K-055

オプティカルフローの方向特徴履歴とサポートベクトルマシンを用いた読唇の基礎的検討

Fundamental Study of Lip-Reading Using Record of Optical Flow Directions and Support Vector Machine

高橋　昌平†　　大谷　淳†
Shohei Takahashi　Jun Ohya

1．まえがき

近年、音声認識の研究は目覚ましい発展をしており，様々なアプリケーションに使用されている.しかし，音声認識アプリケーションの認識率はノイズに大きく影響されるため，オフィスや工場，車内などの騒音の激しい場所では使用が難しいのが現状である.

また，難聴者の方や,失声症,ストレスや病気などが原因で声を出せない，出すのが難しい人々は音声認識の恩恵を利用するのが難しい.

人間は会話内容を理解するとき，音声だけではなく，様々な情報を用いて情報を整理し利用している.発話者の唇の動きを用い，読唇を行うことによって騒音環境下，または声を出していなくても何を言っているか理解することが可能である.

人間と同様に，コンピュータが人間の会話の内容を理解するためには，音声だけではなく、唇の動きをとらえることが非常に重要である.

そこで，本稿では，人間が会話をしている画像から唇の動きを読み取り，会話の内容を認識する方法を検討する.

2．研究手法

画像を用いた読唇法では特徴量が重要である.画像読唇システムの手法は，唇の動き情報を用いたオプティカルフローを用いた手法[1]，口の形情報を用いたモデルベース手法[2]，固有空間法などを用いたイメージベースの手法[3]に大きく分けられる.本稿では、唇周辺の特徴点の動きの情報に基づくオプティカルフローを用いた手法を用いる.

本手法ではまず，Active Shape Model(ASM)という手法を用いることによって，動画像から顔および唇の検出と追跡を行う.ASM は，トレーニングセットの幾何学的情報と，テクスチャーを利用した，テンプレートベースの柔軟な特徴検出の方法である.

ASM によって検出された唇には，ASM の特徴点があり，唇周辺の特徴点のオプティカルフローを時系列に並べたものを，機械学習の特徴量として SVM(Support Vector Machine)に学習させ，認識を行う.動画像中の顔画像，唇の大きさによってオプティカルフローの大きさは変化するため，機械学習させる前に，画像の大きさによって特徴の大きさが変化しないように正規化をしておく.

3. Active Shape Model

ASM は Cootes らによって提案された手法であり[4]，特徴点の集合である Shape を主成分分析することによって、比較的低次元のパラメータによって，物体の様々な形状を表現し、トレーニングした物体を検出する手法である.

今回は、顔の輪郭 15 点，眉 12 点，目 10 点，鼻 12 点，唇１９点の合計 68 の特徴点を用いた.図 1 に ASM の特徴点と検出結果の例を示す.

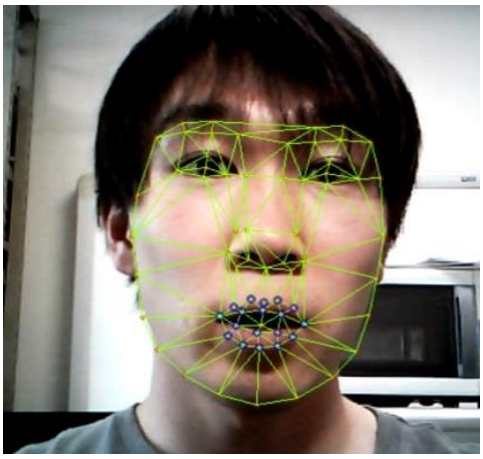

図 1 顔の特徴点と ASM.

ASM の学習画像に与えられた特徴点のベクトルを s とすると，

$$s=(x_1, x_2 \ldots, x_n, y_1, y_2, \ldots, y_n)^T \quad (1)$$

と表すことができる.ここでx_i, y_i(i=1,2...n)は、物体の特徴点の座標であり，n は特徴点の数である.

ここで物体の平均形状を$\bar{S}$とすると，ある物体の形状 s は平均形状からの偏差を主成分分析することによって得られる固有ベクトルE_sを用いることによって

$$s = \bar{S} + E_s b_s \quad (2)$$

と表すことができる.ここでb_sは Shape パラメータと呼ばれ，平均からの偏差を表すパラメータとされる.このパラメータを変化させることによって，ASM で表される形状を様々に変化させることができる.ただし，学習に含まれなかった形状に変化させることはできない.

ASM で表された特徴点は，画像空間に以下のように写像することができる.

$$(x_i', y_i')^T = kRot(\theta)(x_i, y_i)^T + (X, Y)^T \quad (3)$$

ここで k はスケーリング係数，Rot(θ)は回転変換、X,Y は平行移動の大きさである.

画像中の物体と、写像された特徴点で表される Shape モデルのコスト関数が最少になるようにパラメータを調整することによってモデルのフィッティングを行う.コスト関数には、ASM のモデルの輝度を用いる．l 個目の画像のモデルの特徴点によって定義される法線から、それぞれの方向に k ピクセル輝度値をサンプリングし，正規化したベクトルg_lとする.全てのトレーニング画像のベクト

† 早稲田大学国際情報通信研究科
Waseda University Graduate School of Global information and telecommunication Studies

ルの平均値$\bar{g}$とその共分散行列S_gを用いて、新しい画像のモデルとのコスト関数は

$$f(g_s) = (g_s - \bar{\bar{g}})S_g{}^{-1}(g_s - \bar{\bar{g}}) \qquad (4)$$

と表される.新しいモデルにフィッティングを行うとき，この関数が最少になるように計算を繰り返し，パラメータを調整する.ASM の特徴点のうち唇周りの特徴点を用い，オプティカルフローを SVM に機械学習させることによって，発話文字の認識を行う.通常，カメラに映る顔の大きさは異なり，また，人によって唇の大きさが異なるため，そのまま特徴とするとオプティカルフローの大きさが大きく異なってくる.そのため,今回は異なる顔の大きさのデータでも認識できるようにオプティカルフローと唇の長さ，幅との比を特徴とした.

4 実験

実験にはランダムに選んだ１～３桁の数字１５種類を用いた.被験者男４名，女２名の計６名にそれぞれの数字を３回発音してもらい，学習データとしてそのうち２回分を使用し，１回分をテストデータとして用いた.SVM は，15 クラスの分類となる.データの顔と唇の大きさは，実験ごとに異なっている.入力データは,発話する直前から発話後唇を閉じるまでとする.撮影された発話者の映像は 30fps で撮影される.

オプティカルフローは，全ての唇周りの点ではなく，唇の左右端の２点，真ん中の上下の２点のオプティカルフローを用いた.また、SVM に関しては、線形カーネルを用いて，パラメータについては、正（正解）のデータと負（正解でない）のデータそれぞれの正しい認識率の合計が最大になる値を用いた(1).同じく ASM を用いて，初期フレームの話者の唇の幅と高さとの現フレームの幅と高さの比，現フレームでの幅と高さの比，を特徴として SVM を学習させ、認識させたときの比較をする(2).

正のデータの認識率の平均は(1)では 97.7% ,(2)では 98.8%となり，わずかに(2)の手法の方が優れていることが分かる.また，負のデータでは平均で(1)79.6%、(2)では 81.2%のデータが負のクラスとして認識された.

実験に用いた、唇の幅、高さとオプティカルフローを計算する４点は同じ点であり，データに相関があると考えられるため，異なる点を使用すればさらに認識率の上がる可能性がある.

また，一つのクラスに関し負のテストデータに比べ正のデータの数が少ないため，正のデータの認識が下がると認識率の和が大きくなる.そのため，正のデータをより高い確率で認識するよう SVM が学習されている.データ数が増えれば，正のデータの認識率が少し下がり，負のデータの負クラスへの認識率が上がると考えられる.

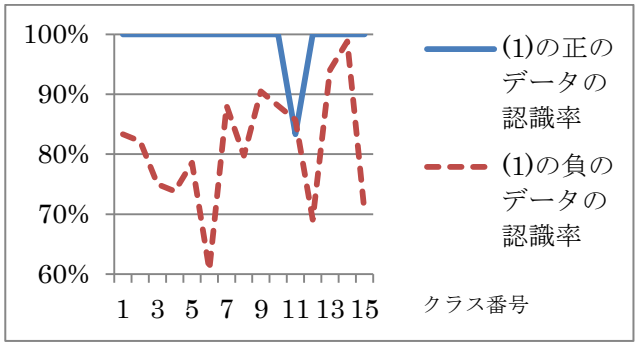

表 1 手法(1)の認識率とクラスの関係

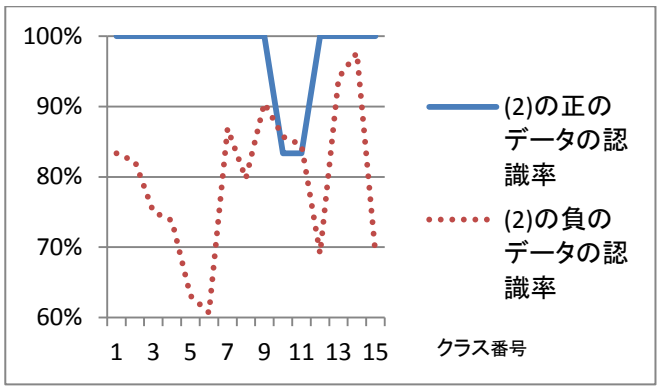

表 2 手法(2)の認識率とクラスの関係

5 まとめ

今回の実験では，ASM を用いることによって，唇の幅と高さ，唇の特徴点のオプティカルフローを特徴点とし，SVM を用いて発音の認識を行った.高い認識率から本手法の有効性が分かる.

また，カメラ中に映った顔，唇の大きさに依存せずに認識することができることが分かった.

今後の課題として，実験データの増加、認識単語数の増加などが考えられる.

参考文献

[1]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 D-II, Vol.J73-D-II, No.6, pp.796-803 (1990)

[2]斉藤 剛史,小西 亮介, “トラジェクトリ特徴量に基づいた単語読唇”,信学論D,Vol. J90-D, NO.4,pp. 1105-1114,(2007)

[3]中田 康之,安藤 護俊,”色抽出法と固有空間法を用いた読唇処理”,電子情報通信学会論文誌、Vol.J85-D-Ⅱ,No12(2002), pp1813-1822

[4] T.F.Cootes, K Walker, C.J.Taylor, “View-based Active Appearance Models", Image and Vision Computing (2002), pp.657-664, 2002.