルギーの最小化", 91 春季信学全大, **D-554**, (1991).

[93] 上田修功, 間瀬健二, 末永康仁: "弾性制約モデルによる輪郭追跡手法", 91 信学秋季全大 (1991).

[94] 上田修功, 間瀬健二, 末永康仁: "エネルギー最小化に基づく輪郭追跡", 情処研資, **CV73-5**, (1991).

[95] 佐藤敦, 上田修功, 間瀬健二: "動きの滑らかさを用いた腕の運動軌跡の抽出", 42 情処全大, **3D-8**, (1991).

[96] 佐藤敦, 間瀬健二, 末永康仁: "x-t 時空間画像からのロバストな物体抽出法", 91 秋季信学全大 (1991).

[97] 福本雅朗, 間瀬健二, 末永康仁: "画像処理を用いた指示動作検出の実験システム", 91 春季信学全大, **A-251**, (1991).

[98] 福本雅朗, 間瀬健二, 末永康仁: "Finger-pointer: 画像処理を用いた仮想指示棒", 信学技報, **HC91-12**, (1991).

[99] 福本雅朗, 間瀬健二, 末永康仁: "動作と音声を統合したマルチメディアインタフェース", 91 秋季信学全大 (1991).

[100] 福本雅朗, 間瀬健二, 末永康仁: "動画像処理による非接触ハンドリーダ", 第 7 回 HI シンポジウム (1991).