

顔画像処理による頭の動作識別法

間瀬 健二

末永康仁

NTT電気通信研究所

人間の動作をTVカメラ等で計算機に取り込んで、そこから意味のある動きを抽出して解釈させることは、手でキーボードやマウスを扱う煩わしさから私達を解放してくれるという、マンマシンインタフェースの向上に欠かせないものである。

本文では、ディスプレイに向かって作業をする時の基本的な命令を、頭の動画から抽出する方法を検討し実験を行ったので報告する。抽出を試みた命令動作は、計算機からの質問に答えるための「ハイ/イイエ」と、スクロールやズームを行うためのディスプレイ上の注視している位置を与える「上・下・左・右・前（拡大）・後（縮小）」である。これらの命令を理解する手法を、複雑な処理を使わずに単純な画像処理の組合せで実現した。

A Study on Head Motion Detection
by Facial Image Processing

Kenji MASE Yasuhito SUENAGA

NTT Electrical Communications Laboratories, NTT
1-2356, Take, Yokosuka-shi, Kanagawa, 238-03 Japan

Human motions enhance communication meaningfully. If machine can directly read human intention from one's head or hand motion with image sensor, man-machine interaction will be drastically improved.

A man-machine interaction method based on head motion detection is presented here with some experimental results. A simple image processing for a series of TV frames detects eight commands: (1) up, (2) down, (3) right, (4) left, (5) zoom-up, (6) zoom-down, (7) yes and (8) no.

1. はじめに

人間の動作をTVカメラ等で計算機に取り込んで、そこから意味のある動きを抽出して解釈させる試みが行われている[1],[2],[3]。唇の動き[1]や身振り[2]から言葉や意味を抽出して計算機と会話を試みたり、目や口の動きで表情や感情を解釈する[3]、[4]試み等である。いずれも、キーボードやマウスに頼らずに計算機に命令や情報を与え、マンマシンインタフェースの向上を図ろうとするものである。

人間同士の会話において、動作や表情は、言葉と同じくらいに重要な役割を果たしている。幼児にいたってはまだ言葉を満足に話せない時期に、頭を振って意思を表示する。各国の習慣によって、同じ動作でも全く異なる意味を持っていることはあるが、例えば、注視する方向に顔を向けるなど、必ず同一の意味を持つものも多い。これらの動作を計算機に理解させることは、より自然なマンマシンインタフェースを作るうえで重要なポイントである。

本文では、図1のようにディスプレイに向かって画像データベースを検索するユーザーの命令を、頭の動画像から抽出する方法を検討し実験を行ったので報告する。抽出すべき命令動作は、計算機からの質問に答えるための「ハイ/イイエ」と、スクロールやズームを行うためのディスプレイ上の注視している位置を与える「上・下・左・右・前(拡大)・後(縮小)」である。これらの命令を理解する手法[5]を、複雑な処理を使わずに単純な画像処理の組合せで実現した。以下、まず2章では、顔画像の処理において利用できる性質について考察する。3章では処理の内容を、画像処理部と動作解析部に分け説明する。4章では実験結果を示

す。

2. 顔画像の性質

顔は人間にとって非常に親密な対象である。そのため頭の動きや表情を読み取るときに、私達は知らず知らずのうちに、頭や顔の解剖学的な構造を利用している。例えば、顔は頭の前面にあり、後頭部とは容易に区別できる。私達は、相手の顔をはっきりとみなくとも、相手の顔の向きや動きを掴むことが出来る。また動作には習慣のように後天的なものと、刺激に対する自然な反応(反射)とがあると考えられる。これらも、相手の考えや状態を知る上で利用される。

一方、顔画像からの動作抽出処理においては、1枚1枚の画像を処理して特徴点を抽出[6]した後、その特徴点の動きの特徴をとらえるという手法が基本となる。ここで、認識すべき動作の種類が限られてくると、おのずと動きの特徴は簡素化でき、必要な特徴点の数も少なく済む。従って、形状や動きの性質などから必要最小限のものだけを使って、動作識別を行うような処理系を考えることができる。そこで、動作に関する性質と形状に関する性質とにわけて、どの様な性質が利用できるかを考察する。

なお、認識すべき意味(あるいは命令)は、

- ①「ハイ」、②「イイエ」、
 - ③「右スクロール」、④「左スクロール」、
 - ⑤「上スクロール」、⑥「下スクロール」、
 - ⑦「ズームアップ」、⑧「ズームダウン」、
- の8種類とする。

2. 1 動作に関する性質

上記の8種類の意味をあらわす動作には、表1のようなものが考えられる。

表 1 意味を表す動作

意 味	動 作 (→理由または状況)
「ハイ」	<u>頭を縦に振る。</u>
「イイエ」	<u>頭を横に振る。</u> そっぽを向く。
「右スクロール」	<u>画面の左側に視線を向ける(左を向く)。</u> →左隅に注目 左から右へシャープに振る。→アゴで頁をめくる感じ。
「左スクロール」	<u>画面の右側に視線を向ける(右を向く)。</u> →右隅に注目 右から左へシャープに振る。→アゴで頁をめくる感じ。
「上スクロール」	<u>画面の下側に視線を向ける(下を向く)。</u> →下隅に注目
「下スクロール」	<u>画面の上側に視線を向ける(上を向く)。</u> →上隅に注目
「ズームアップ」	<u>画面に近付く。</u> →細部を調べる。
「ズームダウン」	<u>画面から離れる。</u> →大局的に眺める。

(下線) が今回対象とした動作。

1つの意味を表すにも、いくつかの動作がありうる。ここでは簡単で、しかも他との区別が容易である動作を抽出して、意味を識別することにする。つまり表1の下線で示したものを使う。

例えば、「イイエ」の場合には、頭を横に振ったり、そっぽを向くという動作がある。これらは比較的似かよった動作である。しかし、そっぽを向く動作は、個人によって向きが異なるし、他の動作との区別が困難である。「右スクロール」のように同じ意味を表しても動作の方向が正反対のものもある。「右スクロール」上段の、左隅に注目するような動作は、固定された大スクリーンを、自由に動かせる窓越しに見るときの窓に対する動作になろう。このとき、視線はみようとすものを追いかけることができる。一方、同下段の動作は固定された窓の向こうのスクリーンを動かすことになり、視線は目的でない1点を注視してスクリーンを引っ張ることになってしまう。

この様に識別しようとする意味に対する動作を考えると、前段の画像処理において取り出すべき頭の動き情報の

種類が次のように限定できる。

- i) x, y 軸回りの回転角の変化
- ii) z 軸方向の移動量の変化

ただし図2に示す様に、水平方向をx軸、垂直方向をy軸、カメラの光軸方向をz軸とする。

2. 2 形状に関する性質

動き情報を抽出するための形状に関する特徴点(特徴量)について考察する。顔の形状に関する性質をまとめると表2のようになる。以下に述べる考察により、表中から、①首が細い、②頭は色の違う顔と髪でつくられる、という2つの性質に注目すれば、上で述べた必要な動き情報を少ない処理で得ることが出来る。特徴点は顔の重心位置となる。

(1) 頭部の抽出

頭の部分を胴から区別するために、顔画像が首の所では細くなる性質を使う。

(2) 回転角の変化

顔そのものの回転、顔の中の目、口などの動きで示される回転等があ

る。目や口の抽出には相当の処理が必要な上、これらを抽出する前には、まず顔の範囲を決定する処理をおこなっており、顔だけの回転を抽出した方が処理量は少ない。顔を頭部の髪の毛のない部分と考えれば、肌と髪を色空間で分離することで顔の抽出処理が可能である。

回転角の決定は、抽出される顔が面積をもっていることと、輪郭の抽出精度をことさらに上げる必要がないことから、顔の面積重心の動きで表す。ただし、顔は楕円体状の頭の表面についているため、重心の動きがそのまま厳密な回転角を与えることにはならない。本目的のために必要なのは回転角の変化であるから絶対的な回転角は不要である。他の理由で必要であれば、頭の形状に似せた三次元モデルを導入することによって解決できると考える。

(3) z 方向の移動量

TVカメラによる撮影が頭の透視像を作ることを利用する。すなわち、頭が近付けば大きく写り、遠ざかれば小さく写ることになる。(1)で抽出されている頭部の大きさを調べてカメラモデルに当てはめれば、次式により移動量を得ることが出来る。

$$z \text{ 移動量} = (z_1 - z_0)$$

$$= \frac{(r_0 - r_1)}{r_1} z_0$$

z_0 : 基準位置

z_1 : 移動位置

r_0 : 基準位置における
基線の画像上の長さ

r_1 : 移動位置における同長さ

表 2 形状に関する性質

部 品	性 質
頭	首で胴体と接続。 <u>顔部と頭髪部からなる。</u> 卵に似た形状である。
首	頭と胴体を接続。 <u>頭や胴体より細い。</u>
顔	顔面とその上の目、鼻、口、耳、眉からなる。 まれにひげが鼻下等にある。 目、鼻、口、耳、眉の位置関係は固定的。
頭髪	形状、長さ、色などまちまちである。 <u>頭部の限定された位置。</u>
その他	眼鏡をかける事がある。 化粧により実際の色と違うことがある。

下線が今回利用した性質

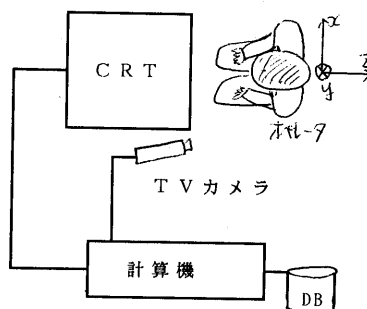


図 1 画像データベースの検索モデル

3. 頭の動作識別の処理

図1のようにCRTの正面に坐った人物の上半身をCRT横のTVカメラから撮影する。実験では、2-3秒の動画像から一つの動作を取り出して意味の識別を行う。すなわち全体の処理の流れは図2のようになる。ここで背景画像の入力は人物像の切り出しを容易にするために行う。

以下、画像処理部、動作識別部および意味抽出部の処理概要を述べる。

3. 1 画像処理部

形状に関する性質をつかって、後段の処理に必要な特徴量は、頭の位置（特にz方向）、および頭の回転角（顔の重心の移動量）の2個である。そこで、各入力画像（原画像）から以下の手順によりこれらの特徴量を抽出する。

step 1) 原画像と背景画像の差分をとり、人物像のシルエット画像をつくる。

step 2) シルエットから頭（首から上）の領域を自動的に切り出して、頭の大きさと重心を求める。頭の大きさからz方向の位置を求める。

step 3) 頭の領域の中から顔の領域を自動的に切り出して、顔の大きさと重心を求める。頭と顔の重心の位置関係から回転角を求める。

なお、これらの処理を可能とするため、撮影に関して以下の簡単な条件を置く。

条件1「頭の写らない背景画像が利用できる」、

条件2「頭の基準位置は通常、画像のほぼ中心にあり、たとえ動いても肩から上の上半身が画面内にある」。

頭と顔の切り出しは次のような処理を行う。

(1) 頭の切り出し

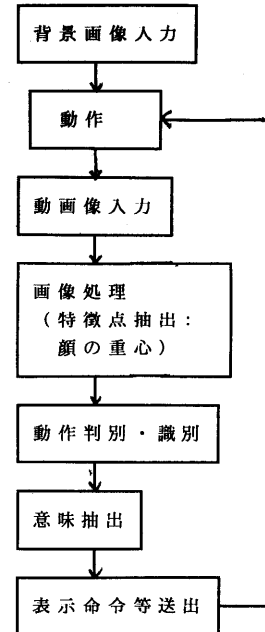


図2 動作識別処理の流れ

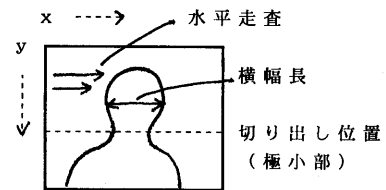


図3 シルエット画像からの頭部切り出し

正面からみると、上半身像は首の所で細くなる。図3のようにシルエット画像を上から下へ水平走査して、各走査線上の頭部の横幅長を調べ、それが極小となるy座標を頭の切り出し位置とする。頭部領域はシルエット画像をしきい値処理して得る。

(2) 顔の切り出し

顔の切り出しを動画像の各フレームに対して順番に行う。図4に処理手順を図示する。第1フレーム（正面を向いた画像）だけはしきい値自動設定のための学習用サンプル及び動作に対する基準画像として処理する。

① 各フレームのシルエット画像をつかって、原画像の中の頭部領域に相当する部分を切り出す（頭部画像とよぶ）。第 n フレーム ($n > 1$) での切り出し時には第 ($n - 1$) フレームの境界座標を使って、その周辺だけを探査する。走査範囲の減少による高速化とともに、切り出しの精度をあげることが出来る。（図 4 - 2）

② 第 1 フレームに対しては、切り出した部分の濃度ヒストグラムをとり、大津の方法 [7] により頭部と顔部を分けるしきい値を求める。

③ 頭部画像において頭部領域の両端から走査して、②のしきい値による最初に見つけた境界を顔部の両端点とする。（図 4 - 3）

さらに切り出された領域から、各部の面積重心を計算して、移動量と回転角を出す。

（3）動きの計算

頭部の重心を (lh, jh)、顔部の重心を (lf, jf)、頭部の面積 S とするとき、次式で頭の各軸回りの回転 R_x, R_y と z 軸方向の移動量 T_z を計算する。

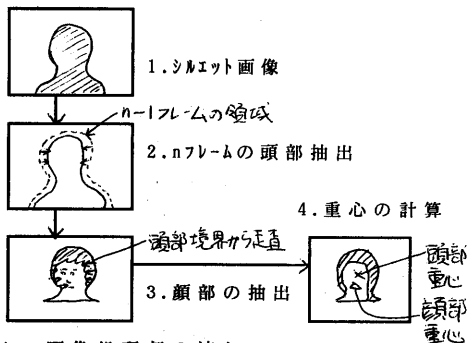


図 4 画像処理部の流れ

$$R_x = \tan^{-1}(K * ((lh - lf) - (lh0 - lf0)) / \sqrt{S}) * \pi$$

$$R_y = \tan^{-1}(K * ((jh - jf) - (jh0 - jf0)) / \sqrt{S}) * \pi$$

$$T_z = (S0 - S) / S0$$

ただし ($lh0, jh0$), ($lf0, jf0$), $S0$ は第 1 フレームの特徴量、 K は定数を示す。

3. 2 動作解析部

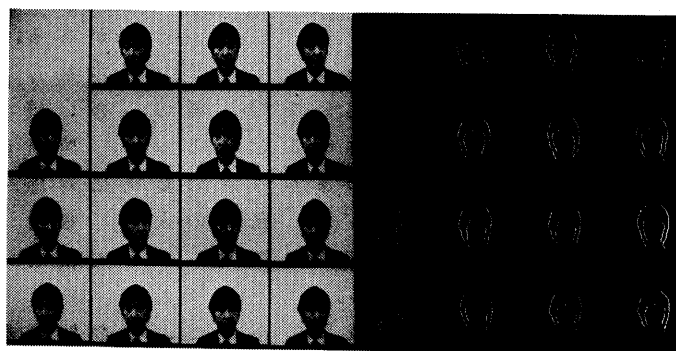
画像処理部で得られたパラメータ: R_x, R_y, T_z の変化を用いて、「縦に振る」、「横に振る」、「横を向く」、「上（下）を向く」、「顔が近づく（遠ざかる）」、「その他」の動作の識別を行う。まず、入力した区間（2～3秒）内に各パラメータ P が 2 つのしきい値区間 ($Tmin(P), Tmax(P)$) をはずれた場合にフラグ $Fmin(P), Fmax(P)$ をたてる。すなわち、 $P < Tmin(P)$ のとき $Fmin(P)$ を "TRUE"、 $P > Tmax(P)$ のとき $Fmax(P)$ を "TRUE"、その他の場合は "FALSE" とする。ただし $Tmin(P), Tmax(P)$ はパラメータ $P: P = \{R_x, R_y, T_z\}$ のしきい値、 $Fmin(P), Fmax(P)$ はパラメータ $P: P = \{R_x, R_y, T_x\}$ のフラグである。動作はこれらのフラグの組みあわせで判別する。判別式は次のものを用いる。

- 1 近づく = $Fmin(T_z)$;
- 2 遠のく = $Fmax(T_z)$;
- 3 横振り = $Fmin(R_y) \cdot Fmax(R_y)$;
- 4 右向き = $Fmax(R_y) \cdot Fmin(R_y)$;
- 5 左向き = $Fmax(R_y) \cdot Fmin(R_y)$;
- 6 縦振り = $Fmin(R_x) \cdot Fmax(R_x)$;
- 7 上向き = $Fmax(R_x) \cdot Fmin(R_x)$;
- 8 下向き = $Fmax(R_x) \cdot Fmin(R_x)$;

なお上記の数値は判別式を利用する優先順位（1 が優先）を示す。また、 $\cdot$ は論理積を、 x は否定を表す。

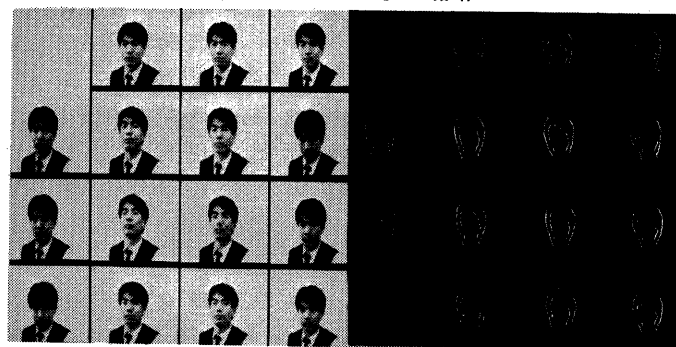
4. 実験

顔を動かしたところを TV カメラで撮り、大きさ 512×512 の画像を 1 枚



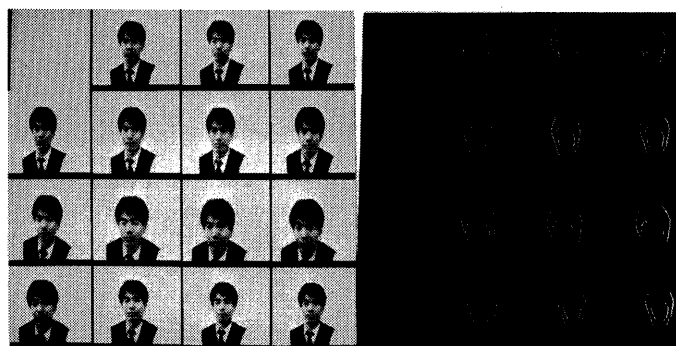
原画像(a) 頭、顔領域抽出結果(b)

図5 「イエエ」の動作



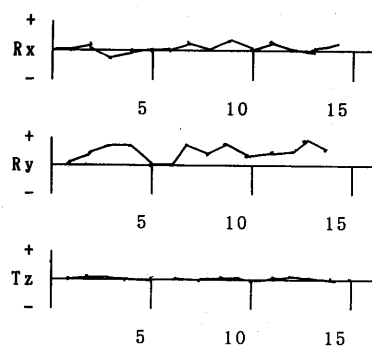
(a) (b)

図6 「ハイ」の動作

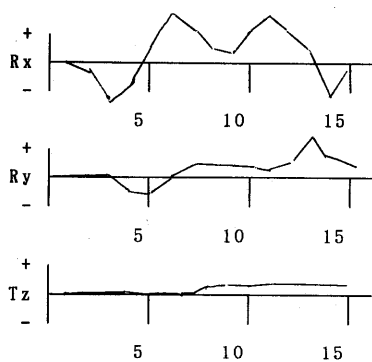


(a) (b)

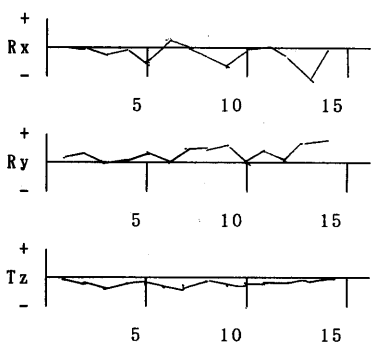
図7 「ズームアップ」の動作



(a) 「イエエ」

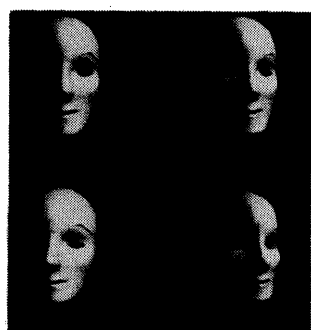


(b) 「ハイ」



(c) 「ズームアップ」

図8 Rx, Ry, Tzの時間的変化



2	3
1	4

図9 動きの再現(図5 1~4フレーム)

(約0.2秒/フレーム) 取り込む。これを128×128にサブサンプルして入力動画像として特徴量の抽出を行い、動きの分類を行った。図5(a,b)は「イエ」のときの原画像(a)と頭、顔領域抽出結果(b)である。同じく図6(a,b)は「ハイ」、図7(a,b)は「ズームアップ」の場合である。

いずれも頭部は原型に近い形で抽出されている。顔部は眉、目、鼻などの影響で輪郭が不安定であるが、動きの計算を重心で行うため、その影響は小さい。顔領域が頭に対して小さめなのは頭の境界からマージンをとって抽出しているためである。また本実験では抽出処理を下方向への水平走査で行っているため髪のみだれが処理を失敗させる原因となっている。これについては改良の余地がある。

図8(a-c)はRx, Ry, Tzの時間的变化をグラフ化したものである。この図からも頭の下上, 左右, 前後の運動をよみとることができる。図9は図5(a)の第1~4フレームの動きを再現したものである。[8]

実際の処理は汎用ミニコンVAX11/780(VMS/C言語)で行った。処理時間は1フレームあたり約0.18秒(第1フレームを除く、ディスクI/O除く)であった。

5. まとめ

頭の動きを簡単な処理で認識し識別する手法を提案し、その実験結果を示した。対象とする動きを限定することによって、処理を簡単にし、高速処理を目指した。対象とした動きは基本的ではあるが重要なものであり、対話をしているときなど、動きのかなりの部分をカバーできると考える。処理手順が単純であるため、パソコンでの実時間処理の可能性もある。現在、画像処理部、動作認識部の何れも認識率、正

答率は十分でなく改良の余地がある。

第1フレームを学習サンプルとして、顔部の切り出しを行う際のしきい値の自動設定を行うことによって、照明環境や、対象のあかるさの変化に対応できる処理とした。任意の人物(眼鏡、髭、長髪)、背景に対しても実験を行う必要がある。

今後は、表2にあげたような性質を画像処理、動き識別においてプログラム化したり知識として利用する事によって、細かな動きや、表情を認識できるシステムの構築が望まれる。また、連続した複数の動作の分離識別のために、しきい値処理と条件判別だけでなく、パラメータの変化を標準パターンと比較するパターンマッチングを利用する必要がある。

(謝辞) 日頃、貴重な御意見を頂く、NTT複合通信研究所知的画像処理研究グループの皆様にご感謝いたします。特に動きの視覚化のための合成画像作成にご協力頂いた同グループ秋本高明研究主任にご感謝いたします。

(参考文献)

- [1] E.D. Petajan: "Automatic Lipreading to enhance speech recognition", IEEE, pp.265-272 1984.
- [2] J. Loomis, et.al: "Computer Graphics Modeling of American Sign Language", ACM-Siggraph, 17, 3, pp.105-114 1983
- [3] 石井、岩田: "濃淡画像による顔の表情の自動認識", 情処29回全大4M-12
- [4] 石井、岩田: "動画像処理による人の視線の追跡", 情処30回全大6M-8
- [5] 間瀬、末永: "顔画像の動き検出の一手法", 情処30回全大7M-5
- [6] 坂井、長尾、金出: "計算機による顔写真の解析", 信学論(D), 56-D, 4, pp.226-233, 1973
- [7] 大津: "判別及び最小2乗基準に基づく自動しきい値選定法", 信学論(D), J63-D, pp.349-356, 1980
- [8] 秋本: "あごの3次元モデルと表皮の自動変形による表情と動きを持つ顔画像の生成", NICOGRAPH'86