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

令和　４ 年　 ６ 月 １６ 日現在

機関番号：２４４０３
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２０２３

研究課題名（和文）Vocabulary acquisition and 3D avatar approach for Japanese sign language communication

研究課題名（英文）Vocabulary acquisition and 3D avatar approach for Japanese sign language communication

研究代表者

ロイ　パルサプラティム（Roy, Partha Pratim）

大阪府立大学・研究推進機構・客員研究員

研究者番号：１０８３７２２２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：単語レベル手話認識(WSLR)の性能向上のために、最初の手法では、大域情報、局所情報および骨格情報に焦点を当てたシステムを提案した。局所情報は手の形と顔の表情から成る。骨格情報は体の位置に対する手の相対位置を表す。これら３つの情報を組み合わせることにより、提案方法は最先端の方法よりも高い認識性能を達成した。
2番目の手法では、元々動作認識のために提案されたI3Dネットワークを、WSLRの性能向上のために変更した。ここでの改善には、inceptionモジュールを改良したDIMと、ジェスチャーの本質的な特徴を特定するための注意メカニズムに基づくTAMが含まれる。

研究成果の学術的意義や社会的意義
Word-level Sign Language Recognition (W-SLR) systems overcome the communication barrier between people with speech impairment and those who can hear. In our approach, we combined these local and relative position of body parts and achieved higher performance on most W-SLR datasets.

研究成果の概要（英文）：To improve the performance of existing word-level Sign Language Recognition (W-SLR), in our first approach, a system with a multi-stream structure focusing on global information, local information, and skeletal information was proposed. The local information comprises of handshape and facial expression. The skeleton information captures hand position relative to the body. By combining these three streams, the proposed method achieves higher recognition performance than the state-of-the-art methods.
In the second work, the original I3D network which was originally proposed for action recognition problems has been modified to improve the WSLR performance. The improvement includes an improved inception module named dilated inception module (DIM) and an attention mechanism-based temporal attention module (TAM) to identify the essential features of gestures.

研究分野：Pattern Recognition, Image Processing, etc.

キーワード：Sign lang. recognition　3D conv. neural networks　Deep learning　Attention Network

１．研究開始当初の背景

Our primary goal is to develop a prototype system for a Sign Gesture Recognition and Translation system that will be able to convert a performed sign gesture into its corresponding meaning in the form of text/speech. Similarly, it will be able to convert an entered text/speech input (available for conversation) into the corresponding gesture with the help of an animated actor ('Avatar'). It will help to bridge the communication with normal people and speech-impaired people. We plan to release the system public so that many people who really require the system can benefit from it. Also, we will write research papers based on our contributions and submit them in peer-reviewed International/National conferences and journals.

２．研究の目的

Sign language may also help to bridge the communication gap between the speech and hearing of the majority community. However, gaining expertise for daily sign language use demands a significant amount of effort and learning time. Training the majority of the population in sign language is not a practical solution. Additionally, sign language also depends on spoken language and culture. For example, people in the United Kingdom and China have different sign languages, which further limits its potential to become popular globally. Low cost and the availability of mobile devices equipped with RGB sensors brings the sign language recognition (SLR) field inside the domain of computer vision. With the advancement of machine learning (ML), significant progress has been achieved in computer vision-related tasks. Hence, it is an excellent idea to explore SLR, which can automatically interpret sign language to support deaf and speech impaired people for smooth and effective communication with the hearing majority community. In the last decade, several studies have been conducted that have targeted SLR using ML techniques and deep learning on images/video captured by RGB sensors

３．研究の方法

In the first project, The multi-stream neural networks (MSNN) are designed for the Word-level sign language recognition (WSLR). As shown in Fig. 1, the multi-streams consist of three streams: 1) a base stream, 2) a local image stream, and 3) a skeleton stream. Each stream is trained separately, and the recognition scores extracted from each stream are averaged to obtain the final recognition result. The designed model incorporates both global and local features that help to achieve an improvement of 10%–15% in Top-1 accuracy compared with conventional methods on the WLASL[2] and MS-ASL[3] datasets.

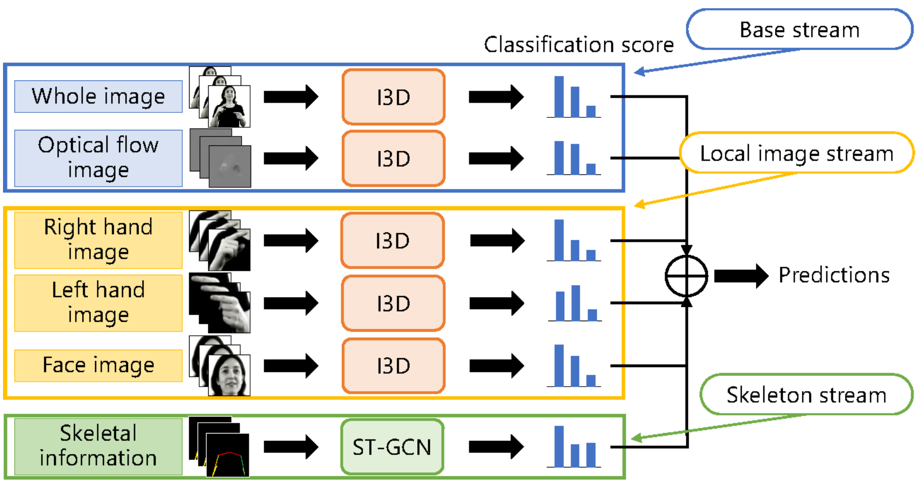

Figure 1: Proposed multi-stream networks

In other work, we revisit the renowned I3D network [1], which was proposed for action recognition problems. The network is modified for the WSLR task. We revisit the I3D model to extend its performance in three essential design aspects. They include an improved inception module named dilated inception module (DIM) and an attention mechanism-based temporal attention module (TAM) to identify the essential features of signs. Additionally, we propose eliminating a loss function that deteriorates performance. The extensively validated the proposed method on the WLASL[2] and the MS-ASL[3] public datasets. The proposed method outperformed state-of-the-art approaches on the WLSAL dataset and produced competitive results on the MS-ASL dataset. The improvement in the top-1 accuracies of the proposed method compared with the I3D model for SLR on WLASL100 and MS-ASL100 is around 15% and 10%, respectively.

４．研究成果

We designed a word-level Sign Language Recognizer in two different works. In the first work, "Word-level Sign Language Recognition with Multi-stream Neural Networks Focusing on Local Regions and Skeletal Information," experimental results of proposed multi-stream models achieved 81.38%, 73.43%, 63.61%, and 47.26% in Top-1 accuracy on the WLASL100, WLASL300, WLASL1000, and WLASL2000 datasets, respectively, which are higher than the results of the conventional methods using only global information. Moreover, to verify the effectiveness of the local image and skeleton streams, the recognition performance of the model with and without each stream was compared. As a result, the models with all three streams achieved higher recognition accuracies than the other models. This confirms that the three streams used in our method were effective for WSLR. Moreover, in the experiments on the MS-ASL dataset, the proposed method achieved 83.86%, 80.72%, 65.46%, and 49.06% in Top-1 accuracy on the MS-ASL100, MS-ASL200, MS-ASL500, and MS-ASL1000 datasets, respectively. These results were better than those of the conventional methods on all datasets except for the MS-ASL100 and MS-ASL1000 datasets. Therefore, these results confirm that the proposed method is not a data-specific method, but a highly versatile method for WSLR and 83.86%, 80.72%, 65.46%, and 49.06% in Top-1 accuracy on the MS-ASL100, MS-ASL200, MS-ASL500, and MS-ASL1000 datasets, respectively. In the second work, "Revisiting I3D for Sign Language Recognition," experimental results is 79.08%, 68.75%, 49.32%, and 34.55% in Top-1 accuracy on the WLASL100, WLASL300, WLASL1000, and WLASL2000 datasets, respectively, and 82.78%, 77.24%, 69.13% and 50.83% in Top-1 accuracy on the MS-ASL100, MS-ASL200, MS-ASL500, and MS-ASL1000 datasets, respectively. The second work increased the Top1 accuracy by around 13% and 9% on WLASL100 and MS-ASL100, respectively, compared to the I3D model.

References:-
1. J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
2. D. Li, C. Rodriguez, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1459–1469.
3. H. R. Vaezi Joze and O. Koller, “MS-ASL: A large-scale data set and benchmark for understanding american sign language,” in proceedings of the British Machine Vision Conference, 2019.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 丸山 瑞己, Shuvozit Ghose、井上 勝文、Partha Pratim Roy、岩村 雅一、吉岡 理文
２．発表標題 局所領域に着目したMulti-stream Neural Networks による手話単語認識
３．学会等名 電子情報通信学会技術研究報告 パターン認識・メディア理解研究会(PRMU)
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	岩村　雅一 (Iwamura Masakazu) (80361129)	大阪府立大学・工学（系）研究科（研究院）・准教授 (24403)	
研究分担者	井上　勝文 (Inoue Katsufumi) (50733804)	大阪府立大学・工学（系）研究科（研究院）・准教授 (24403)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関