2003－CH－58 (8)
2003／5／30

ビデオ映像による人物間対話解析のための顔領域抽出と顔方向の推定

町野太一　　八村広三郎
立命館大学 理工学部 情報学科

インタビューシーンなどのビデオ映像を用いた、複数人物間の対話を解析するための処理の一部として、人物の顔の向きや表情などによる非言語コミュニケーションに関する情報を動画像処理により抽出する。本論文では、顔領域および顔内部の特徴点の抽出と、これらの情報を用いた顔方向の推定の手法および実験結果について報告する。

Detecting Faces and Face Orientations from a Video Image Sequence for the Discourse Analysis

Taichi Machino　　Kozaburo Hachimura
Department of Computer Science
Ritsumeikan University

Non-verbal communication plays an important role in man to man communications. A purpose of this research is to provide a tool for analyzing non-verbal communications in a discourse between humans. The research focuses on non-verval communications carried by face orientations. This paper describes a video image processing method of detecting human faces and face orientations from a video image sequence.

1　はじめに

人物間の対話解析においては、音声言語による対話の内容を対象とすることが基本である。このためには、一般に、音声認識技術と自然言語処理により、対話の内容と質的な情報の抽出が行われる。言語による会話に対する情報処理研究は古くから行われてきた。音声認識技術の進展により、対話に使われている情報すなわち言語情報が、文字化できるようになったことがその大きな理由である。

しかしながら、人と人の対話、あるいは、対話の概念をより広くとらえた「コミュニケーション」においては、言語によるコミュニケーションだけでなく、言語以外の情報を利用する、「非言語コミュニケーション（ノンバーバル・コミュニケーション）」も大きな働きをする。非言語コミュニケーションでは、声の調子、顔の表情、身振り、ジェスチャなどが重要な要素となるが、さらには、相手との距離のとり方、接触、香り、服装、アクセサリーまでもがコミュニケーションに利用されると指摘されている [1]。

このような非言語コミュニケーションの要素の中で重要なものが、人間の動作やジェスチャによるものであるが、特に顔の表情や顔を含む頭部の動作が重要である。

このことから、本研究においては、人物間の対話の様子をビデオカメラで撮影し、これを画像処理することによって、非言語コミュニケーションの様子を解析し、コミュニケーションの質的情報を得ることを目的とする。対話の様子を映像から求めるためには、それぞれの人物の姿勢や動作も重要な要素となるが、まず、ここでは、対象人物の顔を抽出し、その向きや姿勢、さらには視線の方向など、頭部における特徴を計測・検出することを目標とする。これらを抽出できれば、２人は向き合って対話しているのか、あるいは２人ともまったく別の物を見ながら話しているのか、などの情報を得ることができる。

ここでは、対話の現場でのフィールドワークによって得られるビデオ映像を対象とすることを想定し、移動型の単眼ビデオカメラを用いて撮影された実環境下でのビデオ映像から、複数人物の顔の抽出と顔向きの推定を行うシステムを目標とする。当面は、大学の研究実験室内における対話シーン、および、実際の TV 放送のインタビューや討論番組などで得られる映像を対象とした処理が行えるように検討した。

2 顔画像処理

動画像中から顔の位置を特定し、追跡を行うことはロボットビジョンや、ヒューマンインタフェースの分野で中心的で重要な課題であり、多くの研究が行われている。これらの研究において問題になるのは、環境変化に対するロバスト性とリアルタイム性である。実環境下における顔認識では、背景や照明条件などの周辺環境の変動に対して頑強な動作を保証する必要がある [2]。

顔の検出のためのの手掛かりとしては、顔の肌や髪の色彩情報、動きに関する情報などが用いられることが多い。これらの情報は、背景や照明条件撮影環境により大きく影響され、正確な検出は必ずしも容易ではない。特に移動型のカメラを利用する場合にはその環境変動の影響は顕著であり、これらに影響されないようなロバスト性が求められる。

文献 [3] では、ほぼ正面を向いている顔画像から色情報を用いて肌色と髪の領域を抽出し、さらにエッジ情報から各顔器官の領域を検出する。これらに対して、あらかじめ作成しておいた頭部の３次元モデルのフィッティングを行い、頭部の３次元姿勢を推定することを提案している。しかし、この研究では、画像の背景には別の物体が映りこんでいないこと、１人の顔を対象としていること、また、あらかじめ対象人物ごとの顔の３次元モデルを用意しておく必要のあることなどの問題がある。

また、文献 [4] では、肌色抽出により顔領域を、またその結果を利用して髪領域を抽出し、最後に両目の位置を抽出する。これらの情報に対して遺伝的アルゴリズム（GA) を用いて３次元の頭部モデルを適合させ、頭部の姿勢を推定している。しかしながら、この研究では、背景は既知であって変化しないものとして、背景画像をあらかじめ記憶しておき、実際に撮影される映像から背景を除去していること、また、GA によるモデル適合に時間を要し、ビデオレートでの処理が不可能なことが問題としてあげられる。

コミュニケーション解析の実現のためには、顔内部の各種器官、特に目の領域の抽出が必要である。上記 [3] ではエッジ情報と SUSAN オペレータを用いて、また [4] では、暗部に対する閾値処理によって目の領域を抽出している。さらに、特徴的なものとして、文献 [5] では、目そのものではなく、まず、その間に存在する「眉間」をリングフィルタを用いて抽出し、その位置を元にして両目の位置を抽出することによって、ロバストネスを向上させている。

3 ビデオ画像からの人物顔領域の検出と顔方向の推定

3.1 処理の概要

前述のように、本研究では、最終的には社会調査のフィールドで撮影されたような対話シーンのビデオ映像までを対象とすることを想定している。しかしながら、最初の段階からこのような一般的なものまで対象とすることは困難である。したがって、ここでは、背景に特別な制限を設けないこと、対象とする人数についても制限しないこと、の２点を前提に、一般的な大学の実験室環境で得られるビデオ映像と、テレビ放送番組のインタビューや対談シーンの映像程度までを対象として考えることにした。上述のように人数には特に制限はないが、本研究では対話の解析を主目的としているので、基本的には、ビデオ映像中に収まる人物の顔の数は、３名かせいぜい４名程度になることを想定してる。

全体の処理の流れを図 1 に示す。

3.2 色情報による顔領域の検出

ビデオ映像からの顔領域抽出の手掛かりには、色、形状、動きという 3 つの要素をあげることができる。本研究では、実時間処理に有利であり、また顔領域の大まかな位置決めが可能な色情報を用いることにし、顔領域を抽出するための重要な特徴である肌色情報に注目する。

表色系としては、本研究では、デジタルカラー画像の表現で一般的に用いられている RGB 表色系でなく、これよりも人間の色の知覚特性への適合性が高い YUV 表色系を利用した [6]。

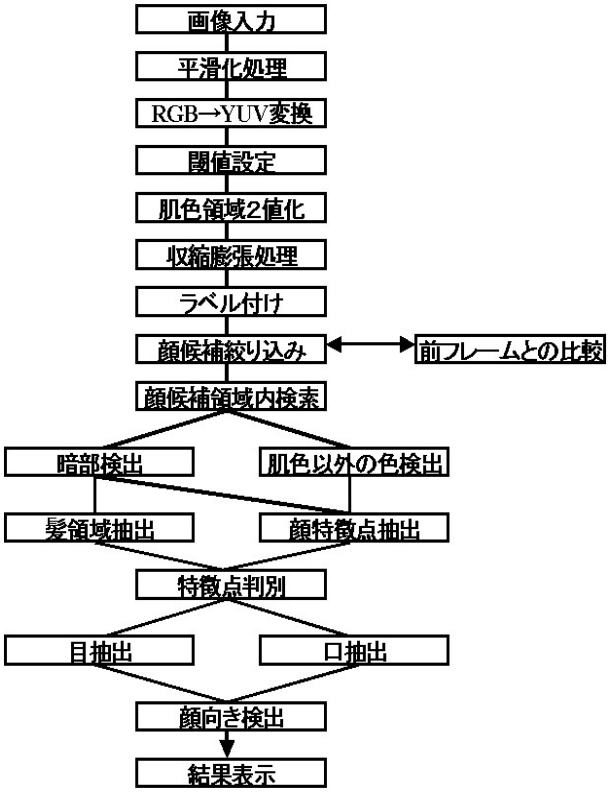


図 1: 処理手順

画像中の各画素の色は YUV 軸で構成される３次元色空間の中にマッピングされるが、Y 軸はその色の輝度（明るさ）に対応し、UV 平面内の位置により、色味（色相）と鮮やかさ（彩度）が決まる。彩度は、UV 平面の原点からの距離により、また色相は、U 軸から計った原点周りの角度により決まる。すなわち、UV 平面上で、色相 h と彩度 s は、それぞれ、以下の式を用いて求める。

$$h = \sqrt{U^2 + V^2}$$

$$s = tan^{-1}\frac{U}{V}$$

なお、対象とするビデオ画像は、見た目にはきれいなノイズの無い映像に見える。しかし、実際の信号としては、各画素における輝度や色には、微細ではあるが、時間的・空間的な変動がのっており、後述する肌色領域の抽出などの際の閾値処理の結果に、強く目立つノイズ領域として残ることが多い。したがって、ここでは、前処理として移動平均による平滑化によりノイズの低減を行う。すなわち、ビデオ入力から得られる RGB 画像の R,G,B の各色成分画像に対して、3×3 の単純移動平均による平滑化を施したものを YUV 変換し、その後の処理に利用する。

3.3 肌色領域の抽出

同じ人種であれば、肌色の変動はそれほど大きくはないので、UV 平面内にある適当な範囲を設定することによって肌色の領域を抽出することができる。

しかしながら、同じ人種であっても、さらには、同じ人の場合であっても、照明などの撮影環境によって肌の色情報は大きく変化する。また、ここでは、一般の環境で得られるビデオ画像を対象とすることを想定しているので、背景に例えば段ボール箱などがあると、これを肌色と混同することがあり得る。肌色の顔領域を正確に抽出するためには、そのための閾値を的確に定める必要があるが、単に色の類似性だけで顔領域を抽出するのは難しい。

したがって、ここでは、後述するように、顔領域候補の抽出の過程では、目や口など、顔の中のいくつかの特徴領域の情報も利用する。一般にこのような特徴領域は肌色の部分よりやや低い輝度を持っている。しかし、この部分においても、肌色領域が照明などの環境条件によって影響を受けるのと同じように、相対的に影響を受けるので、常に固定的な閾値を設定して利用することはできない。

ここでは、対象とするビデオのシーケンスの最初の段階で、画像をモニターしながら、肌色領域と暗部領域を検出するための閾値を設定する方法をとっている。図 2 はこれらの閾値の設定のための画面である。モニタ上に表示される肌色領域と暗部領域を見ながら、その撮影環境に適した、YUV 値に対する閾値を求め、これらを以後の処理で利用する。

3.4 顔候補領域の抽出

肌色領域抽出処理によって得られた 2 値画像には、顔領域だけでなく、背景部分に多くの微小領域を含んでいる。ここでは、このようなノイズ状の微小領域を除去するために、領域の収縮膨張処理を利用する。８近傍での定義で、２回収縮処理を繰り返した後、この結果に対してさらに２回膨張処理を繰り返す。この処理により、背景中のノイズ成分を除去する。

膨張収縮処理結果の２値画像内に残った孤立領域に対して、ラベリング処理を行う。各ラベル付き孤立領域は、人物の顔候補領域となる。これらの各顔候補領域の外接矩形を求め、この矩形の縦横比と面

図 2: 色抽出のための閾値設定画面

積で候補領域の絞込みを行う。

まず、縦横比 (Y_s/X_s) が、1.6 以上あるいは 0.5 以下の領域は顔領域である可能性が低いとみなし、顔領域の候補から除外する。また、画面全体の画素数の 14%程度以下の小さな領域も候補から除去する。さらに、顔領域内には、目、口や髪の毛などの、いくつかの特徴的な暗部領域が存在するものとし、矩形領域内にこのような暗部が全く存在しないものは、顔候補領域として残さない。

さらに、直前のビデオフレームで求まっている顔領域との位置関係などの情報を利用して顔候補領域の絞り込みを行う。すなわち、現フレームの外接矩形内に、前フレームでの顔領域の外接矩形の重心が入っておれば、現フレームの外接矩形は顔候補領域とする。これによって、後ろをすばやく横切る人物の顔の影響を防ぐことができる。

以上のようにして、いくつかの顔候補領域の中から、顔領域だけを取り出すことができる。画像中の顔領域は１つとは限らず、条件を満たすものがあれば複数でも抽出される。図 3 は、このようにして抽出した顔領域（この場合は１つ）の外接矩形を示している。

3.5 顔方向の推定

3.5.1 髪の毛領域と顔方向の仮推定

顔領域の外接矩形の中で、目、口、髪等に相当する領域を抽出する。このため、図 2 で設定した閾値を用いて外接矩形内での暗部を求める。図 4 に暗部領域の抽出処理例を示す。

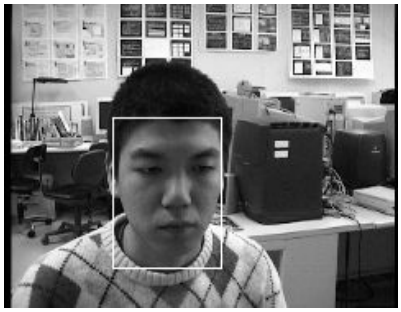


図 3: 顔領域の外接矩形

図 4: 外接矩形内の暗部の抽出

抽出した暗部領域から、髪の毛領域を抽出する。すなわち、上で求めた外接矩形の上半分の部分を探索範囲とし、この中に存在する暗部領域を求め、これが、以下の式の条件を満たせば、その領域を髪の毛の領域とする。ここで、X_s は顔候補領域の外接矩形の横幅を、Y_s は同様にその高さを表しており、X_k、Y_k は、対象の暗部領域の、それぞれ横幅と高さを表している。式中の定数の値は実験により設定した。

$$\left.\begin{array}{l} X_k > X_s * 0.4 \\ Y_k > Y_s * 0.17 \end{array}\right\}$$

髪領域抽出の処理結果の例を図 5 に示す。

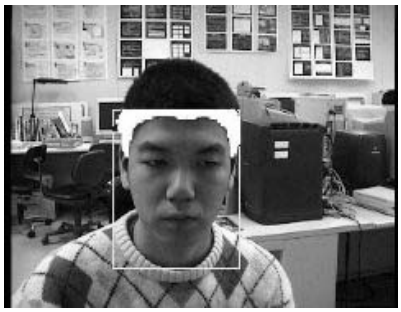


図 5: 髪領域の抽出

次に、この髪の毛領域の分布状況により顔方向の仮推定を行う。図 5 のような髪の毛領域の画像に対

して、顔領域外接矩形内上部において、左半分、右半分のそれぞれで髪の毛領域の面積を求める。この面積の大小により、顔の向きを「仮に」推定し、この後の、目や口の領域の抽出に利用する。右側の面積が左側の面積より大きければ顔の向きは（向かって）左向き、左側の面積が右側の面積より大きければ（向かって）は右向きとする。

3.5.2 顔特徴点候補の抽出

次に、右目と左目および口唇に対応する領域（顔特徴候補領域と呼ぶ）を抽出する。

まず、外接矩形内において、図 2 で設定した肌色の閾値を用いて、肌色ではない色部分の抽出を行う。肌色以外の領域を抽出した結果が図 6 である。

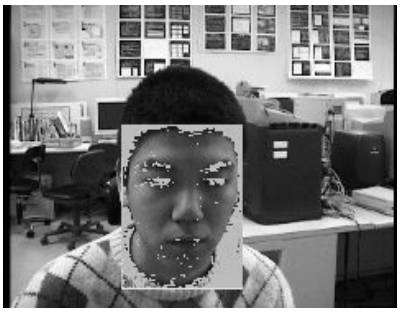


図 6: 肌色以外の領域の抽出

外接矩形内で、暗部抽出で得た結果 (図 4) と肌色以外の領域 (図 6) の画素単位の論理積をとり、残った領域を顔特徴候補領域とする。なお、この時、髪の毛の領域 (図 5) は除去する。このようにして抽出した顔特徴候補領域を図 7 に示す。

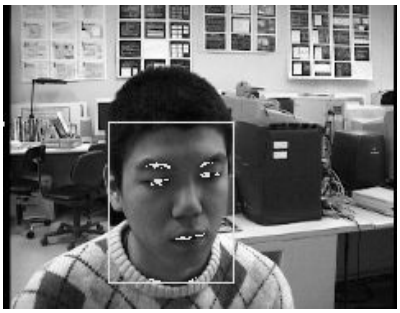


図 7: 顔特徴候補領域の抽出

顔方向の仮推定結果に基づき、目の領域の探索範囲を決める。右目領域は図 8 示す領域を探索範囲とする。探索範囲内で、向かって一番左側に存在する特徴候補領域を右目領域とする。また、ある範囲内に 2 つの候補領域が存在し、これらの領域の重心の間の縦横の座標値の差が一定値以内である場合は、上の方の領域は眉毛の領域とし、下の方の領域を目の領域とする。最後にその領域の重心位置を求め、目に対応する特徴点とする。左目領域についても、同様に行う。

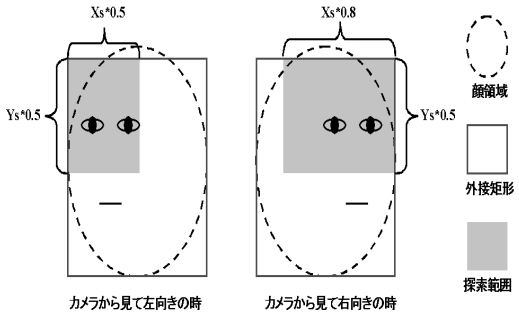


図 8: 右目探索範囲

目領域の場合と同様に、顔方向の仮推定の結果に基づき、口唇領域の探索範囲を決める。図 9 を探索範囲とする。この探索範囲内に存在する顔特徴候補領域のうち、形状が横長のものを口唇とする。ここでは、探索範囲内での顔特徴点候補の横幅を X_m、高さを Y_m とするとき、$(X_m/Y_m) < 2$ のものを口唇領域としている。このようにして抽出された口唇領域の外接矩形の重心を求め、これを口に対応する特徴点とする。

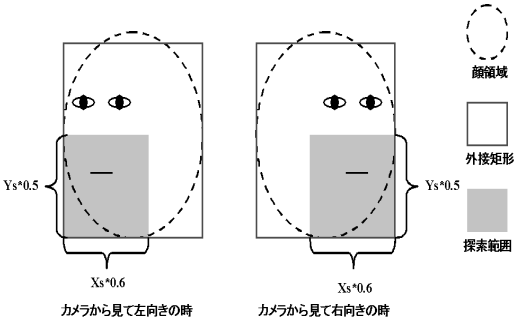


図 9: 口唇領域の探索範囲

顔が横方向を向いている場合、目に対応する特徴点が 1 つしか抽出できない場合がある。この場合、この特徴点が対象人物の右目か左目かを判定する必要がある。ここでは、口の特徴点の位置との関係に基づき目の左右を判定する。口の特徴点より、向かって右側に目の特徴点があれば、その特徴点を左目とする。向かって左側に目の特徴点があれば、その特徴点を右目とする。

3.6 顔方向の決定

いままでに求めた、顔領域内の 3 つの特徴点の位置情報と外接矩形の形状を利用して、最終的な顔の方向を決定する。得られた顔の外接矩形とその中の特徴点の例を図 10 に示す。

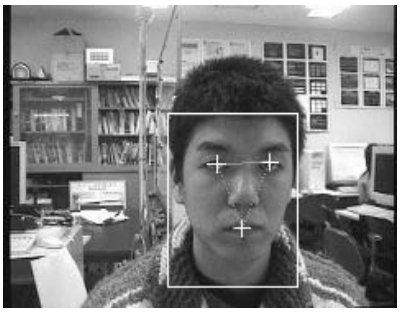


図 10: 顔特徴点の抽出結果

これらの顔特徴点の位置から、図 11 に示すような場合分けを行い、これに従って顔方向の推定を行う。図に示すように、現在のところ、7 方向の顔の向きを推定する。同図においては、右目、左目の特徴点位置を、それぞれ、(x_reye, y_reye) と (x_leye, y_leye) としている。Y_s、X_s は、それぞれ、外接矩形の高さと幅であり、X_min、X_max は、それぞれ、外接矩形の左辺、右辺の X 座標値である。

結果は、画面上の顔領域を取り囲む矩形領域の外側に方向を示す三角形の印で表示される。また、「右向き」、「正面」、「左向き」などの文字でコンソール上にも表示される。なお、ここでは、顔の向きは画面に向かってカメラから見たときの向きの右、左を示している。正面を向いていると判断された場合は三角形の印は表示されない。

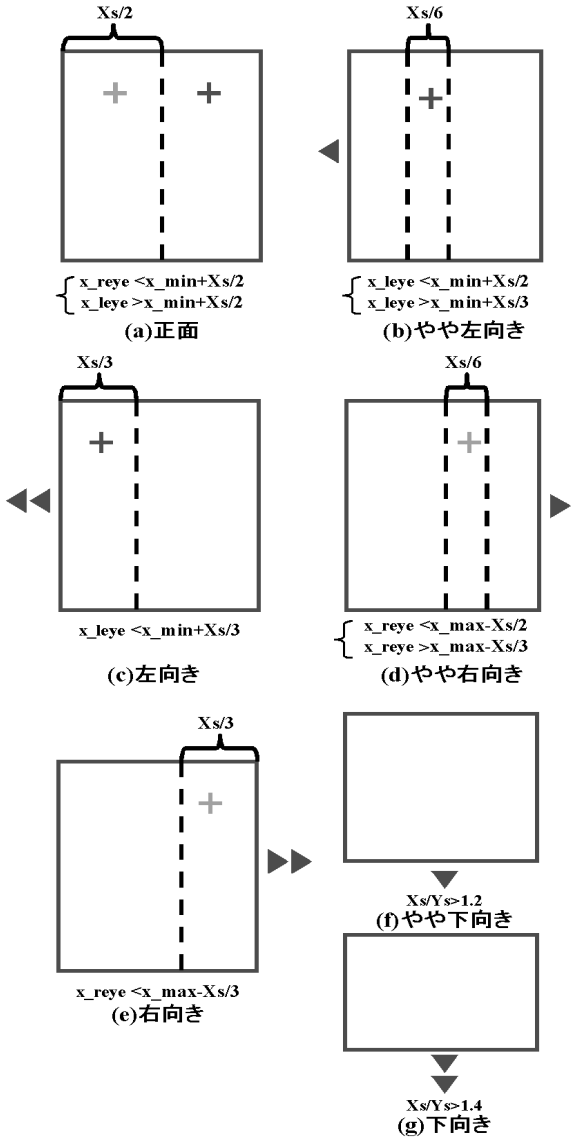


図 11: 顔向き方向の場合分け

4 実験結果と考察

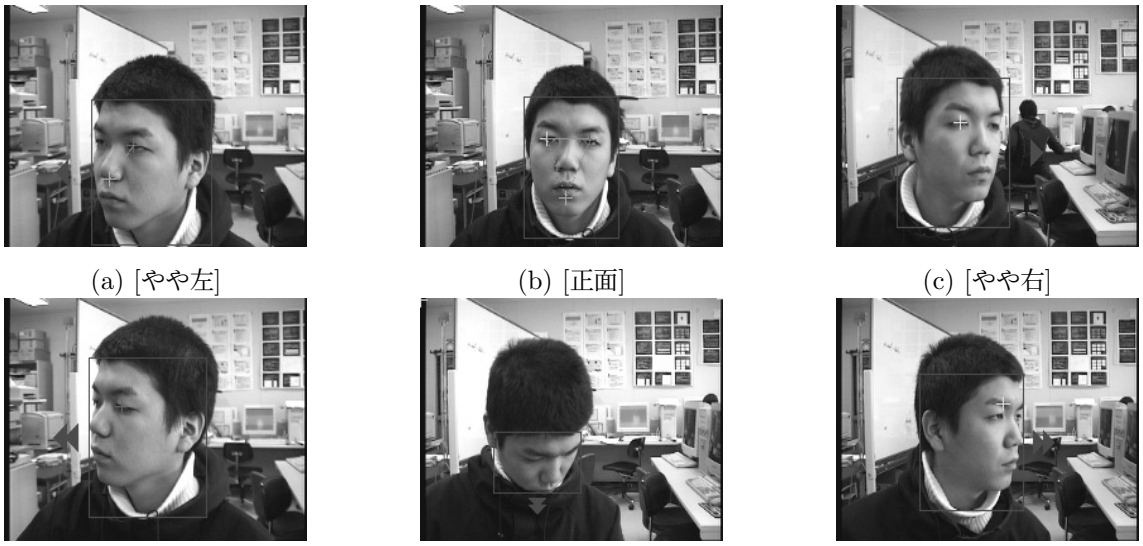
4.1 実験結果

できるだけ、一般的な状況での実験とするため、大学の研究室内で特別な照明などを行わない、実環境において実験を行った。背景には、人物の肌色と類似した色の、段ボール箱、壁、床なども存在している。

処理システムとしては、Intel Pentium 1.0 BGHz プロセッサを用いた PC で、OS には Linux を、またビデオ入力ボードからの画像入力には Video4Linux システム [7] を利用した。ビデオ入力の解像度を 240 × 180 に設定して行い、この場合の処理速度は約 20 フレーム/秒で可能であった。

図 12 に研究室内で 1 人の対象人物を撮影した映像についての実行結果を示す。この場合、顔領域の外接矩形内から顔特徴点を抽出し、正しく顔方向の推定が行えているのがわかる。実際の映像では、これらは色分けして表示されているが、白黒の印刷では判読がしにくいので、本論文中では、それぞれの図の下に [左]、[右] などの文字で結果を表示している。

画面内に 2 人以上の人物が写っている状況での実験の結果を図 13 に示す。左右 2 人の人物の顔領域を正しくとらえ、それぞれの顔の向きもほぼ正しく推定していることがわかる。また同図 (a) では、右側の



(a) [やや左] (b) [正面] (c) [やや右]

(d) [左] (e) [下向き] (f) [右]

図 12: 実験結果

人物の右手は肌色領域として検出しているが、顔として誤認識することはなかった。さらに同図 (c) では、中央の人物も、「やや下向き」と判定されている。

次に、実際のテレビ放送の映像に対して処理を行った結果を図 14 に示す。同図 (a)〜(d) では、顔領域の外接矩形内から顔特徴点を抽出し、正しく顔方向の推定が行われているのがわかる。一方、図 14(e) では、向かって右側に、大きな黄色の背景領域があり、黒い髪の領域がないこともあって、顔候補領域が抽出できていない。同図 (f) の場合は、対象人物の服の胸元が開いているため、顔領域を縦長に広くとっており、また、ちょうど目を閉じた瞬間であるため、口を目と混同する結果となっている。

4.2 考察

ビデオによって撮影された人物の顔の向きを推定する方法として、対象人物の髪の領域や目、口唇の領域を抽出し、それらの重心位置の組み合わせで推定するという、簡便な方法を試みた。実験により、大学の研究室内という環境での、雑然とした背景の中でも比較的良好に顔の領域と方向の抽出ができることを確認した。また、将来、街頭でのビデオインタビューなどでの応用を想定し、類似の形態のテレビ放送の番組からのビデオ映像に対しても適用し、ある程度の結果が得られることも確認した。

しかしながら、もちろん、すべての場合について完全な処理結果が得られるわけではない。

特徴点の分布状況から顔方向を判定するための、図 11 の場合分けのパターンについては、きわめて単純化したものであり、これでだけで十分であると考えてはいない。

当然のことながら、将来的には、対象人物がカメラから見て反対方向を向くような場合についても対応する必要もあるであろう。さらに、今回の処理手法では、頭部を傾ける、すなわち「首をかしげる」ような動作も想定していないが、これも対話の質的解析には重要なファクターであろう。

今回は、顔候補領域内の、目や口などの特徴的器官に対応する領域を抽出し、顔候補領域の絞込みと顔方向の推定に利用したが、これらの器官に対応する領域の抽出の精度が不十分な場合もあった。今後、これらの顔領域内の特徴的な器官をより的確に抽出して顔領域の判定の確度を向上させることを検討する必要がある。

また、現在のところ、ビデオフレーム間の相関は簡単なものしか利用していないが、前フレームでの結果を利用しながら、顔領域を追跡することも考慮に入れるべきであると考えている。

5 おわりに

本研究では、人物間の対話の様子を解析する際の手がかりの一つとして、人物の顔の向きに着目し、実環境下で撮影されるビデオ映像において、リアルタ

(a) [右]、[やや左]

(b) [やや右]、[やや右]

(c) [右]、[やや下向き]、[正面]

図 13: 複数の人物を含む場合の結果

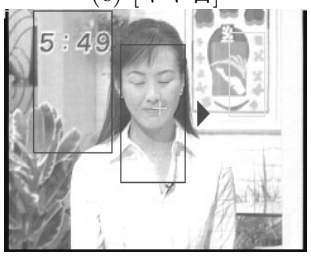
(a) [右]

(b) [やや左]

(c) [やや右]

(d) [やや左]

(e) [下]



(f) [やや右]

図 14: テレビ放送による映像での実行結果

イム（ビデオレート）で人物の顔の抽出と、顔の向きの推定を行う手法を提案した。特に、リアルタイム性を重視し、できるだけ単純な処理手法を用いることによって、高速な顔領域の検出と顔向きの推定を可能にした。使用した画像処理アルゴリズムはいずれも基本的なものばかりであるが、一応の成果は出せたと考えている。

2 人または複数人の対話シーンから、個々人の顔向きの変動データを求め、これより、その対話形態の変化、ひいては対話の質的解析にまでつながる情報を抽出することが本来の最終的目標である。今後、更に顔向き抽出の精度を向上させて、信頼できるデータの抽出を可能にし、発話音声の言語処理から得られる情報と組み合わせることのできる、対話分析システムのためのシステムの一部へと発展させることを考えている。

参考文献

[1] 黒川隆夫: ノンバーバルインターフェース, オーム社, 1994.

[2] M.-H. Yang, D. J. Kriegman, and N. Ahuja: *Detecting Faces in images: A Survey*, IEEE Trans. PAMI, Vol.24, No.1, 2002.

[3] 稲田純也、呉海元、塩山忠義: 顔特徴点を用いた頭部の動きの追跡及び 3 次元姿勢推定、信学技法、Vol.PRMU2001, No.195, pp.9-16, 2002.

[4] 阿部友一、萩原将文: 単眼視動画像からの人物頭部動作の解析と認識、信学論、Vol.J83-D-II, No.2, pp.601-609, 2000.

[5] 川戸慎二郎、鉄谷信二: 目のリアルタイム検出と追跡、信学技法、Vol.PRMU2000, No.63, pp.15-22, 2000.

[6] S. J. Sangwine and R. E. N. Horne: The Colour Image Handbook, Chapman & Hall, 1998.

[7] 飯尾淳: Linux による画像処理プログラミング, オーム社, 2000.