動的輪郭モデルによる唇形状抽出と単語認識のハードウェア実現について

Hardware Realization of Lip Shape Extractions by Using Active Contour Model and its Application for Word Recognition

佐々木　悠介 †　　川村　尚生 ††　　菅原　一孔 ††

Yusuke Sasaki † Takao Kawamura †† Kazunori Sugahara ††

† 鳥取大学大学院　工学研究科　知能情報工学専攻　　†† 鳥取大学　工学部　知能情報工学科

1　はじめに

工場内など高雑音環境下でも、製造装置や搬送用ロボットなどを音声により制御したい場面がある。このような場面では、大語彙の音声認識は必要なくある程度の語数の単語を識別できれば十分な場合も多い。しかし、現在各所で研究が進んでいる音声認識手法では、周囲の雑音が少ない場合には有効であるが、高い雑音環境では認識率が極端に低下してしまう問題がある。この点、人の発話時の唇形状を認識する、いわゆる読唇手法では周囲の雑音の影響はまったくなく、また手話などを使う方法に比べると、人が特別な訓練をする必要がないなどの有利な点があり有効な手法のひとつと考える。

読唇システムを構築するためには、入力動画像から高速に唇形状を抽出するための手法が不可欠である。画像中の色のヒストグラムを集計し、それにより唇領域を抽出するなどの従来手法では、画像を蓄えているメモリへのアクセス数は、画像中の画素数に相当しその数は膨大である。メモリへのアクセスには、最低でも 1 クロック必要であり、結果としてアクセス数の多い従来手法の動作速度は、低下してしまっていた。この点、動的輪郭モデルによる手法では、メモリへのアクセスは少なくてすみ、高速な唇形状抽出処理が期待できる。

本稿では読唇によるシステムの制御を目指し、動的輪郭モデルにより唇の形状を抽出し、その形状情報を用いた母音認識の手法を提案する。また、本手法の様々なシステムへの組み込みを目指し、FPGA 上にハードウェアとして実現した。それによる母音認識結果を示し、本手法の有効性を確認する。

2　動的輪郭モデル

Kass らは画像中のエネルギーの最小化により、画像中の特定領域を抽出する動的輪郭モデル (Snakes) を提案し、その後様々な画像エネルギーを考慮した手法が提案された [1]。しかしこれらの手法では、エネルギーの最小化に多大な計算処理が必要であり、本稿で扱う組み込みシステムへの応用は困難である [2]。その後、エネルギーの最小化を、動的輪郭モデルに加わる力の釣り合いとして捕らえる Sampled-ACM が橋本らにより提案され [3]、さらに菅原らにより振動項を考慮することにより耐雑音性能を向上させたものが提案された [4]。本稿では、振動項を考慮した Sampled-ACM を用いて唇形状を抽出することにする。

振動項を考慮した Sampled-ACM(以下、動的輪郭モデル) では仮想的な閉曲線上の複数の動作点に、圧力、引力、反力および振動項と呼ばれる 4 つの力が動作点に働くことにより、閉曲線が収縮し領域を抽出するが、本稿で抽出対象である唇の形状がおおむね全領域にわたり外側に凸であるという特徴を考慮し、圧力は考慮せず引力と反力および振動項により収縮動作をするものとした。

引力は図 1 に示すように隣り合う 2 つの動作点間に働く力であり、その間の距離に比例した大きさを持つものとした。振動項は引力の合力 F_a に対し直角方向に働く力であり、収縮のたびにその方向を反転する。なお、この振動項の大きさは一定の値 F_v を持つものとした。

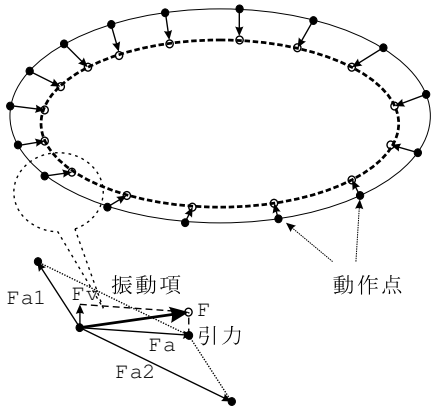


図 1: 引力と振動項による収縮動作

図 2 に示すように、反力は動作点が対象の画像領域に接した際に働く力であり、引力 F_a と振動項 F_v の合力の抽出領域に対する垂直成分を打ち消す働きをもつ。ただし、反力はその大きさに閾値を持ち、ある一定以上の垂直成分は打ち消すことができないものとする。これらの力の働きにより、画像中の雑音をすり抜けたり、あるいは突き抜けたりする動作を実現することが可能となり、画像中の雑音に強い領域抽出手法となる。

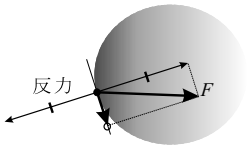


図 2: 反力の働き

3　唇形状の抽出

動的輪郭モデルを用いて唇の形状抽出を行う。本稿ではより正確な唇形状の抽出をするため、唇の外側と内側の形状を抽出する。

本稿で扱う入力画像は画像サイズ 640 × 480[pixel] のカラー画像を想定している。この入力画像を HLS 表色系により 2 値化し、画像中の黒色部分を抽出領域として動的輪郭モデルを適用すると、図 3 に示すとおり唇の外側形状を抽出することができる。なおこの場合、動的輪郭モデルの初期輪郭としては画像全体を囲むように配置した。動的輪郭モデルに基づく手法では画像が記録されているメモリへのアクセスは、動作点が移動する画素のみですみ、色による判別手法などの他の手法に比べ極端にメモリのアクセス数を低減させることが可能である。

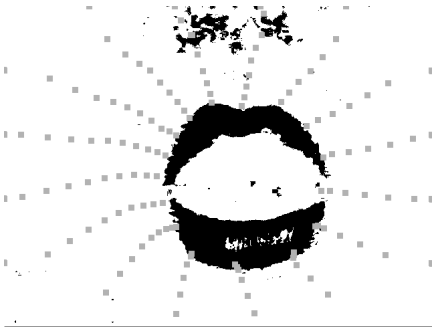


図 3: 唇の外側形状を抽出している様子

続いて、唇の内側の形状を抽出するために、動的輪郭モデルが抽出領域とする画像中の色を白黒反転し、再度適用することを考える。ただし、2 回目の動的輪郭モデルの動作点の初期位置は、1 回目の収束結果を用いることにする。唇の内側の形状を抽出している様子を図 4 に示す。

このようにして得られた母音発話時の唇の外側および内側形状の抽出例を図 5 に示す。図 5 では良好な唇形状が抽出できているが、中には図 6 のように、唇の外側形状の抽出の際に画像の明るさや陰の影響により、鼻やあごの部分を唇の一部としてとらえてしまうこともある。このような抽出結果をもとに読唇を行うと、その識別率の低下を招くことが予想される。

しかし、鼻やあごの部分を唇の一部としてとらえてしまった場合でも、顔の頬の部分については、それを唇ととらえてしまう動作例はほとんど無いことから、

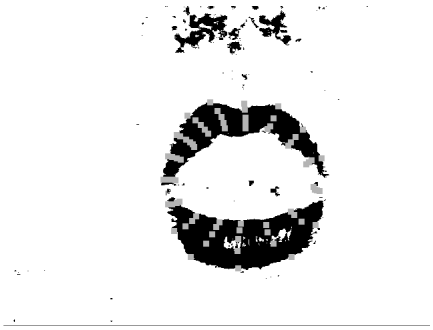


図 4: 唇の内側形状を抽出している様子

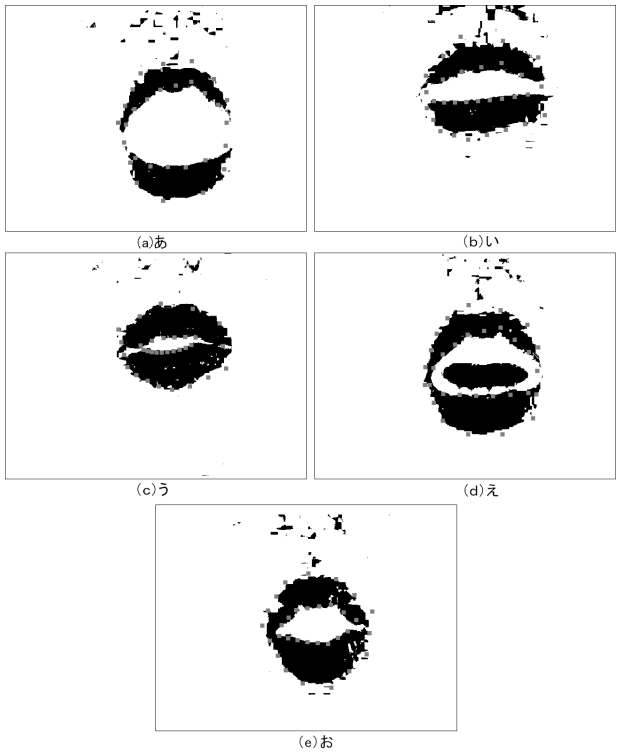


図 5: 母音発話時の唇の外側および内側形状の抽出例

唇の幅はある程度正確に求められていることが多い。

次に、唇の内側形状を抽出する場合、鼻やあごの部分に引っかかり正しく唇の外側形状を抽出できていなかった場合でも、あごと唇の領域は連続している場合が多いことなどから、唇内側下半分の形状は比較的正確に抽出できている。これらの点から本稿における読唇は、唇下半分形状を用いて行うこととする。

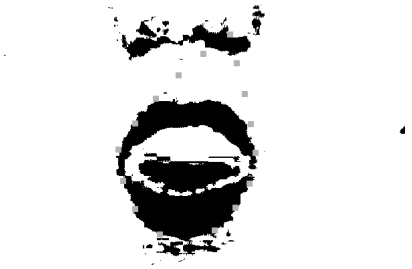


図 6: 唇の外側形状の抽出の失敗例

4 母音認識の手法

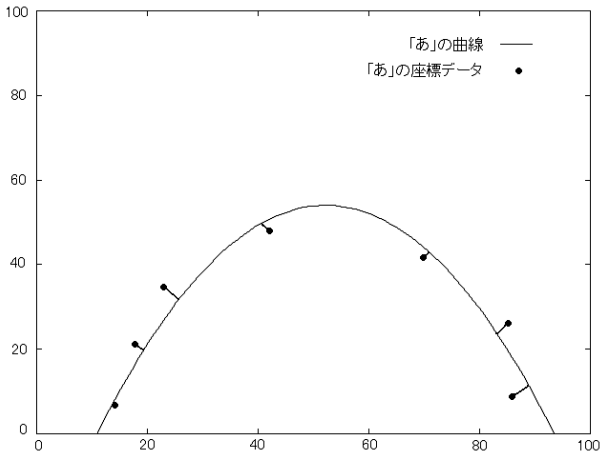
動的輪郭モデルの収縮により得られた動作点の座標を基に、母音を認識する手法を提案する。本手法では、各母音ごとにその唇形状を表現する二次曲線を用意し、母音が未知の唇画像に対して動的輪郭モデルを適用した時の動作点との距離を計算することで母音を認識する。

二次曲線の作成手順としては、はじめに母音を発話している画像に動的輪郭モデルを適用し、収縮した動作点の座標データを取得する。続いて、唇が画像中のどの位置にあっても認識ができるよう、座標データを唇の幅より導出した中心について正規化する。このとき、曲線の基となる座標データは、3 章で述べたように比較的正確に形状を抽出できている唇下半分形状のものを用いることにする。最後に正規化したデータをそれぞれの母音ごとに平均し、二次曲線に近似する。曲線を用いることで、認識対象の唇形状距離計算の手間を軽減できると考える。

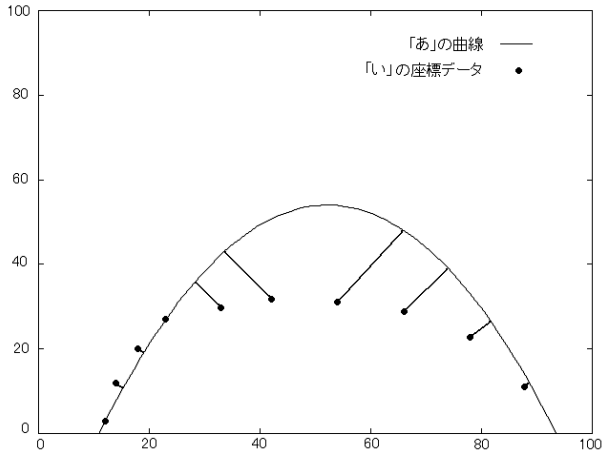
作成した曲線と認識対象の座標データとの距離を比較している様子を図 7 に示す。認識対象の座標データとの垂直距離の合計が最も小さい曲線が示す母音を、その画像の唇形状が示す母音とする。

5 唇形状抽出システムのハードウェア実現

3 章で述べた唇形状抽出手法を、様々なシステムへ組み込むためにそのハードウェア化を試みた。ハードウェア化はハードウェア記述言語、VHDL を用いて FPGA 上に行うこととし、画像入出力回路、FPGA と画像メモリを搭載した装置を開発した。画像の入出力は NTSC 規格に準拠した信号を取り扱うこととし、また FPGA には ALTERA 社の APEX20KC を用いた。この FPGA のロジックエレメント数は 8,320 であり、



(a)「あ」の曲線と「あ」の座標データとの距離



(b)「あ」の曲線と「い」の座標データとの距離

図 7: 唇形状を表す曲線と座標データの距離比較例

20 万ゲートに相当する規模の FPGA である。システム全体の構成を図 8 に示す。

動作の流れとしては、システムの入力として発話者の映像を NTSC カメラより取り込み、NTSC デコーダを介して RGB 値で FPGA ボードのメモリへと格納する。1 フレーム分の画像が格納され次第、メモリに格納された画像に対し動的輪郭モデルを適用する。その結果唇形状を抽出した動作点の座標データが得られる。このデータを母音認識回路に渡し、認識結果を出力する。母音の認識は、ある母音が一定の期間続けて認識された時に完了するとした。

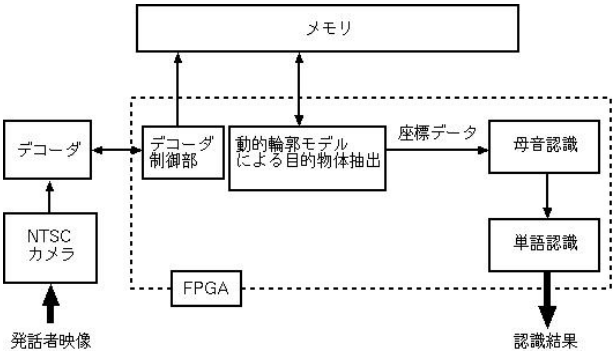


図 8: システム構成

4 章で述べた読唇手法ならびに画像入出力回路の制御回路を実装した結果、65%のハードウェア量でシステムを実現できた。動的輪郭モデル部の 1 フレームの処理に必要なクロック数は約 267,000 であり、FPGA のクロックが 48MHz である条件下で毎秒 30 フレームの動画像を処理するに十分な処理速度を実現できた。またメモリアクセス数については、640 × 480[pixel] の入力画像に対し、外側形状抽出の最大収束回数を 50 回、内側形状抽出の最大収束回数を 30 回としたとき、最高でも約 15,000 回となり、画像中のヒストグラムを集計する手法のメモリアクセス数 307,200 回に比べ大幅に低減できた。

6 母音認識実験

4 章で述べた手順にしたがって母音認識の実験を行った。なお、本稿で用いた唇画像はすべて同一人物のものである。実験に用いた母音を表す曲線は、それぞれ 10 枚の画像に動的輪郭モデルを適用し、結果を平均したものを使用して作成した。この認識システムにそれぞれの母音 10 枚づつ、計 50 枚の画像を与えた時の認識結果を表 1 と表 2 に示す。表より本手法による母音認識が全体としてはある程度の認識率を有していることが分かる。個々で見ると一部の母音認識が単一の母音への誤認識に集中しているため、その差異をつける特徴を抽出する必要がある。

表 1: 母音認識結果

母音	認識数	誤認識数	認識率 (%)
あ	10	0	100.0
い	10	0	100.0
う	9	1	90.0
え	7	3	70.0
お	6	4	60.0
全体	42	8	84.0

表 2: 母音認識結果内訳

	認識された母音				
実際の母音	あ	い	う	え	お
あ	10	0	0	0	0
い	0	10	0	0	0
う	0	0	9	0	1
え	3	0	0	7	0
お	0	4	0	0	6

7 おわりに

本稿では、発話時の顔画像に動的輪郭モデルを適用し、取得した形状を基に母音認識を行う手法を提案した。提案した手法は FPGA 上にハードウェア実現し、唇形状の抽出実験と母音認識実験を行い、その有効性を確認した。今後の課題は、母音認識の時間的な組み合わせによって単語認識を実現することである。

8 参考文献

[1] M. Kass, A. Witkin and D. Terzopoulos, “Snakes: Active Contour Models,” International Journal of Computer Vision, pp.321-331, 1998.

[2] 須賀 弘道, 羽鳥 好律, 植松 明, “SNAKE を用いた顔画像からの構成部品の輪郭抽出,” 電子情報通信学会論文誌 A Vol.J79-A, No2, pp.298-301, 1996.

[3] 橋本 昌寛, 木下 宏揚, 酒井 善則, “Sampled Active Contour Model による輪郭抽出法,” 電子情報通信学会論文誌 D-Ⅱ, Vol.J77-D-Ⅱ, No.11, pp.2171-2178, 1994.

[4] 菅原 一孔, 新地 俊幹, 小西 亮介, “振動項を持つ動的輪郭モデル,” 電子情報通信学会論文誌 D-Ⅱ Vol.J80-D-Ⅱ, No.12, pp.3232-3235, 1997.

[5] T. Shinchi, Kazunori Sugahara and Ryosuke Konishi, “Vowel Recognition According to Lip Shapes by Using Neural Network,” Proceedings of 1998 IEEE International Joint Conerence on Neural Networks, pp.1772-1777 (1998). Anchorage Alaska