コンピュータビジョン 98－3
（1996. 1. 18）

動き情報を用いた唇の抽出法

岡野 健治　宮崎 敏彦　奥村 晃弘　藤井 明宏

沖電気工業株式会社 研究開発本部 関西総合研究所

〒540 大阪市中央区城見1−2−27

輪郭抽出により唇を抽出する方法や色情報を用いて唇領域を抽出する方法は、照明条件が異なったり、肌や唇が想定している色と異なっている場合には、唇の抽出を失敗してしまう場合がある。本稿では、このような問題を解決するために、唇の輪郭抽出や色情報による唇領域の抽出等は行なわず、顔の中心線上の点を正確に追跡することにより、唇の代表点及び唇の上下方向の動きを抽出するアルゴリズムを提案し、評価実験により有効性を示す。

A Method for Lip Detection by Extracting Partial Motion of Face

Kenji Okano　Toshihiko Miyazaki　Akihiro Okumura　Akihiro Fujii

Kansai Laboratory,Research & Development Group,
Oki Electric Industry Co., Ltd.

1−2−27 Shiromi, Chuo-ku, Osaka, 540 Japan

Methods for lip extraction using models of its contour shape or color information are apt to fail due to its sensitivity for light condition. In this paper, we propose a new method for lip extraction which can apply more various situations without using any contour or color models. This method consists of extracting partial motion of face by tracking points which are laid vertically onto the center of the face. In the latter of the paper, we demonstrate the effectiveness of this method through some experimental results.

1 はじめに

唇の動き情報を用いることで音声認識の性能向上が期待できる。例えば、唇が発話中に何回閉じているなどの情報が確実に分かれば、発話された単語を予め絞り込むことが可能になる。

唇を抽出する方法には色情報を用いる手法[1]や輪郭を抽出する手法[2]等が考案されている。しかしながら、いずれの方法も肌や唇の色の変化、照明条件の変化などの影響を受けやすく、実際に使用する場合には、専用の照明を設置したり、処理パラメータの変更が必要となる。

本研究は上唇、下唇それぞれの上下方向の動き情報を正確に抽出する、照明条件にロバストな方式を開発し、機械読唇や音声認識の性能向上に応用することを目的としている。

提案する手法は、肌や唇の色の変化や照明条件の変化の影響を受けにくくするために、色情報を用いた唇領域の抽出や、唇の輪郭抽出等は行わず、顔の中心線上の点を正確に追跡することにより、顔の動画像からダイレクトに唇の上下方向の動きを抽出する。

2 唇抽出アルゴリズムの概要

本アルゴリズムでは顔の中心線上に等間隔に配置した点を顔画像が入力されるたびに正確に追跡し、最も距離変化の大きい区間の端点を上唇、下唇の代表点とする（図1）。又、抽出された代表点の追跡結果を上唇及び下唇の動きと解釈する。つまり、アルゴリズムでは唇の位置の決定と動き情報の抽出が同時に行なわれる。

中心線上の点の追跡はテンプレートマッチングにより行った。ただし、化粧などをして肌の模様に特徴がない場合や、追跡している点が消滅したりする場合（歯や舌）、テンプレートマッチングで点の追跡を行うと、実際とは全く違う点を処理結果として返すことがある。このようなテンプレートマッチングの処理ミスによる悪影響を軽減するために、正誤判定処理を設けた。

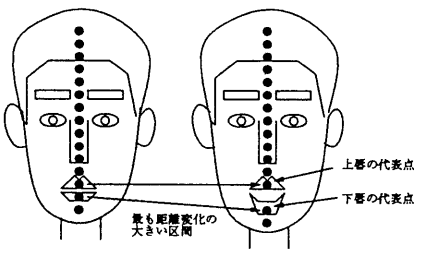

図 1: アルゴリズムの概要

図2に処理の概略フローを示す。入力画像の第1フレームが入力されると、まず顔の中心線を求め、中心線上に等間隔に追跡点を設定する。設定された各点に対して、テンプレートマッチングのためのテンプレートをそれぞれ作成する。第2フレーム以降の画像が入力されると、テンプレートマッチング処理により、それぞれの点がどの位置に移動したかを追跡する。次に、正誤判定処理により追跡結果が正しいかを判定する。追跡結果が誤りと判定された点はこの時点で追跡対象からはずされ、正しいと判定されれば、引き続き以降のフレームに対して追跡処理を繰り返す。ただし、結果が正しいと判定された点の中でマッチング時の距離が閾値を越えているものは、現在のフレーム中の追跡点上の画像をテンプレートとして追加する。上記の処理を最終フレームまで繰り返す。最終フレームの処理が終了したら、唇の代表点を決定する。唇の代表点の決定処理は、「発話中は口が開閉するため、顔の中心線上で上唇と下唇の間の区間が最も伸び縮みする」という仮定に基づいて行う。すなわち、最終フレームまで残った点で隣あうもの同士の距離変化が最も大きい2点を上唇、下唇 の動きを代表する点とする。ここで決定した代表点の追跡結果が上唇、下唇それぞれの動きとなる。

以下、各処理について詳しく説明する。

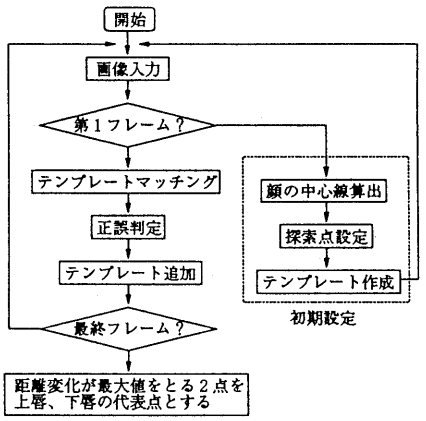

図 2: 処理の概略フロー

2.1 処理する画像について

処理画像はフレームレート 30f/s, サイズ 320 × 240 ドット、24bit/pixel のフルカラーの動画像を対象とした。以降の説明では、画像の幅を W、高さを H、フレーム番号を F とし、入力画像の座標系の原点は左上、座標 (x,y) に於ける入力画像の赤、緑、青のそれぞれのピクセル値を R(x,y)、G(x,y)、B(x,y) とする。（ただし、$0 \leq x < W$、$0 \leq y < H$）

2.2 初期設定

初期設定処理は、第1フレームに対してのみ処理を行う。まず、顔の左右対称性を利用して顔の中心線を求め

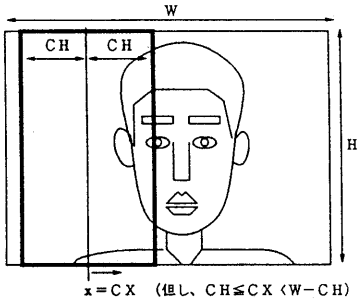

図 3: 中心線算出処理の説明図

る。図3に示すように、x=CXという直線を仮定し、直線との距離がCH以内で、仮定した直線に対して線対称な点同士の画素値の距離の総和 cdist(x) を以下の式により求める。

$$dr(x,y,i) = (R(x-i,y) - R(x+i,y))$$
$$dg(x,y,i) = (G(x-i,y) - G(x+i,y))$$
$$db(x,y,i) = (B(x-i,y) - B(x+i,y))$$

$$cdist(x) = \sum_{y=0}^{H-1} \sum_{i=1}^{CH} \sqrt{dr(x,y,i)^2 + dg(x,y,i)^2 + db(x,y,i)^2}$$

このcdist(x)を CH≦x＜W−CH を満たす全てのxについて求め、cdist(x)が最小値となるxを顔の中心線とする。ここで求めた中心線の値を CX とする。

次に、求めた中心線上に図4に示すように、等間隔（DH）に追跡する点を設定する。

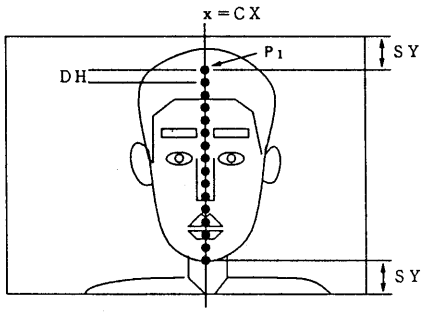

図 4: 追跡点設定の説明図

ここで設定した点を追跡点 P_i とする（i=1,2,…）。但し、Y座標の最も小さい点をi=1とする。

以降、追跡点 P_i の第Fフレームの処理結果の座標は $(rx(i,F), ry(i,F))$ で表現する。よって初期位置の座標は以下の式で表される。

$$rx(i,1) = CX$$
$$ry(i,1) = SY + (i-1) \times DH$$

（ここで SY はY方向の探索範囲である。）

次に設定した各点に対してテンプレートマッチング用のテンプレートを作成する。図5に示すように、座標(x,y)に関するテンプレートを作成する場合、(x,y)を中心として幅2TX、高さ2TYの矩形領域を入力画像より切り出す。

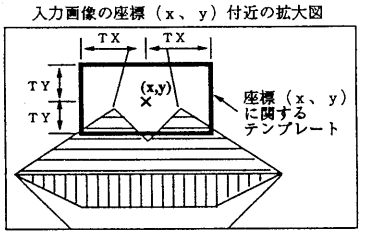

図 5: テンプレート作成の説明図

追跡点 P_i について作成したテンプレートを $T(i,tn)$ で表すとすると（tnはテンプレートの番号）、

$$T(i,1).R(m,n) = R(rx(i,1)+m, ry(i,1)+n)$$
$$T(i,1).G(m,n) = G(rx(i,1)+m, ry(i,1)+n)$$
$$T(i,1).B(m,n) = B(rx(i,1)+m, ry(i,1)+n)$$

となる。（但し、$-TX \leqq m \leqq TX$、$-TY \leqq n \leqq TY$）

2.3 テンプレートマッチング

第2フレーム以降のフレームが入力されると、各点が直前のフレームで位置していた場所からどこに移動したかをテンプレートマッチングによって探索する。探索範囲はパラメータをSX、SYとすると図6のようになる。

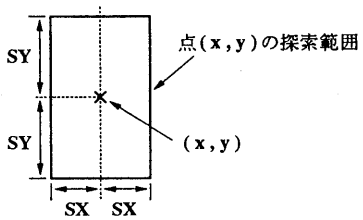

図 6: テンプレートマッチングの探索範囲

第Fフレームにおける、追跡点 P_i のテンプレートマッチング処理は、x,yを以下の範囲で変化させて、

$$(rx(i,F-1) - SX) \leqq x \leqq (rx(i,F-1) + SX)$$
$$(ry(i,F-1) - SY) \leqq y \leqq (ry(i,F-1) + SY)$$

各テンプレート(tn)と座標(x,y)の周辺領域との距離計算を行う。

$$tdist(x,y,tn) = \sum_{n=-TY}^{TY} \sum_{m=-TX}^{TX} \sqrt{TR^2 + TG^2 + TB^2}$$

$$TR = (T(i,tn).R(m,n) - R(x+m,y+n))$$
$$TG = (T(i,tn).G(m,n) - G(x+m,y+n))$$
$$TB = (T(i,tn).B(m,n) - B(x+m,y+n))$$

計算した距離 $tdist(x,y,tn)$ が最小になる座標 (x,y) を追跡点 P_i の追跡結果 $(rx(i,F),ry(i,F))$ とする。

2.4 正誤判定処理

正誤判定処理ではテンプレートマッチングの追跡ミスによる悪影響を軽減するために、連続する追跡点同士の位置情報により、追跡結果が正しいかを判定する。

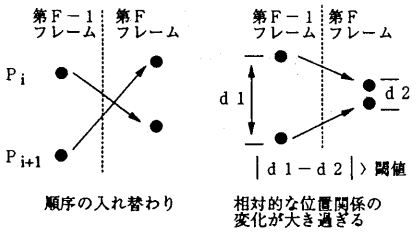

図 7: テンプレートマッチングの処理ミス

図7に示すように、テンプレートマッチングで追跡を誤ると、連続する2つの追跡点の位置関係が直前のフレーム (F−1) と現在のフレーム (F) で入れ替わってしまったり、2点間の距離が極端に変化してしまう場合がある。本処理では連続する3点の追跡点を用いて上記のミスを検出することにより、唇の抽出性能の向上を図る。

連続する3つの追跡点を用いた正誤判定の条件を以下の様に設定した（図8）。

- 追跡点 P_i,P_{i+1},P_{i+2} の3点は直前のフレームの正誤判定で全て正しいと判定された（3点は生き残りの点である）
- 追跡点 P_i,P_{i+1},P_{i+2} の3点の相互の位置関係（Y方向の順序）は変わらない
- 追跡点 P_i,P_{i+1} の2点間の距離の変化はある閾値以下である
- 追跡点 P_{i+1},P_{i+2} の2点間の距離の変化はある閾値以下である

以上、4つの条件を全て満たした場合には、追跡点 P_i, P_{i+1}, P_{i+2} の3点の追跡結果は正しいと判定する。この判定処理を全ての追跡点について行い、一度も正しいと判定されなかった点は結果が誤りであると判定する。

本処理で結果が誤りと判定された点は、正しい追跡結果が得られないため、以降のフレームが入力されてもテンプレートマッチング等の処理は行わない。

2.5 テンプレート追加処理

発話している時の唇の形を観測すると、口の開き具合や、発話内容によって唇の形は変化する。よって、最初のフレームで作成したテンプレートのみでテンプレートマッチング処理を続けると、形の変化が許容範囲を越えてしまった場合に追跡を誤ることがある。一方、1/30秒で唇の形を観測すると、フレーム間では形は極端には変化せず徐々に変化する。このことから、唇の形の変化がテンプレートマッチングの許容範囲を越える前に、新たなテンプレートを追加することにより、追跡誤りの軽減が期待できる。実際の処理では、テンプレートマッチングで計算した距離 $tdist(x,y,tn)$ の最小値が設定した閾値を越えた場合に、現在のフレームを用いてテンプレートの追加を行うようにした。

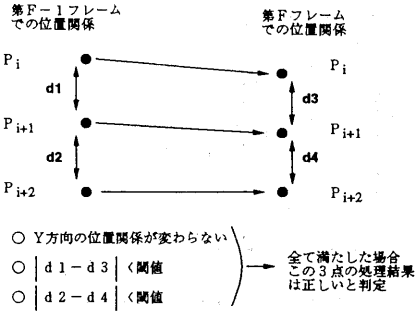

図 8: 正誤判定の説明図

2.6 唇の代表点の決定

最終フレームの処理が全て終了したら、最終フレームまで生き残った点（全てのフレームにおける正誤判定処理で正しいと判定された点）のみを対象として、隣あう追跡点間の距離の変化が最も大きい区間を見つける。

図9に示すように、追跡点 P_{n1} が生き残った点だとすると、次に生き残っている追跡点 P_{n2} を見つける。

（図9の例では $n2 = n1 + 3$ である。）

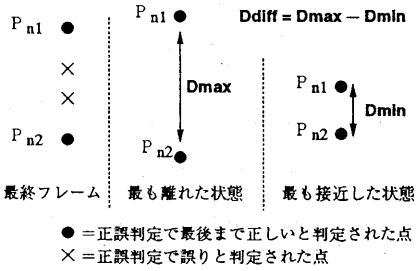

図 9: 唇の代表点決定の説明図

次に過去のフレームの追跡結果をチェックし、追跡点 P_{n1} ,P_{n2} の2点間の距離の最大値 Dmax、最小値 Dmin を求め 、その差分 Ddiff を求める。この Ddiff の値を生き残った全ての追跡点について求め、Ddiff が最も大きい追跡点 P_{n1},P_{n2} を処理結果とする。

つまり、追跡点 P_{n1} が上唇の代表点であり、その点の動き（追跡結果）が上唇の動きに相当する。同様に、追跡点 P_{n2} が下唇の代表点で、その点の動きが下唇の動きに相当する。

3 評価

3.1 評価1：唇の代表点の決定性能

予め決められた単語を発話してもらい、ビデオカメラで撮影を行った。画像データ（動画）は音声の開始点から終了点までを1つの単位として切り出しA/D変換を行った。データ数は4人が発話した合計821の単語である。

処理結果の正誤は図10に示すように、上唇の代表点 P_{n1} は上唇及び上唇の上方5画素以内に存在すれば正解とした。同様に下唇の代表点 P_{n2} は下唇及び下唇の下方5画素以内に存在した場合に正解とした。

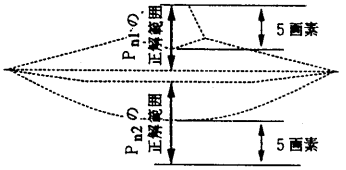

図 10: 唇の代表点の正解範囲

表1に評価時のパラメータ、表2にそのパラメータを用いた場合の評価結果を示す。表2には正誤判定処理を行った場合と、正誤判定を行わずに処理した場合（テンプレートマッチングの処理結果は常に正しいと仮定した場合）の2つの結果を示している。

パラメータ	値
画像 (24bits/pixel)	320 × 240 画素
フレームレート	30 frames/sec
中心線算出 (CH)	100 画素
初期座標間の幅 (DH)	3 画素
テンプレートサイズ (TX)	16 画素
テンプレートサイズ (TY)	8 画素
マッチング範囲 (SX)	1 画素
マッチング範囲 (SY)	15 画素
正誤判定の閾値	3 画素
テンプレート作成の閾値	40 (1 画素当り)

表 1: 評価時のパラメータ

	代表点の抽出率	
正誤判定あり	93.8 %	(770/821)
正誤判定なし	88.7 %	(728/821)
正誤判定の効果	+ 5.1 %	(+ 42)

表 2: 評価1の結果

正誤判定処理がある場合とない場合の抽出率を比較してみると、正誤判定処理を行った方が5％程度、抽出率が向上していることが分かる。

正誤判定処理ありの場合で、唇の代表点がうまく抽出できなかった51個は

- 5画素以内という条件を満たさなかったもの（上唇及び下唇の位置検出はうまくいっている）→ 12個
- 唇以外の場所を選択してしまったもの（顎や鼻の上など）→ 39個

の2つに分類できる。

唇以外の場所を選択してしまった処理結果を見てみると、顎の皮膚の伸縮に反応してしまい最終的に顎を選択してしまったものや正誤判定処理で口より下の点が全て誤りとなってしまい、鼻より上の点しか残らなかったものなどがあった。

3.2 評価2：動きの抽出性能

評価1に用いたデータの中で唇の代表点の抽出に成功した動画データの中から28個を選び、それぞれのデータの各フレーム毎の唇の輪郭と顔の中心線との交点（上下2点）の座標データを作成した（自動抽出して、手作業で修正）。これらの座標データを上唇及び下唇の動きの正解データとした。

これら28個のデータに対して評価1と同様の抽出処理（正誤判定あり）を行い、抽出結果の座標データが正解データとどの程度同じような動きをするかを評価した。この評価には以下の式を用いた。第 F フレームに於ける上唇の正解データのY座標を $ay(F)$、最終フレームの番号を $Fmax$ とすると

$$av = \frac{\sum_{F=1}^{Fmax}(ry(n1,F)-ay(F))}{Fmax}$$

$$dy = \sqrt{\frac{\sum_{F=1}^{Fmax}(ry(n1,F)-av)^2}{Fmax}}$$

上記の式で求めた dy の値が小さいほど、正解データと同じ動きをした（等距離を保って移動した）ことになる。

同様の方法で下唇の評価値 dy も求めた。

	dy	av
上唇	1.17	+4.96
下唇	1.25	−11.11

表 3: 評価2の結果

表3に評価2の結果を示す。表3の値は、28個のデータそれぞれについて求めた dy 及び av の値を平均した値である。

この結果より、代表点の追跡結果は正解データに対して平均して1画素程度、近付いたり離れたりしていることが分かる。

ここで、位置関係（正解データとの差分の平均値）avに着目してみると、上唇の抽出結果は+4.96つまり、5画素程度唇の輪郭よりも内側（下方向）の点を選択していることが分かる。一方、下唇のavは−11.11つまり、11画素程度輪郭よりも内側（上方向）の点を選択している。入力画像中の上唇のY方向の幅が10画素前後、下唇のY方向の幅が15画素前後であることから、テンプレートのY方向の幅TYを8として処理しているのでテンプレートマッチング時にテンプレートの一部が唇よりも内側の領域（歯や舌）にかかっていると考えられる。これにより、抽出された唇の代表点P_{n1},P_{n2}の追跡結果は、正解データに対して1画素程度近付いたり離れたりしていると考えられる。

この問題を解決するために、上唇の代表点としてP_{n1}が選択された場合、P_{n1}をそのまま採用せず、P_{n1}よりも上に位置する追跡点を利用することが考えられる。一方下唇に関しては、代表点としてP_{n2}よりも下に位置する追跡点を採用する。

この考えに基づき再度処理を行ってみた。表4に評価結果を示す。結果を見てみると、上唇及び下唇共に選択された代表点P_{n1},P_{n2}よりも2～3だけ外側の追跡点の動きが唇の輪郭の動きに最も近い動きをすることが分かる。

代表点	上唇 dy	代表点	下唇 dy
P_{n1}	1.17	P_{n2}	1.25
P_{n1-1}	0.73	P_{n2+1}	1.12
P_{n1-2}	0.66	P_{n2+2}	1.02
P_{n1-3}	0.66	P_{n2+3}	0.96
P_{n1-4}	0.71	P_{n2+4}	1.06

表 4: 外側の点を採用した場合

図11の人が「ハネムーン」と発話したデータの処理結果を図12に示す（図11の画像はそのデータの第1フレームである）。但し、パラメータは評価1と同じ値を用い、最終的にP_{n1-3},P_{n2+3}を代表点として採用した。図12には正解データと追跡結果を重ねて表示しているが、抽出結果はほぼ唇の輪郭と同じ動きをしていることが分かる。最も距離の離れている点で1.6画素である。

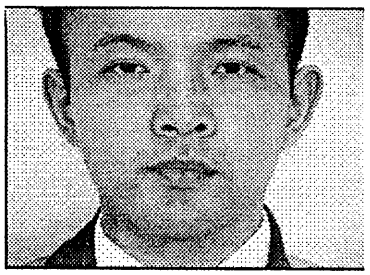

図 11: 処理画像

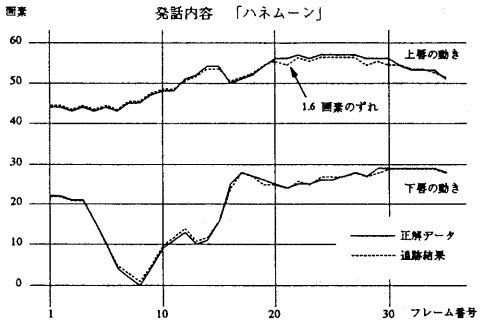

図 12: 追跡結果の一例

4 今後の課題

我々は、今回抽出した唇の上下方向の動き情報を用いて音声認識の性能を向上させる研究や、唇の動き情報により単語を認識する研究を行っている。

音声認識の性能を向上させる手段の一つとして両唇音(/b/,/m/,/p/)の検出が考えられる。両唇音を確実に検出できれば、音声認識時に候補単語を絞り込むことが可能になるので、音声認識の性能向上が期待できる。実際に唇の上下方向の動き情報を用いて両唇音の検出実験を行なった結果、約98%の検出率が得られた。今後は両唇音以外の情報を検出するアルゴリズムも開発する予定である。

又、唇の上下方向の動き情報のみを用いて単語認識実験を行った。5音節からなる30単語を用いた特定話者の認識実験では、約90%の認識率が得られた。さらに、唇の上下方向の動きに加えて口の水平方向の端点の動き情報（手入力により作成）を用いることにより5%程度認識率が向上することが分かった。よって、今後は口の水平方向の端点の動きを自動抽出するアルゴリズムの開発が課題となる。又、現状では入力画像は単語単位に既に切り出されたものとして処理を行っているので、今後はワードスポッティング等の技術開発が必要である。

5 まとめ

本研究では、テンプレートマッチング及び正誤判定処理により顔の中心線上の点を正確に追跡することにより唇の代表点を決定し、合わせて唇の動き情報も抽出する手法を提案した。また、実験により本手法の有効性を確認した。

参考文献

[1] 黒田, 渡辺 : “色彩情報処理による顔画像の唇抽出法”, Human Interface,Vol.10 , pp. 13-18 (1995).

[2] 田村, 梶見, 岡崎, 光本, 河合, 副井 : “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”, PRU89-20, pp. 9-16 (1989).