International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 07 | July 2022 www.irjet.net p-ISSN: 2395-0072

Sign Language Recognition using Facial Gesture and Skeleton Keypoints

Sheela N[1], Kiran Raghavendra[2], Shashank C[2], Sanjana S[2], Dhanush N S[2]

[1]Assistant Professor, Dept of Computer Science Engineering, JSS Science and Technology University, Mysuru
[2]Dept of Computer Science Engineering, JSS Science and Technology University, Mysuru

---

Abstract - *Sign Language is used by people who are speech impaired or hard of hearing. Sign Language Recognition aims to recognise signs performed by the signer in the input videos. It is an extremely complex task as signs are performed with complicated hand gestures, body posture and mouth actions. In recent times, skeleton based models are preferred for sign language recognition owing to the independence between subject and background. Some sign languages make use of mouthings/facial gestures in addition to hand gestures for signing specific words. These facial expressions can be used to assist the convoluted task of sign language recognition. Skeleton based methods are still under research due to lack of annotations for hand keypoints. Significant efforts have been made to tackle sign language recognition using skeleton based multi-modal ensemble methods, but to our knowledge none of them take facial expressions into consideration. To this end, we propose the usage of face keypoints to assist skeleton based sign language recogntion methods. As a result, skeleton based methods on addition of facial feature information achieves an accuracy of 93.26% on AUTSL dataset.*

Key Words: **Sign language recognition, Skeleton based methods, Face expression, SLGCN, SSTCN, Wholepose keypoint estimator, AUTSL dataset**

1. INTRODUCTION

Sign Language is a means of communication for people who are hard of hearing or speech impaired. It is a visual language involving hand gestures, body posture and mouth actions. Comprehending sign language requires remarkable effort and training which is not feasible for the general public. In addition sign language is affected by the language of communication (e.g., English, Chinese, Italian) and region of usage (e.g., American Sign Language, Indian Sign Language). With advancements in computer vision and machine learning it is essential to explore sign language recognition (SLR) which translates sign language and helps the deaf/speech impaired community to communicate easily with others in their daily life.

In comparison with action recognition or pose estimation, SLR is a extremely challenging task. Firstly, SLR requires information of global body motions and intricate movement of hands and fingers to express the sign correctly. Similar signs can interpret different meanings depending on the number of times it is repeated. Secondly, different signers perform signs differently (e.g., speed, body shape and posture, left handed or right handed) thus making SLR challenging.

Inspired by the recent developments on SLR using multi-modal methods [4], we propose the usage of pretrained facial keypoint estimators to provide additional facial gesture information to the SLGCN + SSTCN ensemble framework proposed in [4].

We propose the usage of SLGCN [4] and SSTCN [4] to exploit facial gesture information using face keypoints generated using pretrained estimators thus assisting the complex task of sign language recognition.

2. RELATED WORK

In this section, we review existing publicly available datasets for sign language, and existing state-of-the-art algorithms for sign language recognition.

2.1 Sign Language Datasets

A Word Level American Sign Language (WLASL) dataset is proposed in [1], containing over 2000 words performed by over 100 signers.

A Turkish Sign Language dataset is proposed in [6]. The dataset consists of 226 signs performed by over 43 different and 38,336 sign video samples in total. Samples contain a variety of videos in different backgrounds (both indoor and outdoor environments).

Reference [3] introduces a 3D hand pose dataset based on synthetic hand models.

2.2 Sign Language Recognition Approaches

An appearance based approach and 2D human pose based approach is proposed in [1] creating baselines that aid method bench marking. In addition, [1] proposes a pose-based temporal graph convolution networks (Pose-TGCN) that models spatial and temporal dependencies.

A Two-Stream Inflated 3D ConvNet (I3D) is proposed in [7]. It is based on 2D ConvNet inflation: filters and pooling kernels of very deep image classification convnets are

expanded into 3D, thus making it possible to learn spatio-temporal features from video.

Skeleton Aware Multi-modal SLR framework (SAMSLR) is proposed in [4] to take advantage of multi-modal information thus aiding sign language recognition. RGB and depth modalities are also added and assembled into the final framework to provide information that is complementary to the SL-GCN and SSTCN methods.

Reference [2] focuses on the translation system and introduces the STMC-Transformer which improves on the current state-of-the-art by over 5 and 7 BLEU respectively on gloss-to-text and video-to-text translation of the PHOENIX-Weather 2014T dataset [8]. An approach that estimates hand pose from RGB images has been proposed in [3]. Refernce [3] proposes a deep network that learns a network-implicit 3D articulation. Together with detected keypoints in the images, this network yields good estimates of the 3D pose. A method to automatically find compact and problem-specific topology for spatio-temporal graph convolutional networks in a progressive manner has been proposed in [5]. Summary of usage of facial expressions by signers and an account of the range of facial expressions has been proposed in [9]. It is done by making use of the three dimensions on which facial expressions vary: semantic, compositional, and iconic. In [12], a sign language recognition model is created using Convolutional Neural Networks (CNNs), Feature Pooling Module and Long Short-Term Memory Networks (LSTMs). In the CNN part, a pre-trained VGG-16 model is used, after adapting its weights to the dataset. The extracted features are used to generate multi-scale features. The features matrices are reduced to feature vectors, using Global Average Pooling (GAP). The features that are obtained passed to the LSTM architecture after instance normalization, which generates the text.

Reference [14] investigates the task of 2D human whole-body pose estimation, with the aim of localizing dense landmarks on the entire human body including face, hands, body, and feet. Different deep models are trained independently on different datasets of the human face, hand, and body. In addition, [14] introduces COCO-WholeBody which extends COCO dataset with whole-body annotations i.e 133 dense landmarks with 68 on the face, 42 on hands and 23 on the body and feet. Neural network model ZoomNet is proposed in [14], which takes into account the hierarchical structure of the full human body to solve the scale variation of different body parts of same person.

Table -1: Statistical summary of AUTSL dataset

Subsets	Samples
Train	28142
Validation	4418
Test	3742

3. AUTSL DATASET

The Ankara University Turkish Sign language dataset [6] is gathered to perform general Sign Language Recognition tasks. It consists of 226 signs performed by 43 different signers and 38,336 isolated sign video samples in total. Samples contain a wide variety of backgrounds recorded in indoor and outdoor environments. Moreover, spatial positions and the postures of signers also vary in the recordings.The collection method uses Kinect V2 sensor [10],[11]. Each sample is recorded with Microsoft Kinect v2 and contains color image (RGB), depth, and skeleton modalities. Specifically it is split to training, validation, and testing categories as shown in Table. 1.

Fig-1: Surprise Sign

4. PROPOSED METHOD

In this section we discuss various deep learning model benchmarks. We consider two baseline models , the model benchmarked in [6] as well as the SLR challenge leaderboard (Baseline RGB and Baseline RGB-D) model. In paper [12] CNN + LSTM structure is employed for the construction of the model and VGG-16 model is used to extract the features for each video clip, which is given input to the LSTM to generate text.

We consider the SLGCN and SSTCN model used in [4] as a baseline and create an ensemble of the two models with addition of facial features generated using wholepose keypoint estimator proposed in [14].

The SLGCN model used in [4] uses 27 key points. 7 keypoints to model the upper body and 10 keypoints to model each arm are considered. From these keypoints bone data, joint data, bone motion data and joint motion data is extracted. This is given input to the SLGCN network. The SSTCN model considers 33 key points i.e. 4 landmarks on the lips and 1 landmark on the nose has been considered in addition to the 27 keypoints mentioned above. As mouth actions and facial gestures are essential for better understanding of sign language, we explore sign language recognition by using face landmarks to capture variety of facial expressions e.g., eye brow raise while signing "surprise" as shown in Fig. 1.

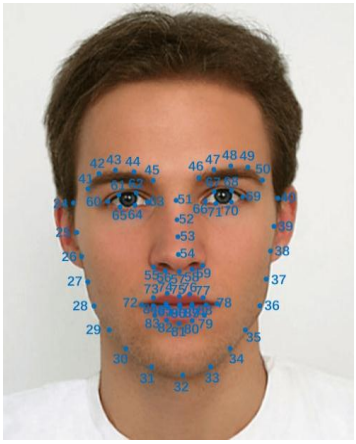

Fig-2: Face keypoints using wholepose estimator

We make use of wholepose keypoint estimator proposed in [14] to generate whole body keypoints. Specifically we use 61 keypoints in total i.e 6 keypoints for left eye, 6 keypoints for right eye, 5 keypoints for left eye brow, 5 keypoints for right eye brow, 12 keypoints for mouth as shown in Fig. 2 and 27 key points defined in [5](10 keypoints for each hand and 7 keypoints for the upper body). The ensemble of the two models is used to make sign language predictions as shown in Fig. 3.

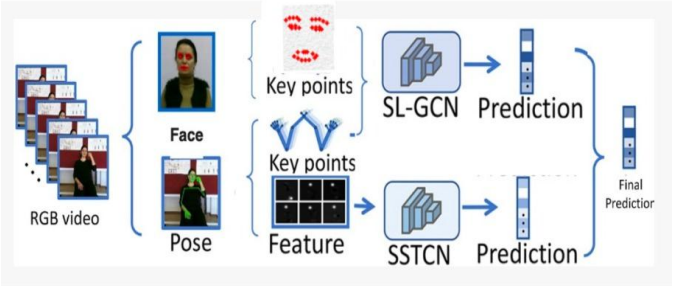

Fig-3 : SLGCN + SSTCN ensemble with facial features

5. RESULTS

In this section, the performance of SLGCN+SSTCN ensemble model with facial features is presented. For testing, we make use of the test set of the AUTSL dataset [6] consisting of 3742 examples. In an attempt to capture facial expressions and mouth gestures, we have added 34 face key points. On this addition, we noticed a Top1 accuracy of 93.26 percent and Top 5 accuracy 95.66 percent as shown in Table 2.

The above results can be attributed to the imbalance between signs with facial expressions and signs without facial expressions in the AUTSL dataset i.e, the AUTSL dataset consists very few signs having facial expressions, thus resulting in an absence of clear patterns in facial expressions.

Table - 2 : Performance of SLGCN+SSTCN ensemble

SLGCN+SSTCN Ensemble	Top1	Top5
inclusion of facial features	93.26	95.66

6. CONCLUSION

In this paper, we propose the addition of facial features to capture facial gestures and mouth actions, thus adding valuable information to the task of sign language recogntion.

We construct a skeleton graph for Sign Language Recognition using pretrained whole pose estimators and make use of SLGCN [4] to model temporal and spatial dynamics of facial keypoints and skeleton keypoints and SSTCN [4] to extract information from skeleton features. The proposed methods has the ability to provide fruitful results on sign languages that heavily relies on facial expression.

REFERENCES

[1] Li, Dongxu & Rodríguez, Cristian & Yu, Xin & Li, Hongdong. (2019). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison.

[2] Yin, Kayo & Read, Jesse. (2020). Better sign language translation with STMC-transformer. 5975-5989. 10.18653/v1/2020.coling-main.525.

[3] Zimmermann, Christian & Brox, Thomas. (2017). Learning to estimate 3D hand pose from single RGB images. 4913-4921. 10.1109/ICCV.2017.525.

[4] Jiang, Songyao & Sun, Bin & Wang, Lichen & Bai, Yue & Li, Kunpeng & Fu, Yun. (2021). Sign language recognition via skeleton-aware multi-model ensemble.

[5] Heidari, Negar & Iosifidis, Alexandros. (2020). Progressive spatio-temporal graph convolutional

network for skeleton-based human action recognition.

[6] Mercanoglu, Ozge & Keles, Hacer. (2020). AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods. IEEE Access. 8. 181340-181355. 10.1109/ACCESS.2020.3028072.

[7] Carreira, J. & Zisserman, Andrew. (2017). Quo Vadis, Action Recognition? A new model and the kinetics dataset. 4724-4733. 10.1109/CVPR.2017.502.

[8] Forster, Jens & Schmidt, Christoph & Hoyoux, Thomas & Koller, Oscar & Zelle, Uwe & Piater, Justus & Ney, Hermann. (2012). RWTH-PHOENIX-Weather: A large vocabulary sign language recognition and translation corpus.

[9] Elliott, Eeva & Jacobs, Arthur. (2013). Facial expressions, emotions, and sign languages. Frontiers in psychology. 4. 115.

[10] Diana Pagliari and Livio Pinto. Calibration of kinect for Xbox One and comparison between the two generations of microsoft sensors. 15:27569–27589, 10 2015.

[11] Clemens Amon, Ferdinand Fuhrmann, and Franz Graf. Evaluation of the spatial resolution accuracy of the face tracking system for Kinect for windows V1 and V2. In Proceedings of AAAI Conference on Artificial, pages 16–17, 2014

[12] Ozge Mercanoglu Sincan, Anil Osman Tur, and Hacer Yalim Keles. Isolated sign language recognition with multi-scale features using LSTM, 2019.

[13] Colin Raffel and Daniel PW Ellis. Feed-forward networks with attention can solve some long-term memory problems. arXiv:1512.08756, 2015.

[14] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. 2020.