Vol.2014-CVIM-191 No.7
2014/3/3

複数画像特徴量を用いた読唇システム ― オプティカルフロー特徴・形状特徴・離散コサイン変換特徴の統合の検討 ―

高橋 昌平[†1] 大谷 淳[†1]

あらまし 本論文では，動画像から唇の情報を読み取り画像特徴のみを用いて会話の内容を認識する手法を述べる．画像による会話認識では、ノイズの影響が大きい車の中や，聴覚，視覚障碍者にも有益である．
提案手法では，初めに顔と唇を含んだ動画像に Active Shape Model を適用し顔と唇領域の追跡を行う.追跡された唇から，オプティカルフロー，形状、離散コサイン変換といった唇の特徴を抽出する.抽出された特徴は階層型 SVM の中間層の SVM によって学習認識され，認識結果が最下層の SVM によって統合され最終認識結果となる.複数の画像特徴を用いることによって，認識結果が向上することが実験結果で示された.

Automatic Lip-Reading by using Multiple Visual Features -Integration of the Shape, Optical Flaw and DCT features-

SHOHEI TAKAHASHI[†1] Jun Ohya[†1]

Abstract In the paper, we present a lip-reading method that can recognize speech by using only visual features. Lip-reading can work well in noisy places such as in the car or in the train. In addition people with hearing-impaired or difficulties in hearing can be benefited. First, the Active Shape Model (ASM) is applied to track and detect the face and lip in a video sequence. Second, three visual features, the shape, optical flow and Discreet cosine transformation of the lip are obtained from the lip area detected by ASM. The extracted features are ordered chronologically so that Support Vector Machine (SVM) is performed so as to learn and classify the spoken words. Hierarchical SVMs are used to recognize the words. Each visual feature is trained by the respective middle-layer SVM, and those outputs of SVM's are integrated by the final SVM. Experimental results show that the integration of these features improves the recognition accuracy.

1. はじめに

音信号を用いた会話認識の研究は以前から研究されており，近年ではコンピュータ,携帯電話,コールセンターなど至る場所で使用されているのを実際に見ることができる.しかし,音信号を用いた会話認識は車や電車の中などノイズが大きな場所では利用が難しい.もし,車の中で会話認識システムが使用できるなら,ハンドル操作に集中し手が使えない運転手も何らかのデバイスを使うことができる.また,聴覚障害者や発話障害者など正常に発音をするのが難しい人は音声認識システムを利用することは難しいため、システムの恩恵を得ることができない.視覚情報を用いた会話認識では,正常に発音ができなくても,唇の動き等を用いて会話の認識をすることができる.

人間は,会話を理解するために音声情報だけではなく,唇の動きなどの視覚情報にも頼っている.視覚情報を利用した会話認識システムができれば,騒音の大きな環境でも使用でき,正常な発音ができない人々でも会話認識を利用することができ,音声認識を用いた会話認識システムの補間をすることができる.

本研究では,音声認識では解決することができない上記の課題を解決するために視覚情報のみを用いた会話認識システムの手法を提案する.

視覚情報を用いた会話認識の研究は,自動読唇と言われ様々な手法が研究されている.

Shaikh らは,動画像からの読唇システムの研究を発表している[1].唇の縦方向のオプティカルフローの情報とサポートベクトルマシンを用いることによって発音の分類を行っている. 間瀬らもオプティカルフローを用いた読唇の研究を発表をしている[2].オプティカルフローを主成分分析し,その固有値から代表的な特徴を割出し特徴量とし,あらかじめ登録しておいた発音の特徴量とマッチングさせることによって認識を行っている.

Chiou らは,スネークを用いて唇の領域を抽出した後,主成分分析を用いて特徴を抽出し,隠れマルコフモデルを用いて発音の分類を行っている[3]. 中田康之らも、固有空間法を用いた読唇処理の研究をした[4].色抽出法と固有空間法

†1 早稲田大学国際情報通信研究科
Waseda University Graduate School of Global Information and Telecommunication Studies.

を用いて,唇の位置を検出し固有空間の時間的変化を記述しマッチングを行っている.
斉藤らは,唇の形状を特徴量とした読唇処理の研究を発表した[5].口内面積や唇の幅と高さのアスペクト比を時系列に並べたものを特徴量とし,データベースに登録された単語とマッチングをさせ認識する.
従来の読唇システムの研究では,特徴量を分類すると,オプティカルフローを利用するもの,周波数空間や固有空間などの画像情報を利用するもの,唇の形状特徴を利用するものに大別できる.
しかし、従来の研究手法ではこれらの特徴は独立して利用されている.これらの特徴量を複数用いて会話認識を行うことができれば,視覚情報を用いた会話認識システムの精度が上がると考えられる.また,従来の手法では,カメラの距離,または個人の身体的特徴から生じる画像上に映りこむ唇の大きさを考慮した研究はみあたらない.

2. 提案手法

本研究では,音信号を用いず,複数の視覚特徴を組み合わせた会話認識システムを提案する.ここで,特徴量は画像に映りこむ唇の大きさを考慮する.

提案手法の処理の流れを図 1 に示す.

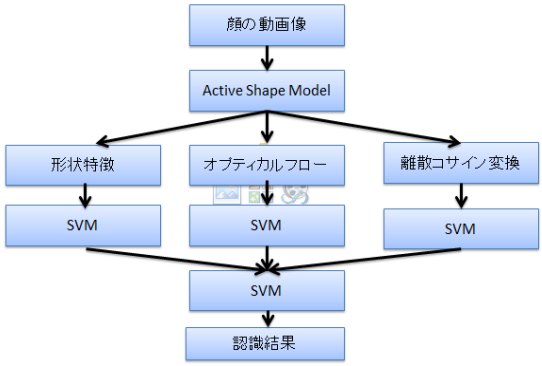

図 1　提案手法の処理の流れ

初めに入力として,ある単語を発音している人の顔の動画像に Active Shape Model(ASM)[6]を適用する.ASM は,複雑な形状を持つ物体やフレームごとに形状が変化する物体でも検出と追跡ができる手法である.様々な形状の物体を用意したトレーニングセットを分析することで,複雑な形状を持つ物体を近似することができるモデルを作成し,物体に合わせることで検出及び,追跡を行う.ASM によって動画像中の顔と唇を検出,追跡する.
検出された唇からは、唇の形状特徴,オプティカルフロー,空間周波数といった特徴量が抽出される.抽出された特徴量は時系列に並べられる.
それぞれの特徴量は,サポートベクトルマシンによって学習,分類される.複数の特徴を統合した認識をさせるため,中間層のサポートベクトルマシンの分類結果を特徴量として、最下層のサポートベクトルマシンが会話の認識を行う.

3. Active Shape Model

Active Shape Model (ASM) は Cootes らによって開発された複雑な形状をもつ物体でも追跡や検出が可能な手法である[6].顔画像及び唇の画像は,人種や個人の特性のみならず,光の方向や強さに大きく影響される.さらに,発音をしているとき顔と唇の形状は大きく変化しており,形状が単純な物体や動きが少ない物体と比較して,追跡及び検出するのは難しい.ASM では,データベースから様々な形状に対応することができるモデルを生成することで様々な形状に変化する物体及び複雑な形状を持つ物体の追跡を可能とする手法である.図 2 に ASM による顔及び唇の追跡結果を示す.

図 2 ASM の追跡結果

対象物体をよく表すモデルを生成するために使用されるデータベースには,対象物体の形状をよく表すランドマークポイントのデータの位置座標の集合が格納されている.ここで,対象物体の形状をよく表すランドマークポイントとは,対象物体の輪郭を構成する線のうち,特徴的な曲線が始まる点や,直線同士の交点である.図 2 には本論文で使用するランドマークポイントを示す.顔と唇のランドマークポイントとして 68 点使用しており、唇が青い点,それ以外が赤い点である.顔の輪郭に 15 点,眉毛に 12 点,目に 10 点,鼻に 12 点,唇全体に 19 点使用されている.また,唇の外側の輪郭に 12 点,内側の輪郭に 7 点の割合で構成される.データベース中の画像は,様々な表情や発音をしている顔画像から構成されており,すべての画像にランドマークポイントが付けられている.

データセット中の画像の n 個のランドマークポイントは, 2n 個の要素をもつベクトル x として表すことができる.

$$\mathrm{x} = (x_1, x_2, \ldots x_n, y_1, y_2, \ldots y_n)^T \qquad (1)$$

ここでx_i, y_iはそれぞれ i 番目のランドマークポイントの x 座標と y 座標である. S 個の顔及び唇画像を含んだデータセットでは S 個のベクトル x を含む.
様々な形状の顔と唇の画像が十分にデータセットに含まれているならば, このランドマークポイントの集合をよく表すモデルを生成することで, データセットに含まれていない新しい画像中に現れる顔と唇の形状によくフィットするモデルを生成できる.
モデルとして, 全てのデータセットを生成し最もよくフィットするモデルを探索するのでは、次元が高く, 時間がかかってしまう. そこで次元削減を行い低い次元で形状を表すモデルを生成する必要がある. そのために主成分分析が適用される. データセット中の全てのベクトル x から分散共分散行列と平均を求めると, 主成分分析によって次元削減されたモデル X は以下のように表すことができる.

$$\mathrm{X} \cong \bar{x} + \mathrm{Pb} \qquad (2)$$

P =$(P_1|P_2|\ldots|P_t)$は, 主成分分析の結果の, 大きさについての上位 t 個の固有値に対応する固有ベクトルを含んだ行列であり, $\bar{x}$はデータベース中のベクトル x の平均ベクトルである. また, b は以下のように表される.

$$\mathrm{b} = P^T(x - \bar{x}) \qquad (3)$$

ベクトル b は, モデルを生成する際のパラメータとして使用され, b の値によって新しく生成されるモデル X の形状が変化する.
データベースに, 十分な数の顔画像から得られたランドマークポイントが含まれる場合, あるベクトル b の値は, 新しい画像中に含まれる顔と唇の画像の形状をよく表すモデルを生成する. この新しく生成されたモデルをモデルの座標空間から, 新しい画像中の顔と唇の座標空間に移動させることで, 入力画像中の顔と唇の画像を検出する. 移動は以下の式で表すことができる.

$$X' = T_{x_t, y_t, s, \theta}(\bar{x} + Pb) \qquad (4)$$

$$T_{x_t, y_t, s, \theta}\begin{pmatrix} X' \\ Y^7 \end{pmatrix} = \begin{pmatrix} x_t \\ y_t \end{pmatrix} + \begin{pmatrix} s\cos\theta & -s\sin\theta \\ s\sin\theta & s\cos\theta \end{pmatrix}\begin{pmatrix} x_i \\ y_i \end{pmatrix} \qquad (5)$$

関数$T_{x_t, y_t, s, \theta}$は, θ度の回転行列, 大きさ s のスケーリングを, (x_t, y_t)は平行移動を表す.

モデルを移動した後, 入力画像中にある顔と唇にフィットするパラメータは下記の二乗誤差を最小化することで推定される.

$$\mathrm{F}(\mathrm{b}, \theta\, \mathrm{s}, x_t, y_t) = |X - X'|^2 \qquad (6)$$

ここで X は探索される画像中の顔と唇のランドマークポイントの位置の点である.

対象に良くフィットするパラメータを推定する処理は, モデルと画像中の対象物体との位置座標だけでは不十分である. そのため, ランドマークポイント周囲の輝度値の情報を用いる.

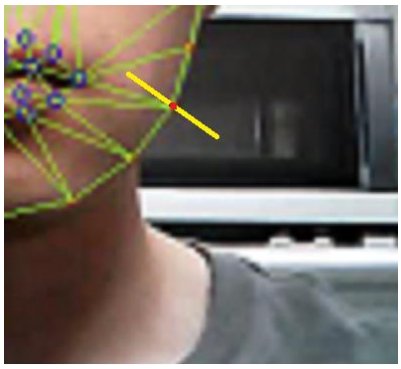

図 2 ランドマークポイント周辺のサンプリングの例

ここでランドマークとして指定した点をつなぐことでできるモデルの境界は強いエッジを持つことを利用する. エッジ方向の輝度値の情報は特徴的なものとなる.

データベース中の i 番目のランドマークポイントの法線方向に k ピクセル分の素値を境界の両側でサンプリングする. ランドマークポイントを含めた 2k+1 個のサンプルをベクトルg_iとする. 図 2 にサンプリングの例を示す. 中央の赤い点は対象のランドマークポイントであり黄色い線は幅 2k のサンプリングの位置である. サンプリングは長さ 2k の黄色い法線に沿って行われる.

ここで画像ごとの輝度値の変動を抑えるために以下の式を用いて正規化する.

$$g_i' = \frac{1}{\sum_j |g_{ij}|} g_i \qquad (7)$$

ここで, j はデータベースのすべての画像を示す.
これらの正規化した輝度値の平均を$\bar{g}_i'$, 共分散をS_{g_i}とする. これらはランドマークポイントの法線方向の輝度値の統計情報となる. これらの処理は全てのランドマークポイント

において反復的に行われる. 全てのランドマークポイントの輝度値の統計情報を並べたベクトルを$\bar{g}', S_g$とする.
式(6)と合わせ, 画像中の新しい顔と唇の画像の輝度値の統計g_sと生成されたモデルとフィットするパラメータは以下のように計算される.

$$\mathrm{F}(\mathrm{b}, \theta \mathrm{s}, x_t, y_t) = |X - X'|^2 = (g_s - \bar{g}')^T \mathrm{S}_g{}^{-1} (g_s - \bar{g}') \quad (8)$$

これはモデルの平均と新しいサンプルとのマハラノビス距離である.
式(8)を最小化するように各種パラメータを変化させることによって画像中の顔と唇を追跡する.

4. 特徴量

ASM によって顔と唇の追跡を行った後に特徴の抽出を行う.特徴抽出は後の機械学習の処理の精度を大きく左右するため重要な処理である.個々の発音の独自性をよく表す特徴を抽出できれば機械学習によって線形分離が可能である.
画像による読唇では,1 章で述べたように,大別して３つの特徴を使用する.即ち,それぞれ形状特徴,オプティカルフロー,空間周波数や固有値を利用した特徴である.本研究の手法では,画像ベースの手法として離散コサイン変換を用いた空間周波数を用いる.これらの特徴はそれぞれ時系列に並べられ後述するサポートベクトルマシンによって学習,分類される.
それぞれの特徴について以下に述べる.

4.1 形状特徴

形状特徴は,唇の形状をよく表す特徴であり,唇の幅や高さ,面積,周囲長などが考えられる.本手法では主に唇の幅と高さを用いる.

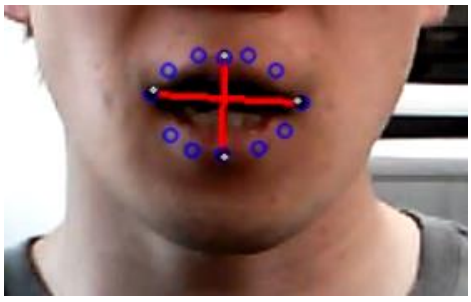

図 3 唇の外側の特徴点と唇の幅と高さ

ASM により唇の外側の 12 点を追跡する.このうち幅と高さを計算するために左右両端の２点と上下の２点を用いて,唇の幅と高さを計算する. 図 3 に ASM の追跡結果の唇の外枠の特徴点と,唇の幅と高さを示す.

求められた幅と高さは個人の唇の大きさやカメラとの距離に依存するため,処理の頑健性の欠如の原因となる.そこで,何も発音をしていない唇を閉じた時の唇の幅と高さを用いて正規化された幅と高さを特徴量とする.
これに加え,現在のフレームにおける唇の高さ/幅の比も形状特徴として用いる.

4.2 オプティカルフロー

オプティカルフローとは画像中のある点の動きを表したベクトルである.移動の距離と方向が動きのベクトルがオプティカルフローに対応する.
オプティカルフローを抽出するために唇の外側の１２点を使用する.また,このオプティカルフローの大きさも個人の唇の大きさやカメラとの距離に依存するため,何も発音をせず口を閉じている時の唇の高さと幅を用いることによって正規化する.

4.3 離散コサイン変換

離散コサイン変換により画像を周波数領域に変換すると,画像の多くの情報がその低周波領域に集中する.そのため,画像を全部使用せずとも低周波領域で画像をよく表現できることが多い.
２次元離散コサイン変換は以下の式で表される.

$$\mathrm{X}(k_1, k_2) = \frac{4C(k_1)C(k_2)}{NM} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} x(i,j) \cos\frac{(2i+1)k_1\pi}{2N} \cos\frac{(2j+1)k_2\pi}{2M}$$

$k_1 = 1,2,3 \ldots M-1, k_2$=0,1,2….N-1 ,

$\mathrm{C}(0)=\frac{1}{\sqrt{2}}$ C(k)=1 (k≠ 0) (9)

ここで x(i,j)は２次元画像の画素(i,j)の持つピクセル値である.
ASM によって計算された特徴点から、唇の中心を計算し,幅と高さから唇を含む大きさを求めることによって唇全体の画像が抽出される.抽出された唇の画像は離散コサイン変換によって周波数量域に変換しその低周波領域の 20×20 ピクセルの領域を離散コサイン変換による特徴として使用する.

5. Support Vector Machine

SVM は正のデータ, 負のデータの２種類の分類法である. 発音を分類するためには一種類の正解とそれ以外の不正解で学習データを構成し, 学習, 分類を行う必要がある.
ここでトレーニングデータのバランスが問題となる, １００種類の発音を分類するとする. １個の正解と９９個の不

正解のデータがある. これらを分類する分類境界を構成するときに,数少ない正解のデータからより多くのサポートベクトルを選ぶ必要がある.
データの構成がアンバランスな問題を解決するために SVM にコストファクターC を用いる[7].

$$\text{Minimize:}\quad \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i:y_i=1}\varepsilon_i + C\sum_{i:y_j=-1}\varepsilon_j$$

$$\text{s.t}\quad \forall k:\ y_k[\vec{w}\cdot\vec{x}_k + b] \geq 1 - \varepsilon_k \qquad (10)$$

i は正のクラス, j は負のクラスに属するデータである.
SVM は階層構造となっており、中間層の SVM では単一の特徴を用いた時の認識結果が出力され,最下層の SVM で中間層の SVM の出力をインプットとし、最終認識結果を出力する.コストファクターC は全ての SVM で独自に計算される.

6. 実験と考察

実験では,日本語を含むデータセットを用いて本論文で提案する自動読唇システムに適用し,認識率を計測する.実験に用いられるデータセットはすべて市販の web カメラで撮影されたものである.カメラは単語を発音している被験者の顔に焦点を合わせ,顔全体が含まれるように撮影する.暗くなるのを防ぐため,被験者の顔にはライトの光が当てられている.被験者は男性６人,女性３人の合計９人で構成ある.表 1 に示されている単語１５種類,数字１５種類をそれぞれ３回カメラに向かって発音してもらった.撮影されたビデオはそれぞれ単語データセット,数字データセットとして扱い,後述の実験で使用する.
ビデオのサイズは幅 480 ピクセル,高さ 640 ピクセルであり,一秒間あたり３０枚のサンプリングレートである.

表 1 データセット中の単語

単語データセット	数字データセット
Daikon, Izakaya, Kimono, Koi, Manga, Origami, Samurai, Shamisen, Sukiyaki, Sushi, Teppanyaki, Teriyaki, Tofu, Tunnami, Yakitori	2, 8, 9, 21, 39, 65, 72(ななじゅうに), 104, 257, 311, 423, 590, 781, 874, 953

6.1 実験 1

実験１では「単語」のデータセットを用いて自動読唇の精度を計測する.実験データに ASM を適用し,顔と唇の動画像を検出および追跡する.検出された画像から形状特徴,オプティカルフロー,離散コサイン変換の特徴量が抽出され,それぞれ時系列に整列される.整列された特徴量はそれぞれ SVM によって学習されるがこの時 SVM のパラメータ C を調整する必要がある.パラメータ C はサポートベクトルマシンにおいて,学習時に使用されるサポートベクトルの数をポジティブデータから多くとるか,ネガティブデータから多く取るかに関係するパラメータである.実験ではパラメータ C は、ポジティブデータを正のデータだと認識する認識率,ネガティブデータを負のデータと認識する認識率の合計が最大となる値とする.
表 2 に実験結果を示す.正の認識率は正しい発音を正しく分類できた割合で,負の認識率は間違ったデータを間違いと認識した割合である.合計値は正の認識率と負の認識率の合計である.負の認識率と合計値では提案手法が最も高いが正の認識率では若干低くなっている.これはサポートベクトルマシンのパラメータ C を調整するときに合計値がもっとも高くなるように学習しているからであり, 提案手法では単一の特徴量を用いた認識よりパラメータ C を調整する数が多いからだと考えられる.

表.2 実験 1 の結果の認識率の平均値

	提案手法	**離散コサイン変換**	**形状特徴**	**オプティカルフロー**
正の認識率	91.85	99.26	93.33	74.81
負の認識率	92.69	77.34	85.29	41.21
合計	184.55	176.6	178.62	116.02

次に表 3 には精度の平均値を示す.精度は以下のように定義される.

$$\text{精度}\quad \text{Accuracy} = \frac{TP+TN}{TP+FP+FP+FN} \times 100\% \qquad (11)$$

ここで,TP,TN はそれぞれ True Positive,True Negative であり,正のデータを正のデータと認識する数,負のデータを負と認識する数である.FP,FN は
False Positive, False Negative であり誤分類の数を示す.それぞれ正のデータを負と認識する数,負のデータを正と認識する数である.精度は提案手法が最も高いのが分かる.

表.3 実験結果の精度の平均値

	提案手法	離散コサイン変換	形状特徴	オプティカルフロー
精度	92.233	81.17	88.07	43.99

6.2 実験２

実験 2 では「数字」のデータセットを用いて自動読唇の精度を計測する.

表 4 に実験結果の平均値を示す.実験 1 の時と同様に合計値と負の認識率では提案手法が最も高いが，正の認識率では若干低くなっていることが分かる.

表.4 実験結果の認識率の平均

	提案手法	離散コサイン変換	形状特徴	オプティカルフロー
正の認識率	94.81	98.51	94.07	51.91
負の認識率	86.98	78.46	79.26	54.49
合計	181.79	176.97	173.33	106.40

また,表 5 には精度を示す.精度も実験 1 と同様に提案手法が最も高くなっているのが分かる.

表.5 実験結果の精度の平均

	提案手法	DCT	形状特徴	オプティカルフロー
精度	87.55626	80.04573	80.14978	57.62755

6.3 実験３

実験 3 では「単語」のデータセットと「数字」のデータセットを,一つのデータセットとして用いて自動読唇の精度がどのように変化するかを計測する.

表 6 には、実験 3 の単語と数字をまとめたデータセットと,実験 1 , 2 の単語のデータセット,数字のデータセットの結果の、正の認識率、負の認識率の平均と合計との比較を示す.正の認識率は,実験 3 の値は実験 1 より大きく,実験 2 より小さい値となっている.これは実験 3 では実験 1，2 で用いた単語と数字のデータセットを合わせた,データセットとなっているため,実験 3 の正の認識率は実験 1 の単語データの正のデータの認識率と実験 2 の数字のデータの正の認識率の間に収まると考えられる.

負の認識率の平均値は実験 3 が高くなっている,同様に合計の平均も高くなっている.これは、データセットをまとめた時に追加されたデータセットが,正のデータとは大きく異なるため,容易に負のデータと認識できたことが理由だと考えられる.

例えば,”311”の認識率を考えた時,実験 3 の時にデータセットに追加されたのは単語のデータセットであり”311”の発音とは大きく異なり,多くのデータは容易に負のデータとして分類できたと考えられる.

表 6　データの認識率の平均の比較

	実験３	**実験１**	**実験 2**
正の認識率	93.75	91.85	94.81
負の認識率	93.89	92.69	86.98
合計	187.64	184.55	181.79

7. 結論

本論文では人間の会話をしている動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した まず, Active Shape Model により, 動画像から顔及び唇の検出と追跡を行う.ASM によって検出された唇から形状特徴，オプティカルフロー，離散コサイン変換の 3 つの特徴量を抽出する.抽出した特徴量は中間層の SVM によってそれぞれ学習,認識され最下層の SVM によって結果を統合され最終的な認識結果を得る.

実験では被験者 9 人により「単語」15 個,「数字」15 個を含むデータセットの各単語を発話させ,自動読唇の実験を行った.

実験 1,2 では 3 つの視覚特徴量を用いることで従来の手法である単一の特徴のみを用いた認識の精度を大きく改善できることがわかった。

実験 3 ではデータセットの単語数を増やしても認識率が下がることはなく,むしろ精度に関しては上がっていることが分かった.30 個のデータセットで 90%の認識率が達成された.

実験結果の高い認識率と精度からから画像情報のみを用いた読唇システムの精度は高く十分に将来利用できる可能性がある技術であることが分かる.

コミュニケーションが難しい障害者の人々にとって 30～100 個の単語を用いてコミュニケーションをできることはまったくコミュニケーションができないことに比べ大きな差がある.

将来的な課題として,データセットをどのくらいまで増やしても高い認識率を保持するのか調査することが挙げられる.

本論文では最大 30 種類の単語の分類を行ったが, 100 種類, 1000 種類の単語でもどのくらいの認識精度を出せるか調査する必要がある.

また,将来的には自然言語処理などの文脈解析の技術と組み合わせて用いることで単語同士の意味の繋がりから,文脈を理解することも課題となる.

また,オプティカルフロー単一の特徴量を使った認識では結果が悪かった.これは従来のオプティカルフローによる

読唇の研究では唇全体のオプティカルフローを用いていたのに比べ,本手法では唇周りの 11 個の特徴点のみを用いていたことが原因であることが考えられる.オプティカルフローよる認識の精度をあげることで提案手法の複数の特徴量を用いた認識も精度が上がることが期待できる.

参考文献

1) Ayaz A. Shaikh, Dinesh K. Kumar, Wai C. Yau, M. Z. Che Azemin, ”Lip Reading using Optical Flow and Support Vector Machines” 2010 3rd International Congress on Image and Signal Processing (CISP2010) 327-330(2010) Word 2007 のヘルプと使い方
http://office.microsoft.com/ja-jp/word-help/CL010072933.aspx
2) 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”, テレビジョン学会技術報告 IETJ Technical Report vol. 13, No. 44, PP. 7-12 (1989)
http://office.microsoft.com/ja-jp/
3) Greg I. Chiou and Jenq-Neng Hwang, “Lipreading from Color Video” IEEE Transactions on Image procesing, Vol. 6, No. 8, 1192-1195 (1997)
4) Nakata, Yasuyuki, and Moritoshi Ando. ″Lipreading method using color extraction method and eigenspace technique.″ 電子情報通信学会論文誌 D-Ⅱ, Vol, J85-D-Ⅱ No.12, pp.1813-1822 (2002)
5) 斉藤剛史,小西亮介,唇および校内領域形状に基づいたトラジェクトリ特徴量による読唇” 第 6 回情報科学技術フォーラム (FIT2007), H-016, pp.39-40, (2007)
6) Cootes, Tim, E. R. Baldock, and J. Graham. ″An introduction to active shape models.″ Image Processing and Analysis (2000): 223-248.Booth, N. and Smith, A. S., [Infrared Detectors], Goodwin House Publishers, New York & Boston, 241-248 (1997)
7) Morik, Katharina, Peter Brockhausen, and Thorsten Joachims. ″Combining statistical learning with a knowledge-based approach-a case study in intensive care monitoring.″ *ICML*. Vol. 99. 1999.