ISSN No:-2456-2165

Sign Language Recognition System

Likhitha K[#], Sahana H J[*],Niharika B R, Abhishek Raju[#] , Prathima M G[#]
[#]Computer Science Department, B.I.T. Bangalore

Abstract:- The goal of vision-based sign language recognition is to improve communication for the hearing impaired. However, the majority of the available sign language datasets are constrained. Real-time hand sign language identification is a problem in the world of computer vision due to factors including hand occlusion, rapid hand movement, and complicated backgrounds.

In this study, we develop a deep learning-based architecture for effective sign language recognition using Single Shot Detector (SSD), 2D Convolutional Neural Network (2DCNN), 3D Convolutional Neural Network (3DCNN), and Long Short-Term Memory (LSTM) from Depth and RGB input films.

Keywords:- *Sign Language Recognition System, Multi Modal Approach, Skeleton Based.*

I. INTRODUCTION

Since sign languages have distinctive linguistic patterns, they are largely employed as a means of communication by the deaf community. A loss of socialisation may occur when deaf-mute people are unable to communicate with members of the hearing community. The caregiver must communicate with the deaf-mute person in these circumstances. Therefore, it is crucial to create a continuous sign language recognition system that does not require any apparatus on the hands in order to give these two populations of individuals similar communication opportunities.

An extensive amount of study has been done to improve the understanding of hand sign language. However, there are still a number of difficulties, including real-time performance, hand occlusion, quick hand movements, and many others.

Several suggested models combine characteristics in various ways. We are expanding on the 2D hand skeleton model, allowing different view points to learn the features of the hand and leveraging the dynamics as a local spatio-temporal as well as long-term features from the LSTM model. While just the features relating to 2D views of the hand have been included in these models. Our model makes use of several and complimentary sources of data, including heatmap features, pixel level appearance, and hand skeleton view. Utilizing 3DCNN and LSTM, we can learn how to take advantage of the preceding 2DCNN model along with the other feature representations. This model also contains a simple identification procedure: SSD to concentrate on the region of interest, 2D CNN to extract spatial discriminative features, 3D CNN to extract local spatio-temporal dynamics from various feature representations, and LSTM to recognise final sign labels. Our model's output demonstrates consistent performance gains for the suggested methodology across all tested datasets. In order to promote two-way communication between the hearing impaired and the general public, we propose a deep learning based architecture for effective continuous sign language recognition. We do continuous sign identification of vast sign vocabularies using real-time input from a webcam on a phone, laptop, or other device. The input's first SSD architecture is utilised to detect the user's hands. Additionally, a 3D hand skeleton is created utilising the 3D keypoints to provide numerous views of the hand. Then, five 3DCNNs are fed with hands,RGB Video , and a Depth Video. Each 3DCNN extracts spatio-temporal complementary information, such as features related to appearance, geometry under various camera angles, and features from hand joint prediction scores. For the final categorization of the hand, the outputs of the five 3DCNNs are concatenated, ensembled and sent to an LSTM. Finally, we can use the expansion of 3D local spatiotemporal features for long-term sign modelling.

II. LITERATURE REVIEW

Sign language recognition system (SLR) has achieved great progress and high recognition accuracy in recent years by developing a practical deep learning architecture and improving computing power [1, 2, 3, 4, 5, 6, 7, 8, 9]. The remaining task of SLR is to capture all the body movement information and local arm,hand,facial expressions at the same time. [1] proposes to use a linear discriminant analysis (LDA) algorithm for hand gesture recognition to convert the recognized hand gestures into text and speech formats. This paper [2] proposes a hand gesture recognition method based on YCbCr color space, COG, and template matching. In [3], the proposed system utilizes LMC as a sensor and SVM and DNN is utilized for data training. Study [4] is creating a network that can effectively classify images of static signs by equivalent text from CNN. They present a glove system that recognizes numbers from [5] sign language. The KNN algorithm is used as a classifier. Model [6] recognizes characters using SVM and FMG algorithms. In [7],RNN is used to capture the long-term time dependence between the inputs and 2D-CNN is used to extract spatial features from the input. The system proposed in [8] converts the cross-domain knowledge into message tokens to improve the accuracy of the WSLR model. The proposed system [9] presents a transformer-based learning system for recognizing continuous sign language and translating it to text , This is achieved using connectionist temporal classification (CTC). These methods are not yet effective enough to get complete motion information.

Multi-modal Approach multi modal approach is an end to end framework which provides users with a convolution architecture to exploit different features captured from an image, word composition and the matching relation between the two modalities. Different modalities might contain different information related to the hand gesture which can complement each other and provide us with a distinctive representation of the action.

The [4] paper proposes a effective method utilizing super vector to fuse different multi view representation together

Skeleton Based Action Recognition Skeleton based action recognition is the process of recognizing action using the skeleton data obtained from the image. The skeleton data is nothing but the information related to the 2D or 3D coordinates of the human skeletal joints. It can also be used along with other modalities to achieve a multi modal representation of the action. Usually we use recurrent neural networks to model skeleton data.

The [10] proposes a Graph based approach to model the changing patterns of skeleton data using Graph convoluted network (GCN). This method is also termed as ST-GCN. But still the skeleton based sign language recognition systems are still under explored. The [12] tried to extend STGCN to SLR but was unsuccessful to achieve higher accuracy and used only 20 sign classes.

III. OUR APPROACH

SSTCN - Separable Spatial Temporal Convolution Network.

This architecture has proposed an SSTCN to further exploit whole body skeleton features, which can significantly improve the accuracy on whole-body key points compared with the traditional 3D convolution. Besides using key point coordinates generated from the whole-body pose network, an SSTCN model to perceive the sign language from whole-body features is also proposed.Features of 33 key points from 60 frames of each video as the input to our model is extracted , which contains 1 landmark on the nose, 2 landmarks on shoulders, 4 landmarks on mouth, 2 landmarks on wrists, 2 landmarks on elbows and 22 landmarks on hands.

Instead of using 3D convolution, the input features are processed with a 2D convolution layer separably, which reduces the parameters and makes it easier to converge.

Pose - Word-level sign language recognition is the fundamental building block for interpreting sign language sentences. signalling a sign language word requires very subtle body movements that make WSLR a particularly challenging problem. The human skeletal motion plays a significant role in conveying what word the person is signalling. Hence a pose-based model is used to tackle the problem of WSLR. Human pose estimation involves localising key points of human joints from a single image or a video.The 27-node skeleton graph was constructed using a pretrained HRNet whole-body pose estimator given by MMPose to estimate 133-point whole-body keypoints from RGB videos. The graph is divided into four streams (joint, bone, joint motion and bone motion). As data augmentations, random sampling, mirroring, rotating, scaling, jittering, and shifting are used.

RGB - To facilitate parallel loading and processing, all frames of RGB videos are extracted and saved as pictures. Based on the important points derived from whole-body posture estimation, RGB and optical flow frames are cropped and scaled to 256 x 256 pixels.

Optical Flow - The TVL1 technique, which is implemented with OpenCV and CUDA, is used to obtain optical flow features.The output flow maps of x and y directions are concatenated in channel dimension.

HHA - stand for horizontal disparity, height above the ground and angle normal. The HHA representation encodes properties of geocentric pose that emphasize complementary discontinuities in the image (depth, surface normal and height) because of which it works better than using raw depth images for learning feature representations with convolutional neural networks.

Depth Flow - HHA features the same way as the RGB frames in data augmentation. Besides, the exact same procedure used for RGB to extract optical flow from the depth modality (named depth flow). The depth flow is cleaner and captures different information compared with the RGB flow

RGB Ensemble - Ensemble is to combine the predictions from multiple neural network models to reduce the variance of predictions and reduce generalisation error. Here, The skeleton-based technique, which incorporates SL-GCN and SSTCN, outperforms RGB + Flow and Depth ensemble models.

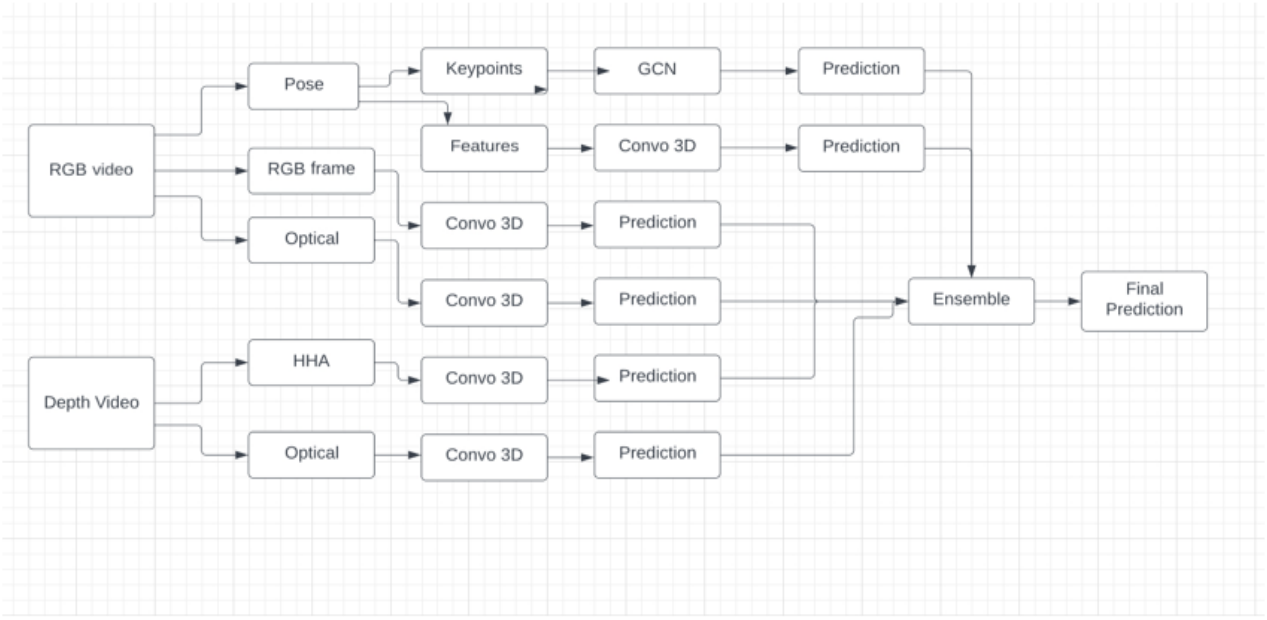

Fig 1:- Our Approach to sign Language Recognition System Using a Multi-Modal Ensemble

The ensemble results of RGB All and RGB-D All demonstrate that the whole-body skeleton based approaches are able to collaborate with the other modalities and further improve the final recognition rate.

Multi ensemble model - We use a simple ensemble method to ensemble all four modalities above. Specifically, we save the output of the last fully-connected layers of each modality before the softmax layer.

IV. IV. RESULT AND ANALYSIS

In this section, we present evaluation of our proposed framework on the AUTSL dataset.

➢ *Autsl Dataset* –

Is collected for general SLR tasks in Turkish sign language. Kinect V2 sensor is utilized in the collection procedure. Specifically, 43 signers with 20 backgrounds are assigned to perform 226 different sign actions. In general, it contains 38,336 video clips which is split to training, validation, and testing subsets. The statistical summary of the balanced dataset, which is used in the challenge, is listed in Table

➢ *Baseline Method* –

Along with the AUTSL benchmark, several deep learning based models are proposed. We treat the best model benchmarked in [12]. Specifically, the model is mainly constructed using CNN + LSTM structure, where 2D-CNN model are used to extract feature for each video frame and bidirectional LSTMs (BLSTM) are adopted on top of the these 2D CNN features to lean their temporal relations. A feature pooling model (FPM) [12] is plugged in after the 2D

CNN model to obtain multi-scale representation of the features.

Modality	Top-1	Top-5
Baseline RGB	42.58	-
Baseline RGB-D	63.22	-
Keypoints	**95.45**	**99.25**
Features	94.32	98.84
RGB Frames	94.77	99.48
RGB Flow	91.65	98.76
Depth HHA	95.13	99.25
Depth Flow	92.69	98.87

Table 1:- Results of single modalities on AUTSL dataset.

	Finetune	Track	Top-1
Baseline	-	RGB	49.23
Baseline	-	RGB-D	62.03
Ensemble	No	RGB	97.51
Ensemble	No	RGB-D	97.68
Ensemble	w/ Val	RGB	**98.42**
Ensemble	w/ Val	RGB-D	**98.53**

Table 2:- Performance our ensemble results evaluated on AUTSL test set.

➢ *Evaluation* –

When training our models on the training set, we adopt an early stopping technique based on the validation accuracy to obtain our best models. Then we test our best models on the test set and use the hyperparameters tuned on validation set to obtain our ensemble prediction. To further improve our performance, we finetune our best models on the union of training and validation set. we stop training when the training loss in our finetuning experiment is reduced to the same level as our best models in the training phase. Our predictions with and without finetuning are evaluated and reported in Table 2. Our proposed SAM-SLR approach surpasses the baseline methods significantly.

V. CONCLUSION

In this paper, we propose a new deep learning-based pipeline architecture that efficiently realises real-time automated sign language recognition by combining SSDs, 2DCNNs, 3DCNNs, and LSTMs. The model gave a new hand skeleton feature representation to 3DCNN for richer features after projecting it onto three surfaces. To obtain the trademark features, we also applied pixel-level 3DCNN and heatmap features. The LSTM is given the concatenated output of all 3DCNNs with stacked inputs in order to recognise sign language completely. Additionally, a thorough study of the single view and multiview projections of the 2DCNN and 3DCNN models is provided.

REFERENCES

[1]. Himanshu Gupta, Aniruddh Ramjiwal, Jasmin T. Jose , "Vision Based Approach to Sign Language Recognition" , IEEE , 2018.
[2]. Mahesh Kumar N B, " Conversion of Sign Language into Text" , Springer link , 2018.
[3]. Teak-Wei Chong and Boon-Giin Lee , " American Sign Language Recognition Using Leap Motion Controller with Machine Learning Approach" , mdpi Journal , 2018.
[4]. Lean Karlo S. Tolentino, Ronnie O. Serfa Juan, August C. Thio-ac, Maria Abigail B. Pamahoy, Joni Rose R. Forteza, and Xavier Jet O. Garcia , " Static Sign Language Recognition Using Deep Learning" , International Journal of Machine Learning and Computing, , 2019
[5]. Rim Barioul ,Sameh Fakhfakh Ghribi , Houda Ben Jmaa Derbel ,and Olfa Kanoun, " Four Sensors Bracelet for American Sign Language Recognition based on Wrist Force Myography" , IEEE Xplore , 2020.
[6]. Paul D. Rosero Montalvo, Pamela Gody Trujillo, Edison Flores Bosemedian, Jorge Carrascal Garcia, Santiago otero potosi, Henry Benitez Pereira, " Sign Language Recognition Based on Intelligent Glove Using Machine Learning Techniques" , IEEE ,2020.
[7]. Dongxu Li , Cristian Rodriguez Opazo, Xin Yu, Hongdong Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison " , Computer vision foundation , IEEE Xplore., 2020.
[8]. Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, Hongdong Li , " Transferring Cross-domain Knowledge for Video Sign Language Recognition" , IEEE Xplore , 2020
[9]. Necati Cihan Camgoz¨, Oscar Kollerq, Simon Hadfield and Richard Bowden , " Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation" , IEEE Xplore , 2020.
[10]. Ozge Mercanoglu Sincan, Anil Osman Tur, and Hacer Yalim Keles. Isolated sign language recognition with multi-scale features using LSTM. In 2019 27th Signal Processing and Communications Applications Conference (SIU), pages 1–4. IEEE, 2019.
[11]. Songyao Jiang, Bin Sun , Lichen Wang, Yue Bai, Kunpeng Li and Yun Fu , "Skeleton Aware Multi-modal Sign Language Recognition" , IEEE Xplore , 2021.
[12]. *Ozge Mercanoglu Sincan and Hacer Yalim Keles. AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods. IEEE Access, 8:181340–181355, 2020.*
[13]. John Bush Idoko. "Deep learning based sign language translation system" KSII Transaction on internet and information systems (TIIS). 2020.
[14]. M. E. Al-Ahdal and M. T. Nooritawati, "Review in sign language recognition systems," in 2012 IEEE Symposium on Computers Informatics (ISCI), March 2012, pp. 52–57.
[15]. Cao, Z., Hidalgo, G., Simon, T., Wei, S., & Sheikh, Y. (2017). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields (pp. 7291–7299). Las Vegas, United States: CVPR.
[16]. Dadashzadeh, A., and Tavakoli Targhi, A., and Tahmasbi, M., & Mirmehdi, M. (2018). HGR-Net: A fusion network for hand gesture segmentation and recognition.
[17]. Elboushaki, A., Hannane, R., Afdel, K., & Koutti, L. (2020). MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. Expert Systems with Applications
[18]. Ferreira, P. M., Cardoso, J. S., & Rebelo, A. (2019). On the role of multimodal learning in the recognition of sign language. Multimedia Tools and Applications, 78(8), 10035–10056
[19]. Cao, Zh. And Hidalgo, G. and Simon, T. and Wei, Sh.E. and Sheikh, Y. (2017). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.
[20]. Dadashzadeh, A. and Tavakoli Targhi, A. and Tahmasbi, M. and Mirmehdi, M. (2018). HGR-Net: A Fusion Network for Hand Gesture Segmentation and Recognition.
[21]. Preetham, C.; Ramakrishnan, G.; Kumur, S.; Tamse, A.; Krishnapura, H. Hand Talk-Implemented of a Gesture Recognition Glove.
[22]. Guo, H. and Wang, G. and Chen, X. and Zhang, C. and Qiao, F. and Yang, H. (2017). Region Ensemble Network: Improving Convolutional Network for Hand Pose Estimation.
[23]. Wu, J.; Sun, L.; Jafari, R. A Wearable System for Recognizing American Sign Language in Real-Time Using IMU and Surface EMG Sensors. IEEE , 2016
[24]. *Cheok, M.J.; Omar, Z.; Jaward, M.H. A review of hand gestures and sign language recognition techniques.*