様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　　５　日現在

機関番号：　１７１０４
研究種目：　挑戦的萌芽研究
研究期間：　2015～2017
課題番号：　１５Ｋ１２６０１

研究課題名（和文）手指動作と非手指動作の重要度を考慮したマルチモーダル手話認識システムの開発

研究課題名（英文）Development of multimodal sign language recognition system considering the importance of manual signal and non-manual signal

研究代表者

齊藤　剛史（Saitoh, Takeshi）

九州工業大学・大学院情報工学研究院・准教授

研究者番号：１０３７９６５４

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：聴覚障害者の主なコミュニケーション手段である手話は、手指動作だけでなく、口唇の動きや顔の表情などの非手指動作を同時に利用する。画像処理技術を利用した手話認識は、非接触センサであるカメラを利用するため実用化が期待されている。しかし従来研究の多くは手指動作のみしか考慮されていない。本研究では手指動作を用いた手話認識だけでなく、読唇技術と表情認識技術についても検討した。研究をスムーズに進めるためにモーションセンサを用いたデータベースについても検討した。さらに手話観察時における注視点を解析することで、顔周辺、手指周辺あるいはその他のどこを注視しているか解析した。

研究成果の概要（英文）：Sign language is the main communication means of hearing impaired people. This uses not only finger movement but also non-fingers movement such as lip movement and facial expression. Sign language recognition (SLR) using image processing technology is expected to be put into practical use because it uses a camera which is a non-contact sensor. However, most related studies have considered only hand movements. In this research, we examined not only SLR using fingers movement but also lip reading technique and facial expression recognition technology. This research also studied database using motion sensor to make research progress smoothly. Furthermore, we analyzed gaze point at observation of sign language scene by using gaze point estimation technique.

研究分野：　人間医工学・リハビリテーション科学・福祉工学

キーワード：　手話認識　注視情報分析　読唇　表情認識

１．研究開始当初の背景

(1)画像処理技術を利用した手話認識は古くから研究されているが、手話動作データを精度よく取得することが困難な課題があった。近年、高精度かつ安価なモーションセンサが広く普及し、手話認識の期待が高まっている。研究代表者も手話認識に取り組み、手話単語認識および指文字認識の成果を公表し始めている。一方、手話者は手指動作だけでなく、口唇や表情などの変化も利用して意思を伝える。研究代表者は読唇技術や表情認識技術を学び、手話認識に必要な各モダリティの画像処理技術を併せもつ。以上の経緯より手話認識、読唇および表情認識を統合するマルチモーダル手話認識システム開発を着想した。

(2)みずほ情報総研と千葉大学の黒岩・堀内研究室は 2013 年より手話認識システムの共同開発に着手しているが手話動作のみを対象としている。また名古屋工業大学の北村は非手指動作に着目している。Koller らは手話者の口唇の動きを解析して視覚素（viseme）の認識について報告しているが精度は 47%と低い。これまでの手話認識で非手指動作が利用されていない要因は読唇技術の難しさにある。

２．研究の目的

聴覚障害者の主なコミュニケーション手段である手話は、手指動作だけでなく、口唇の動きや顔の表情などの非手指動作を同時に利用する。画像処理技術を利用した手話認識は、非接触センサであるカメラを利用するため、手話者と非手話者のコミュニケーション支援や手話教育に期待されている。しかし従来研究の多くは手指動作のみしか考慮されていない。本研究では手指動作だけでなく非手指動作を考慮する。さらに、手話会話における手話熟練者の注視点を解析することで、腕の動き、手の形状、口唇の動き、顔の表情など各モダリティの重要度を確認する。

３．研究の方法

(1)手話シーン観察時における注視点解析に基づく各モダリティの重要度推定

研究代表者らが開発したウェアラブルカメラを用いた注視点推定システムを利用し、手話シーン観察時に手話者の顔周辺、手指周辺、あるいはその他のどこを注視しているか解析する。これにより各モダリティの重要度を確認する。

(2)モーションセンサを利用したデータセット構築

マルチモーダル手話認識用のデータベース（DB）について、二つの DB を用いて研究を進めた。一つは、イタリア手話 20 単語が収録されており、Multi-Modal Gesture Recognition (MMGR) workshop on International Conference on Multimodal Interaction (ICMI)で利用された公開 DB である ChaLearn である。ChaLearn はモーションセンサである Microsoft Kinect で撮影されたカラー画像、距離画像、骨格情報および音声データが含まれる。図 1 に ChaLearn の画像例を示す。もう一つは、Kinect を用いて独自に日本語手話 100 単語の DB である J100word を構築した。特に ChaLearn については他研究グループとの比較として利用可能である。

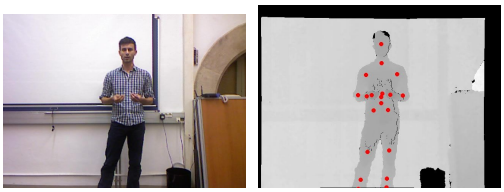

図 1　ChaLearn の画像例

(3)手話認識

本研究では、人手により設計した手の位置、手の動きおよび手型の 3 種類の特徴量を用いて、HMM により認識する手法を提案した（表 1 の Ours1）。また近年、機械学習分野で注目を集めている深層学習を用いたアプローチとして、前述の 3 種類の特徴量を入力として再帰型ニューラルネットワーク（RNN）を用いた認識する手法を提案した(表 1 の Ours2)。

また手話認識に関する従来手法では骨格情報を特徴に用いていた。しかし、Kinect より取得できる骨格情報の内、全て、上半身のみ、あるいは手のみなど考慮する骨格情報に違いがあるものの、手話認識に有効な骨格情報について検証されていない。そこで本研究ではこのことについても検証した。

(4)読唇

手話熟練者より、日本語手話では手指動作だけでなく口唇を動かすと聞いた。そのため、読唇技術を手話認識に導入することは有用であると考える。そこで研究計画に従い、読唇に関する研究に取り組んだ。

読唇分野においても、手話認識同様に深層学習が注目されており、本研究でも手話認識同様に深層学習を導入した。

2 種類のアプローチを検討した。一つは手話認識と同様に人手により設計した特徴量を用いて RNN により認識する手法である。ただし特徴量に関しては、顔特徴点を検出し、特徴点の動きに基づく特徴量（MF）および図 2 に示すような口唇周辺 ROI を抽出し、ROI に対して自己符号化器（autoencoder; AE）を適用して得られる特徴量を用いた。

もう一つの手法として、発話シーンよりフレーム画像をある規則に基づき取り出して連結した図 3 に示すようなフレーム連結画像（CFI）を提案し、CFI を入力として畳み込みニューラルネットワーク（CNN）により認識

する手法を提案した。

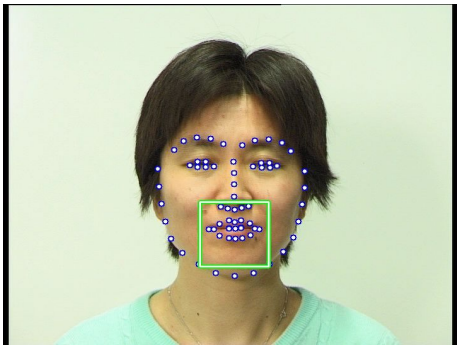

図 2　顔特徴点と口唇 ROI

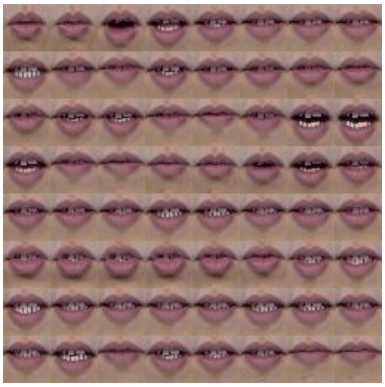

図 3　発話シーンのフレーム連結画像

(5)表情認識

表情認識については既存手法である LBP-TOP を用いた。ただし、従来のように図 4 左に示すような矩形 ROI でなく、図 4 右に示すような顔部位に基づく非矩形 ROI を提案した。

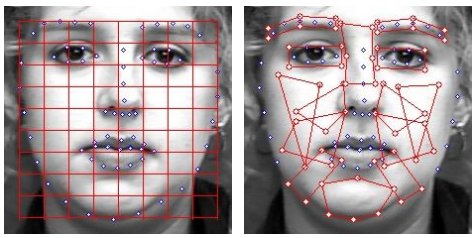

図 4　矩形 ROI と非矩形 ROI

4．研究成果

(1)手話シーン観察時における注視点解析に基づく各モダリティの重要度推定

手話ネイティブ 2 名、手話通訳者 1 名および手話未学習者 6 名、合計 9 名（A～I）の協力のもと図 5 に示すような環境で観察実験を実施した。観察対象の手話シーンは、全国手話検定試験の 4 級レベルと 5 級レベルのテキスト付属の DVD に収録されている 16 文を選んだ。解析結果を図 6 に示す。この図は、各被験者が手話 16 文を観察した際に、顔周辺、手指周辺、あるいはその他の 3 カテゴリについて、どれくらいの割合で注視していたのか表したグラフである。この結果より、手指よりも顔に注視点が集中する傾向が見られ、また手話未学習者よりも熟練者の方がその傾向が強いことを確認した。

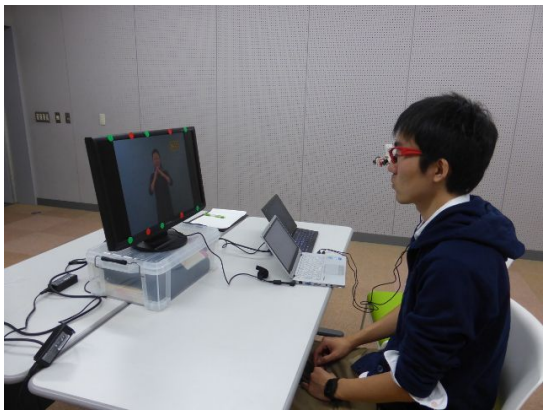

図 5　手話シーン観察実験の様子

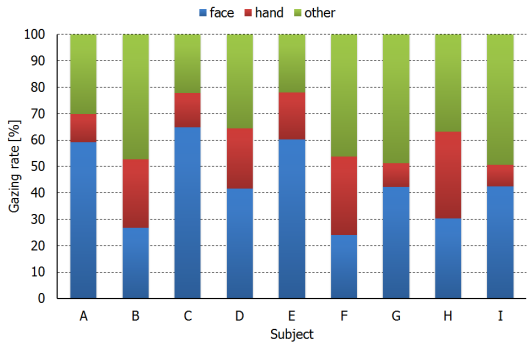

図 6　手話シーン観察時の注視点解析結果

(2)モーションセンサを利用したデータセット構築

ChaLearn は公開 DB であるため、データが整備されており以後の実験で用いた。独自収集した日本語 DB については実験で利用するために整備を進めた。

(3)手話認識

ChaLearn に対して認識実験結果を表 1 に示す。ただし PPTK、SUMO、ANRINKO は他研究グループの認識実験結果である。認識に用いるモダリティ、識別器の違いがあるものの、提案手法である Ours2 で最も高い認識率を得た。

表 1　認識結果（ChaLearn）

method	modalities	classifier	accuracy [%]
PPTK	color, depth, skeleton	GMM, HMM	82.7
SUMO	skeleton	Random forest	68.3
AURINKO	color, skeleton	ELM	36.7
Ours1	depth, skeleton	MS-HMM	65.0
Ours2	depth, skeleton	GRU	**87.1**

また手話認識に有効な骨格情報の検討に関しては、ChaLearn に対して、両手 H、両手首 W、両肘 E、両肩 S および首 N の各骨格を組み合わせた 31 通りに対して認識実験を実施した。実験結果を図 7 に示す。実験結果より、従来研究で用いられている手のみ、あるいは全ての骨格を用いるより両手および両肘の二つの骨格を用いることで高い認識精度を得られる新たな知見を得た。

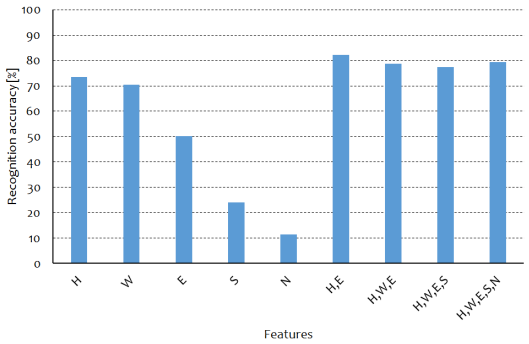

図 7　用いる骨格の違いによる認識結果

手話認識に関する国内外の研究では、Kinect を用いて手話シーンが撮影されてきた。しかし、2017 年 10 月に Microsoft 社が Kinect の生産を終了したことが公表された。このことは当初予期していなかった問題である。一方、深層学習を用いてカラー画像のみから骨格情報を推定する手法を実装したライブラリ OpenPose がカーネギーメロン大学より公開された。研究代表者は、手話シーンに対して OpenPose を適用した結果、正しく骨格が取得できることを確認した。そのため、今後の研究としてはカラー画像のみを用いるアプローチに研究の方向性をシフトする。これにより、新たに手話シーンを撮影せず、手話教材等で利用されていた手話シーンも利用可能となる。

(4)読唇

OuluVS や OuluVS2 など読唇向けの公開 DB に対して二つの提案手法を適用し、他手法よりも高い精度が得られることを確認した。

表 2 は OuluVS2 に対して他手法との比較結果である。OuluVS2 は正面顔のみでなく、カメラ 5 台で撮影した DB である。角度によって精度の違いがあるものの、提案手法は高い認識精度が得られている。

読唇に関しては、手話認識同様に DB 不足が課題である。特に読唇技術で利用可能な日本語の公開 DB は二つしかない。そこで研究代表者はスマートデバイスで撮影した発話シーンを収集し、公開 DB として整備を進めている。

(5)表情認識

実験には公開 DB である Cohn-Kanade（CK）を用いた。CK は 100 名以上の被験者から構成されており、6 表情が含まれている。認識実験の結果、非矩形 ROI よりも矩形 ROI の方が高い認識精度 96.1%が得られることが判明した。また他手法と比較した結果、他手法（Zhao+Pietikainen, 2007）で 96.3%が報告されていた。

表 2　認識結果（OuluVS2）

	front	30	45	60	profile
Baseline	74.8	—	—	—	—
Zhou et al.	73.0	75	76	75	70
Zimmermann et al.	73.1	75.6	67.2	63.3	59.2
Lee et al.	81.1	80	76.9	69.2	**82.2**
Ours	85.6	**82.5**	**82.5**	**83.3**	80.3
Chung et al.	**94.1**	—	—	—	—
Petridis et al.	84.5	—	—	—	—

5．主な発表論文等

〔雑誌論文〕（計　0　件）

〔学会発表〕（計　１２　件）

①児玉 知也、齊藤 剛史、手話認識に有効な骨格情報の検討、電子情報通信学会福祉情報工学研究会、2017

②Tomoya Kodama, Tomoki Koyama, Takeshi Saitoh, Kinect Sensor Based Sign Language Word Recognition by Mutli-Stream HMM, SICE Annual Conference, 2017

③児玉 知也、齊藤 剛史、再帰型ニューラルネットワークを用いたマルチモーダル手話認識、第 20 回画像の認識・理解シンポジウム、2017

④Takeshi Saitoh 他, Concatenated Frame Image based CNN for Visual Speech Recognition, ACCV2016 workshop: Multi-view Lip-reading/Audio-visual Challenges, 2016.

⑤橋村 佳祐、齊藤 剛史、距離画像のフレーム連結画像を用いた Convolutional Neural Network による手話単語認識、電子情報通信学会福祉工学研究会、2016

⑥Masaya Iwasaki, Takeshi Saitoh, LBP-TOP based Facial Expression Recognition using Non Rectangular ROI, International Conference on Information and Communication Technology Robotics, 2016

⑦祐宗 高徳、渋谷 昌尚、川田 健司、齊藤 剛史、手話シーン観察時の注視情報分析、電子情報通信学会パターン認識とメディア理解研究会、2016

〔図書〕（計　０　件）

〔産業財産権〕

○出願状況（計　０　件）

○取得状況（計　０　件）

〔その他〕
ホームページ等
http://www.slab.ces.kyutech.ac.jp/~saitoh

６．研究組織
(1)研究代表者
齊藤　剛史（SAITOH, Takeshi）
九州工業大学・大学院情報工学研究院・准教授
研究者番号：１０３７９６５４