# International Journal Of Scientific & Technology Research, Vol 1, Issue 1

**P. Praveen Kumar, P. V. G. D. Prasad Reddy and P. Srinivasa Rao**

over the traditional and other deep Networks. **Abstract:** Sign language recognition is a potentially difficult process using RGB video data in a deep Learning framework due to data inconsistencies during capturing. To improve the recognition accuracy of SLR systems as a whole, researchers applied variable inputs such as depth and motion frames in addition to RGB video. However, skeleton data was little used along with RGB and depth inputs for SLR, although its been in operation for human action recognition. In this work, we explore the possibility and challenges in using RGB, depth and skeletal sign language data for recognition. The recognition is accomplished using a three stream Convolution Neural Network with score fusion on spatial data. Each CNN stream is made up of 8 convolutional layers, two fully connected and a soft max layer. Finally, score fusion is performed to extract the labels during testing. The proposed framework has shown an improvement in recognition accuracy

**Index Terms:** Sign language recognition, Convolution Neural Network, RGB, Depth, Skeletal data, Deep learning, score fusion.

## 1 INTRODUCTION

Sign language recognition is still considered challenging and complex due to involvement of random human hand movements. These movements are very inconsistent with respect to the degrees of freedom a person has on the hands during the signing process. Sign language is the language of the hearing impaired that uses fingers and hand movements in reference to head, face or body. Past research shows, 2D video-based sign language recognition systems are widely considered due to low cost camera sensors. However, the processing algorithms were complex and computationally intensive to be considered for development of complete sign language recognition system. The major difficulty in sign language recognition compared to speech recognition is to recognize simultaneously different communication attributes of a signer such as hands and head movement, facial expressions and body pose [1]. All these attributes must be considered simultaneously for a good recognition system. The second major problem faced by designers of sign language recognition system is tracking the signer in the clutter of other information available in the video. This is addressed by many researchers as signing space [2]. A sign language space can be created with entities such as humans or objects stored in it around a 3D body centered space of the signer. These entities are executed at a certain location and later referenced by pointing to the space. A major challenge faced by researchers to define a model for spatial information containing these entities created during the sign language dialogue.


- *P. Praveen Kumar is Department of Computer Science and Systems Engineering, College of Engineering Andhra University, Visakhapatnam, India. E-mail: pkpinjala.auce@gmail.com*
- *P. V. G. D. Prasad Reddy is Department of Computer Science and Systems*

*Engineering, College of Engineering Andhra University, Visakhapatnam, India.*

- *P. Srinivasa Rao is Department of Computer Science and Systems Engineering, College of Engineering Andhra University, Visakhapatnam, India.*


Additional difficulties arise in the form of background in which signer is located. Most of the methods developed so far use simple backgrounds in controlled set-up such as dark backgrounds, special hardware like data gloves, restricted sets of actions, restricted number of signers, resulting different problems in sign language feature extraction [3].To make the system adaptive to different signers as in case of automatic speech recognition systems which are able to cope with different dialects. Speaker adaptation techniques known for speech recognition can be used to make the system more robust. While recognizing signs of few signers only the intrapersonal variabilities in appearance and velocity of their hands needed to be modeled. As the number of signers increases, the quantity and diversity of variabilities is extremely increased. In continuous speech recognition as well as continuous sign language recognition, co-articulation effects have to be considered. Currently, machine learning is rapidly transforming speech and image processing methodologies to get higher performances. To this effect, Sign language is not far behind. The present deep Networks use 2D RGB video frames as well as Kinetic based RGB -D data as inputs. In this work, we propose to fill the gap induced by the previously established deep learning algorithms with a three-stream convolutional neural network with RGB, depth and skeletal inputs. The idea of using three different input modalities is to improve recognition by mutually influencing where there is difficulty in extracting features from raw data. In this work only spatial features are exploited for SLR, even though motion features had successfully applied alongside spatial features for action recognition tasks. The following paper is organized as follows. Section 2 gives literature and methodology in section 3. Section 4 and 5 discuss results and discussion respectively.

## 2 LITERATURE REVIEW

The main challenge in any sign language recognition system is to find a computer model that can fully capture the

large scale vocabularies of the language. In this regard, the video based sign language is the most popular mode of designing approaches for machine interpretation, that comes with a large set of challenges for the researchers [4]. Brightness, contrast, background, blurring and occlusions in a sign video, greatly reduces the prediction capability of the machine interpreter algorithms. The 2D video based SLR approaches used so far are based on skin color and saliency maps [5], which functionally localize the human signer. Segmentation of hands of the signer is approached in many ways such as edge, morphological, frame subtraction, gaussian mixture models (GMM) and graph cuts [6]. However, all these models suffer due to complexity in acquiring signs in real time. Researchers also induced tracking features, along with shape and orientation features in sign videos captured with complex video backgrounds [7] for better recognition performance. Like speech recognition, Hidden Markova Models (HMM) are a popular choice for classification of sign language sparse features [8] with its double stochastic model. Moreover, the researchers showed that Artificial Neural Networks (ANNs), Fuzzy systems, Bayes classifier and Support Vector Machines (SVM) are a nice alternative to HMMs [9]. The last decade has seen a transformation in the 2D video processing of sign language with the availability of low cost multi modal sensors. Microsoft Kinect with their powerful SDK's has become a primary choice for multi modal data acquisition. It provides RGB color images, depth maps and skeleton data of the human body at a frame rate of 30fps. Hence, brightness, contrast, blurring and background problems that are related to 2D sensors are handled efficiently to an extent with binocular vision in Kinect. Even sign language is no exception, where better processing models are being proposed to handle both RGB and depth data [10] efficiently. Recent advances in hardware accelerators has given rise to machine learning algorithms such as Convolutional neural networks (CNNs) [11] with Keras and Tensor flow modules, the computer vision based classification flourished multifold. The CNNs successes with RGB video action data [12] with single stream deep CNNs or LSTMs [13] was noticeable. It then migrated to RGB –D [14] and skeleton based multi modal CNN architectures [15]. Many combinations such as CNN– LSTMs [16] and CNNs –RNN (Recurrent Neural Nets) [17] has been in operation for sign language recognition with RGB -D data. Recent 3D technology Advancements have made researchers to improve sign language recognition tremendously using deep Networks [18]. The next section gives the proposed methodology for RGB D Indian sign language recognition framework.

## 3 PROPOSED METHODOLOGY

Unlike the previous works, the proposed method uses three different modalities as inputs to a three stream convolutional neural network. In this section, we describe the CNN architecture with training and testing procedures that were conducted during the experimentation.

### 3.1 3 Stream CNN Architecture

To improve the recognition of sign language machine translator over the past methods, we propose to increase the amount of data used for training. Consequently, all the recorded video data from Kinect was used as inputs. The data consisted of RGB video, depth and skeletal information of the sign. This multi modal information has been used in the past with only two modalities at time for sign language recognition. However, in this work we are able to use all modalities captured from the Kinect for recognition. The database is made from 200 class labels of Indian sign language with 5 subjects in 5 orientations. A total of 5000 sign videos in three different modalities were created for experimentation. This input is fed to a three stream convolutional neural network with each stream absorbing the three modalities. The architecture of the proposed 3 stream CNN used for Sign language recognition is shown in figure 1. Each of the three streams are made up of similar layers. The CNN has 8 convolutional layers and in between them max pooling by 2 layers ending with two fully connected leading to a soft max decision layer. The architecture is inspired from VGG - 16 CNN model, but with half the depth. In contrast our proposed CNN architecture is shallow enough to be computationally efficient and faster to compensate for the extra streams during training and testing. The entire architecture is constructed using tensorflow as frontend and keras as backend in python. During experimentation, the networks parameters and hyper parameters are kept constant across all datasets. We used sign language data set from BVCSL3D and remaining action datasets MSRDailyAction3D, UTKinect, G3D.

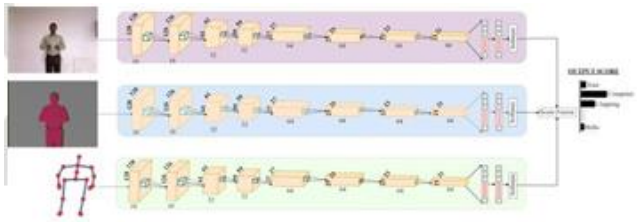

***Fig. 1****. Proposed CNN architecture*

### 3.2 Training the 3 stream CNN

The goal of training is to optimize the multinomial logistic regression objective function by means of mini-batch gradient descent with momentum of 0.5. The penalization multiplier is set to 0.002 to regularize weight decay during training for a video frame size of 128. The RGB video frames, the depth maps and the skeletal frames are resized to 128×128 using down sampling by 2. The max pooling in the RGB stream is set to 0.5 for all 8 conv net layers. However, in depth stream the max pooling was 0.6 and in skeletal stream it was 0.7 for all convolutional layers. The initial learning rate is set to 0.01 and decreased by a factor of 10, when validation accuracy became constant. For a 4800-RGB D sign dataset during training, the learning rate was decreased three times from 0.01 to 0.001 to 0.0001 to 0.00001 when the loss appeared to be constant and the training stopped after 156 epochs. On MSRDailyAction, UTKinect and G3D action datasets the learning rate was decreased twice from 0.01 to 0.001 to 0.0001

during the entire training course. The training was stopped within 100 epochs for action datasets due to less data samples used for training when compared to sign language. Moreover, all the nets on all datasets required less epochs when compared to previous approaches, due to a medium scale frames being used during training. Weights in each layer are randomly initialized with a gaussian distribution function with zero mean and 0.01 variance for all layers. Intermediate validation showed overfitting in the dense layers during training for all the datasets. This is handle by inducing dropout of 0.5 before the dese layer, which improved the validation score. All three streams are trained with same set of hyper parameter values in the initial training phases. Finally, the outputs of the SoftMax layers in the three streams are fused using average the similarity scores for making a decision on the output class labels.

### 3.3 Sign (or Action) Testing

The proposed three stream CNN when trained uses a preallocated batch of RGB D sign videos for testing which classifies the signs (or actions) into corresponding labels. The RGB D video samples in all datasets are video frames of sizes 128×128. The resulting output is a class score vector of probabilities from the SoftMax layer. The probabilities from the two streams are averagely fused to generate a single class score vector for the input query sign (or action) RGB D data. The max probability score in the output vector is the detected class label for the given test sign. We used multiple percentages based training model, where different sample sizes are set aside for testing from the entire dataset. The training loss was monitored during the entire training cycle and was kept as linearly as possible for faster weight updates. Except for UTKinect dataset the loss was on the higher side when compared to the remaining datasets because of the data inconsistency. Lowest training loss was registered for BVCSL3D sign language data due to higher data consistency.

## 4 RESULTS AND DISCUSSIONS

A set of four 3 stream CNNs were created for experimentation on one sign language and three action datasets. Figures 2, 3, 4 and 5 will help visualize the datasets. All these datasets were recorded using Microsoft Kinect producing RGB, depth and skeletal video frames of 3D sign language and human action. Each video is trimmed to 94 frames per sign in all modalities for sign language to maintain uniformity over the three streams. However, for action datasets trimming of videos is 56, 65 and 87 frames for MSRDailyActivity, UTKinect and G3D respectively. The hardware used in all experiments has been a 8GB GPU from NVIDIA with a 16GB memory running on 2.4GHz Intel core processor. The hyper parameters were kept constant throughout the training sessions. Performance measures such as recognition accuracy, precision and recall were calculated across networks on all datasets for checking the robustness of the proposed CNNs against the state of the arts. The proposed method is being tested on subjects, orientations and network type used.

***Fig. 2**. BVCSL3D RGB D sign dataset, (a) Sign "Balloon" in RGB and its*
**(b) Depth, (c) RGB "Circus" and (d) its depth, (e) RGB "Drama" and its (f) depth, (g) RGB "Computer" and its (h) depth map.**

***Fig. 3.** MSRDailyAction3D RGB Video Frames, (a) "Talking on Cell", (b) "Read Book"*

***Fig. 4.*** *UT Kinect action datasets. (a) "Walk", (b) "Clap", (c) "Wave", (d) "Pick"*

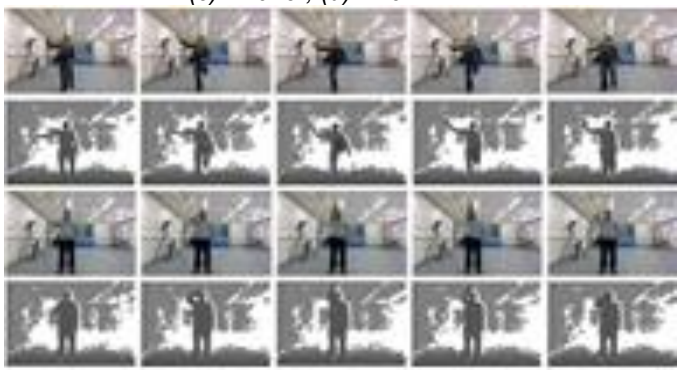

***Fig. 5.*** *Action examples from G3D gaming action set. (a) RGB Video frames of "Kicking with left leg", (b) Depth frames from (a), (c) RGB Video frames "Punching with right hand", (d) Corresponding depth frames from (c).*

**4.1Subject based testing the proposed CNN**

Here the proposed 3 streams are fed with different proportions of training and testing data from the same and cross subject sets. Table 1 gives the average recognition rates calculated on the training set. The table provides results for same subject testing and cross subject testing in various data proportions. The overall recognition rates of the proposed 3 stream CNN are far better than the other methods from literature. The superiority of the proposed method is attributed primarily to the use of 3rd stream, which provided additional resources for recognition. The proposed 3 stream CNN architecture shows the power of complimenting the loss in one stream by the other stream.

***TABLE 1***

*RECOGNITION RATES OF OUR PROPOSED MODEL AGAINST THE STATE-OF-THE-ART MULTI STREAM MODELS ON DIFFERENT TRAINING AND TESTING DATA SPLITS.*

| *Testing and Training data splits* | | *Recognition Rate (%)* | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *50% data for Training; 50% data for Testing* | | | | | *40% data for Training; 60% data for Testing* | | | | | *70% data for Training; 30% data for Testing* | | | | |
| *Dataset* | | *Gao[20]* | *Liu[21]* | *Wang[19]* | *CoT4CN N[22]* | *Proposed* | *Gao[20]* | *Liu[21]* | *Wang[19]* | *CoT4CN N[22]* | *proposed* | *Gao[20]* | *Liu[21]* | *Wang[19]* | *CoT4CN N[22]* | *proposed* |
| *Same Subject* | *BVCSL3D* | *71.5* | *82.3* | *84.6* | *87.9* | *89.24* | *60.1* | *70.8* | *73.2* | *76.2* | *77.16* | *73.6* | *84.4* | *86.5* | *90.2* | *92.34* |
| | *MSRDaily Action3D* | *84.9* | *85.8* | *89.1* | *91.4* | *92.34* | *73.5* | *74.3* | *77.7* | *79.7* | *79.34* | *87.3* | *87.9* | *91.5* | *93.7* | *93.96* |
| | *UT Kinect* | *82.8* | *85.3* | *88.6* | *90.9* | *91.17* | *71.4* | *73.8* | *77.2* | *79.2* | *80.15* | *84.9* | *87.4* | *90.5* | *93.2* | *94.44* |
| | *G3D* | *86.8* | *89.2* | *92.5* | *94.7* | *94.96* | *75.4* | *77.7* | *81.1* | *83.2* | *83.93* | *88.9* | *91.3* | *94.4* | *95.3* | *95.93* |
| *Cross Subject* | *BVCSL3D* | *73.9* | *78.2* | *82.5* | *86.8* | *87.21* | *62.5* | *66.7* | *71.1* | *75.1* | *76.54* | *76.4* | *80.3* | *84.4* | *89.1* | *90.92* |
| | *MSRDaily Action3D* | *84.5* | *84.5* | *87.9* | *90.1* | *91.34* | *73.1* | *73.1* | *76.5* | *78.4* | *79.34* | *86.6* | *86.6* | *89.8* | *92.4* | *94.57* |
| | *UT Kinect* | *83.1* | *83.9* | *87.2* | *89.5* | *90.68* | *71.7* | *72.4* | *75.8* | *77.8* | *78.29* | *85.2* | *86.1* | *89.1* | *91.8* | *92.08* |
| | *G3D* | *84.9* | *87.4* | *90.7* | *93.0* | *93.24* | *73.5* | *75.9* | *79.3* | *81.3* | *82.69* | *87.6* | *89.5* | *92.6* | *95.3* | *95.93* |

The table 1 also gives an insight into the functioning of CNNs on the lines of data availability for training. It shows that larger data training will help in increased recognition rates over the

nets using lesser training data. However, it is also seen that multi stream CNNs can perform better than single stream models in situations where data is sparsely available. Similarity, cross subject validation showed higher recognition rates on the proposed CNN when compared to other models.

### 4.2 Orientations testing

In real situations, the subject is never going to be in the exactly same orientation as in the datasets used for training. Hence an investigation into this aspect is necessary creation to judge the robustness of a training algorithm like CNN. In this regard we have trained our proposed 3 stream CNNs with data in one orientation and tested with subject data in a different orientation. The performance measures here are precision and recall. Table 2 shows averaged values for various datasets of sign language and action. The results from table point to the use of multiple streams that provided better performance over the other methods.

***TABLE 2***

*PRECISION AND RECALL AVERAGES OF OUR PROPOSED MODEL AGAINST THE NETS PROPOSED IN [19],[20],[21] AND [22].*

| Dataset | | Precision (%) | | | | | Recall (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Gao[20] | Liu[21] | Wang [19] | CoT4CNN [22] | Proposed | Gao[20] | Liu[21] | Wang [19] | CoT4CNN [22] | Proposed |
| Same Subject with multiple orientations | BVCSL3D | 70.26 | 84.84 | 87.67 | 88.27 | 89.75 | 75.81 | 86.65 | 91.44 | 92.94 | 93.24 |
| | MSRDaily Action3D | 82.54 | 86.39 | 89.22 | 90.26 | 91.34 | 88.66 | 87.42 | 92.14 | 93.25 | 94.65 |
| | UT Kinect | 81.31 | 84.15 | 86.98 | 88.13 | 88.93 | 87.24 | 87.73 | 92.71 | 93.90 | 94.32 |
| | G3D | 84.03 | 86.93 | 89.76 | 90.83 | 91.57 | 88.79 | 88.33 | 93.60 | 94.66 | 95.48 |
| Cross Subject with multiple orientations | BVCSL3D | 74.24 | 83.09 | 85.92 | 86.87 | 86.94 | 82.57 | 85.25 | 90.01 | 91.56 | 92.67 |
| | MSRDaily Ac tion3D | 83.92 | 82.78 | 85.61 | 86.98 | 87.84 | 86.96 | 85.59 | 90.33 | 91.98 | 92.82 |
| | UT Kinect | 82.57 | 81.42 | 84.25 | 85.23 | 86.31 | 88.54 | 88.41 | 92.25 | 93.33 | 94.28 |
| | G3D | 84.42 | 84.98 | 87.81 | 88.68 | 89.72 | 89.86 | 89.44 | 93.45 | 94.51 | 95.67 |

### 4.3 Network comparison

This section shows results on only sign language data that is being supplied to the network while testing. The results of our study are presented in table 3. It shows of using only RGB data for testing without the use of other modal inputs and so on. We found that in multi stream CNNs sparse testing is a potential possibility as there is no provision of using Kinect everywhere in real time. This calls for the use of RGB videos and its derivatives for testing even though the CNN was trained with depth and skeletal data associated with RGB. Table shows higher percentage of recognition when motion data is also made available for testing along with spatial RGB video frames.

***TABLE 3***
*IMAGE DATA VALIDITY ON ALEXNET, VGG16, GOOGLENET, RESNET AND THE PROPOSED CNN.*

| *Testing Inputs* | *Networks* | | | | | |
|---|---|---|---|---|---|---|
| | *AlexNet* | *VGG16* | *GoogLeNet* | *ResNet* | *onnived Resnet* | *Proposed* |
| *RGB Only* | *80.27* | *83.79* | *84.32* | *84.02* | *84.96* | *85.32* |
| *Depth Only* | *83.64* | *85.13* | *86.49* | *85.77* | *85.30* | *86.27* |
| *RGB+Depth* | *83.99* | *86.05* | *87.44* | *85.98* | *89.45* | *90.58* |
| *RGB+Depth+ Skeleton* | *86.74* | *87.56* | *87.94* | *87.09* | *89.69* | *91.38* |

### 4.4 Individual 3D sign testing

Finally to check the exact amount of recognition obtained by individual 3D signs, we draw confusion matrices in same and cross subject testing in fig 6 and 7. These figures show how some signs missed in their entirety due a strong similarity between them. A more efficient and powerful machine interfaces are required to improve on such situations where two signs have a very strong connection.

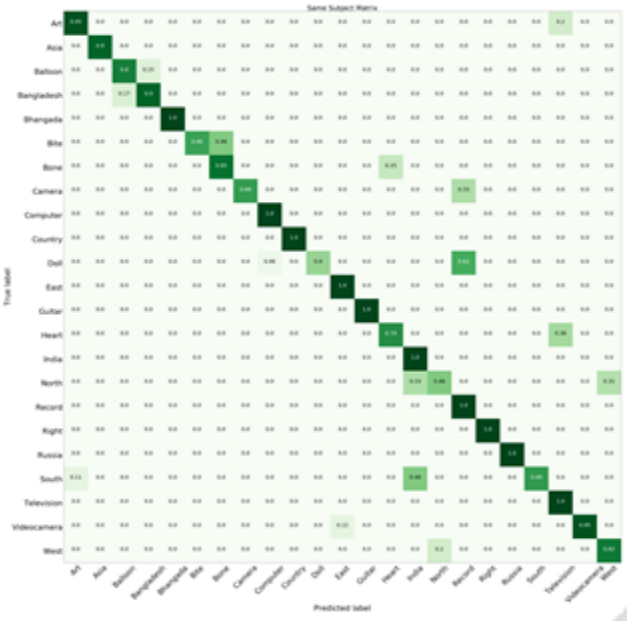

***Fig.6****. Confusion matrix for same subject training and testing.*

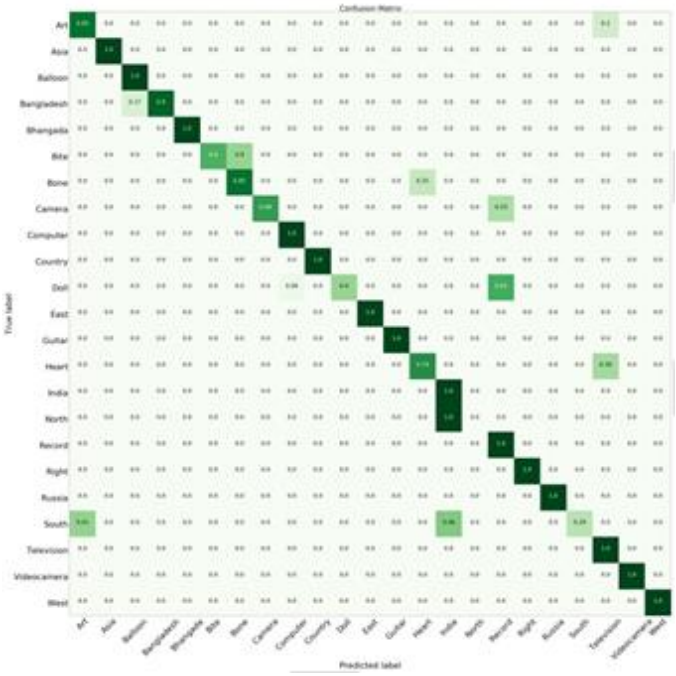

***Fig. 7.*** *Confusion matrix for cross subject testing.*

## 5 CONCLUSION

This work has shown the potential to develop a 3D RGB D based sign language recognition system. To this effect, we designed a 3 stream convolutional neural network architecture with 8 convolutions and 2 fully connected layers. The proposed 3 stream CNN is trained and tested on RGB D sign language data BVCSL3D and three other popular RGB D action datasets. We found that the proposed CNN has the ability to recognize more signs in Indian sign language than the similar methods in the category. This work also shows cross subject and sparse datasets can be made useful with our multi stream CNN model. The work found that the overall recognition rates were higher than the previous averages on the same datasets.

## REFERENCES

[1] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of American sign language. Computer Vision & Image Understanding, 81(3):358–384, March 2001.

[2] G. Yao, H. Yao, X. Liu, and F. Jiang, “Real time large vocabulary continuous sign language recognition based on op/viterbi algorithm,” in Intl.

Conf. Pattern Recognition, Hong Kong, Aug. 2006, vol. 3, pp. 312–315.

[3] Kishore, P.V.V. and Kumar, P.R., 2012. Segment, track, extract, recognize and convert sign language videos to voice/text. (IJACSA) International Journal of Advanced Computer Science and Applications,

[4] Kishore, P.V.V., Prasad, M.V.D., Prasad, C.R. and Rahul, R., 2015, January. 4-Camera model for sign language recognition using elliptical fourier descriptors and ANN. In 2015 International Conference on Signal Processing and Communication Engineering Systems (pp. 34-38). IEEE.

[5] Zamani, M. and Kanan, H.R., 2014, October. Saliency based alphabet and numbers of American sign language recognition using linear feature extraction. In 2014 4th International Conference on Computer and Knowledge Engineering (ICCKE) (pp. 398-403). IEEE.

[6] Rautaray, S.S. and Agrawal, A., 2015. Vision based hand gesture recognition for human computer interaction: a survey. Artificial intelligence review, 43(1), pp.1-54.

[7] Pansare, J.R., Gawande, S.H. and Ingle, M., 2012. Real-time static hand gesture recognition for American Sign Language (ASL) in complex background. Journal of Signal and Information Processing, 3(03), p.364.

[8] Al-Rousan, M., Assaleh, K. and Tala'a, A., 2009. Video-based signer- independent Arabic sign language recognition using hidden Markov models. Applied Soft Computing, 9(3), pp.990-999.

[9] Naglot, D. and Kulkarni, M., 2016, August. ANN based Indian Sign Language numerals recognition using the leap motion controller. In 2016 International Conference on Inventive Computation Technologies (ICICT) (Vol. 2, pp. 1-6). IEEE.

[10] Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H. and Presti, P., 2011, November. American sign language recognition with the kinect. In Proceedings of the 13th international conference on multimodal interfaces (pp. 279-286). ACM.

[11] Hou, R., Chen, C. and Shah, M., 2017. Tube convolutional neural network (T-CNN) for action detection in videos. In Proceedings of the IEEE International Conference on Computer Vision (pp. 5822- 5831).

[12] Xu, Z., Yang, Y. and Hauptmann, A.G., 2015. A discriminative CNN video representation for event detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1798- 1807).

[13] Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M. and Baik, S.W., 2017. Action recognition in video sequences using deep bi-directional

LSTM with CNN features. IEEE Access, 6, pp.1155-1166.

[1] Zhu, G., Zhang, L., Shen, P. and Song, J., 2017. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. IEEE Access, 5, pp.4517-4524.

[2] Chai, X., Liu, Z., Yin, F., Liu, Z. and Chen, X., 2016, December. Two streams recurrent neural networks for large-scale continuous gesture recognition. In 2016 23rd International Conference on Pattern Recognition (ICPR) (pp. 31-36). IEEE.

[3] Tsironi, E., Barros, P., Weber, C. and Wermter, S., 2017. An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition. Neurocomputing, 268, pp.76-86.

[4] Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C. and Xu, W., 2016. Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2285-2294).

[5] Kumar, E.K., Kishore, P.V.V., Sastry, A.S.C.S., Kumar, M.T.K. and Kumar, D.A., 2018. Training CNNs for 3-d sign language recognition with color texture coded joint angular displacement maps. IEEE Signal Processing Letters, 25(5), pp.645-649.

[6] Wang, Pichao, Wanqing Li, Zhimin Gao, Jing Zhang, Chang Tang, and Philip O. Ogunbona. "Action recognition from depth maps using deep convolutional neural networks." IEEE Transactions on Human- Machine Systems 46, no. 4 (2016): 498-509.

[7] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, “Two-Stream 3-D convNet Fusion for Action Recognition in Videos With Arbitrary Size and Length,” IEEE Transactions on Multimedia, vol. 20, no. 3, pp. 634–644, Mar. 2018.

[8] D. Liu, Y. Wang, and J. Kato, “Evaluation of Triple-Stream Convolutional Networks for Action Recognition,” 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Nov. 2017.

[9] Ravi, Sunitha, Maloji Suman, P. V. V. Kishore, Kiran Kumar, Teja Kiran Kumar and Anil Kumar. "Multi modal spatio temporal co- trained CNNs with single modal testing on RGB–D based sign language gesture recognition." Journal of Computer Languages 52 (2019): 88-102.