ISS-A-050

スマートフォンで撮影された発話動画識別の検討

雫　優太†　　赤松　茂†
† 法政大学大学院理工学研究科応用情報工学専攻

1. はじめに

近年，音声認識技術の発達により，携帯電話での音声によるインターフェイスや音声認識による家電の操作などさまざまな音声認識技術が実用化されている．しかし，現在の音声認識技術には，実環境など雑音環境下で頑健に音声認識を行うことが難しいため，その解決手法の一つとして，唇動画像を用いた画像認識が注目され，研究が進められている[1]．従来の画像読唇研究では，室内での固定カメラで撮影されたデータを対象に解析等が行われていた．しかしながら，実際日常生活で撮影される動画は，スマートフォンなどで撮影された手振れなどのノイズが入ったものが多い．そのため，本研究では読唇技術を実用的に利用することを目的として，手振れ等のノイズが入った動画像に対する識別率向上の検討を，従来の研究[2]で有効性が示されているオプティカルフローとSVMを用いて行う．

2. オプティカルフロー

オプティカルフローとは，デジタル画像内の物体の動きをベクトルで表したもので，動画像にいくつかの仮定を置き画素の移動値を検出する方法である

3. Support Vector Machine

Support Vector Machine(SVM)とは，学習データを用いて複数のクラスを分類する線を生成し，その線に基づき新規データがどのクラスに分類されるかを識別する手法である．

4. 実験手法

4.1 データセット

データセットにはスマートデバイスを用いて撮影されたSSSD[3]を用いた．SSSDには25単語を発話した動画像が36名分含まれているデータセットである．本実験では，単語の長さによる影響を最小にするため，”ありがとう”，”おめでとう”，”こんにちは”，”こんばんは”，”さようなら”，”すみません”の6単語を利用した．

4.2 手振れの取得

差分は比較的動きが少ないと考えられる右頬，左頬，鼻の3カ所から得られたオプティカルフローを使用した．取得した領域を図1に示す．領域取得はSSSDに含まれる特徴点をもとに，発話動画の最初のフレームで行い，後のフレームでも同様の領域を取得した．

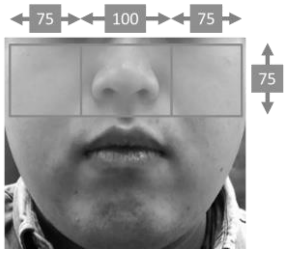

図1．オプティカルフローの差分取得領域

4.3 手振れによる影響を除いた解析

4.2でフレームごとに取得した速度ベクトルの中から，最頻した値をノイズによる速度ベクトルとみなし，本来の発話動画像から得られる速度ベクトルから減算した．その後，得られた速度ベクトルを学習データとしてSVMを作成し，Leave-One-Out法を用いて評価を行った．

5. 結果

結果を表1に示す．差分を除いたオプティカルフローを特徴量に用いることでより良い精度が得られた．

表1．差分を引いた際の識別結果

		差分として速度ベクトルを引いた個所			
		なし	右頬	左頬	鼻
発話単語	ありがとう	0.78	0.79	0.76	0.85
	おめでとう	0.68	0.74	0.80	0.77
	こんにちは	0.63	0.65	0.66	0.70
	こんばんは	0.56	0.59	0.54	0.63
	さようなら	0.67	0.72	0.69	0.73
	すみません	0.72	0.75	0.71	0.77
	平均	0.67	0.71	0.69	0.74

6. 今後の課題

より有効な差分個所の検討及び，不安定な環境における平行移動以外のノイズの存在有無を検証する．

謝辞

本研究の一部には，科学研究費補助金(基盤(C)19K12188)の助成を得た．

参考文献

[1] S. Agrawal, V. R. Omprakash and Ranvijay, ″Lip reading techniques: A survey,″ 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, 2016, pp. 753-757.

[2] 齊藤 剛史，窪川 美智子: SSSD:スマートデバイスを用いた読唇技術向け日本語データベース，電子情報通信学会技術研究報告，vol.117, no.513, pp.163-168, 2018.

[3] A.A.Shaikh, D.K.Kumar, W.C.Yau, M.Z.C.Azemin and J. Gubbi, ″Lip reading using optical flow and support vector machines,″ 2010 3rd International Congress on Image and Signal Processing, Yantai, 2010, pp. 327-330.