DSLR-Net a Depth Based Sign Language Recognition Using Two Stream Convents

P.V.V.Kishore, K.B.N.S.K.Chaitanya, G.S.S.Shravani, M.Teja Kiran Kumar, E.Kiran Kumar, D.Anil Kumar

Abstract*: Sign Language is the only medium of communication among the hearing impaired and mute people. Sign language includes hand movements along with facial expressions to convey their meaningful thoughts. To decrease the communication barriers with the normal people, we propose a novel framework based on two stream convolutional neural network (CNN). Which is a powerful tool with a combination of depth data obtained from low cost Kinect sensors along with RGB data obtained from Video Camera other than Kinect RGB data to ensure better pixels per inch (PPI) for images and latest deep learning algorithm Two stream convolutional neural network (CNN). RGB and Depth data are given as input to both the streams separately in training and for validating the network only RGB data is required to predict the class labels of the sign. Here the features are shared by both the RGB and Depth streams to recover the missing features by convolution operation in CNN. Which results better classification rates by decreasing the training epochs and can be used for real time interpreters for better performance. To validate our method, we created our own dataset BVCSL3D and the publicly available datasets NTU RGB-D, MSR Daily Activity 3D and UT Kinect. To claim the novelty of our model we tested our data with other state-of-the-art models and recognition rates are investigated.

***Index Terms*: Sign language, Depth, RGB, Convolutional Neural networks.**

I. INTRODUCTION

From past few years Sign language recognition is being exercised in which, hand movements are estimated by using segmentation and hand tracking algorithm by using a continuous sign language video data [1]. This method achieves better recognition rates only in non-complex backgrounds. However, in complex backgrounds, the recognition is a challenging task.

Depth data has gained more popularity in recent research as it is captured from low cost Microsoft Kinect [2]. Depth information gives the structural representation of signers in a 3D domain as it can be differentiated by the backgrounds. It has limitation that only depth data must be given for predicting a sign which is very hard to operate in real time.

Local spatio temporal (LST) method is another technique which also gained more popularity because it localizes salient body shapes and recognizes the actions in any complex backgrounds [3]. Thus, LST features mark out the human pose from RGB video data by some pre-selected interest points. Despite of their success LST feature have limitation in the form of inconsistency in video data. Convolutional neural networks [4] also called as ConvNets are popularly used technique for Human Action Recognition (HAR) by most of the researchers around the globe. The major component of this work mainly focusses on the two stream ConvNets and its Deep analysis for Sign Language Recognition. These two stream convnets will have two inputs where, depth data captured from Microsoft Kinect is given to one stream and RGB data digital video camera is inputted to another stream. ConvNets basically extracts the features from low to high levels as the data flows from lower order convolutional layers to higher order convolutional layers. Our ConvNet model consists of six layers and each convolutional layer is followed by ReLU (Rectified Linear Unit) as a activation function and each of two convolutional layers followed by one maxpooling and like in two streams. Output of these two streams are fused to share the features among them such that missed information in one stream can be overcome by another stream to give the better recognition rates. In general, RGB data consists of 80% background pixels density and only 20% is of signer's pixel density. This greatly reduces the probability of finding the signer and hence unable to predict the sign of the signer. To overcome this problem, we are giving depth data as input to one of the stream in addition to RGB data stream of the convnet. In this way we can extract the signer information from the background based on structural information obtained from the depth data,such that our proposed CNN model can recognize the sign based on the signer hand movements with the greater recognition rates. The proposed architecture mainly solves two problems. 1) It can easily locate the signer and can predict the hand gesture movements. 2) Only RGB data is enough to predict signs in real time, i.e., training with RGB and Depth data and testing with only RGB data.

The rest of the paper is organized as follows. Section 2 gives the brief details of our proposed model DSLR-Net for Sign language and human action recognition with other different models. In next section, the detailed explanation of the proposed architecture is given. Finally, section 4 gives the results and analysis of our model.

II. LITERATURE REVIEW

Sign language recognition (SLR) is being carried out in 1D,2D and 3D. 1D is characterized by signals from motion sensing gloves. 2D is characterized by Images or video frames captured by a digital camera. 3D is characterized by depth with addition to RGB to form RGB-D data recorded by Microsoft Kinect.

Revised Manuscript Received on June 05, 2019

P.V.V.Kishore, Professor, Department of ECE, Koneru Lakshamaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, A.P, India.

K.B.N.S.K.Chaitanya, G.S.S.Shravani, M.Teja Kiran Kumar, E.Kiran Kumar, D.Anil Kumar, Department of ECE, Koneru Lakshamaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, A.P, India.

1D SLR data mainly collected by the wireless glove it generates radio frequencies based on the glove movements and these frequencies are transmitted to microcontroller for the recognition [5]. 1D SLR is fast as it deals only with the glove-based hand movements. Beside hand movements, SLR includes facial expressions and some head movements as it is developed as a Visual based Sign Language [6]. 2D SLR produces more recognition rates when compared to 1D SLR.
2D data explores all the visual based sign language approach in terms of both speed and recognition accuracies. In [7] hand shapes are captured, and these captured image sequences are converted to 1D sequence distance vectors by extracting hand shapes and scores an accuracy of 33.8% over 24 Japanese sign by using Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM). HMM is a most commonly used classifier for 2D SLR [8]. HMM's are faster if and only if the sign in 2D video is accurately represented. However, 2D SLR is unable to achieve the better results under low brightness, occlusions and blurring effect by the camera. So far, 2D models are used for SLR by extracting hand shape and hand tracking data as features [9]. Later researchers developed the 2D SLR models these models got better results under complex backgrounds [10], by using Multicamera model and from 2D SLR based mobile interpreters. In [11] hand shapes are extracted by contours using k-curvature algorithm and then Dynamic Time Warping (DTW) algorithm is used for recognizing the hand gestures by reference hand gestures as ground truth.
Later Deep learning algorithms have gained more popularity due to increase in recognition rates. In [12] deep learning based algorithm Convolutional neural network were proposed to recognize the signs in a video. Even though these models also face the same problems as low brightness, occlusions and blurring effect. To minimize these problems, we are moving in to 3D depth data as it contains structural data of the signers. Apart from 2D there are several approaches to 3D sign language recognition by acquiring data from 3D sensors such as TOF (time-of-flight) cameras or Microsoft Kinect sensors. These sensors were developed to interact with games in real time i.e., for tracking of full-body movements and gestures [13]. Many researchers have successfully developed models like interactive displays , robotic assistance, hand gesture recognition [14], sign language recognition [15], etc., by using these sensors.
3D depth-based sign language recognition has gained more popularity due to improved interactivity, user comfort and highly accurate in recognition tasks than 2D based approaches. Sign language recognition from depth data gives good result but when in real time providing a depth sensor for recognition is a major challenge for the researchers other than RGB data. To minimize these problems, we proposed an architecture named DSLR-Net which shares a information form depth to RGB data. The data comprises of 3D depth frames which captured using Microsoft Kinect and RGB video frames captured using Video Camera by setting frame rate to 30 frames per second(fps) to maintain the capture sync between them. The main advantage of using the both RGB and depth gives both 2D and 3D information so that architecture shares the weights of both information for final classification as 2D gives the visual information and 3D gives the structural information of the signers.

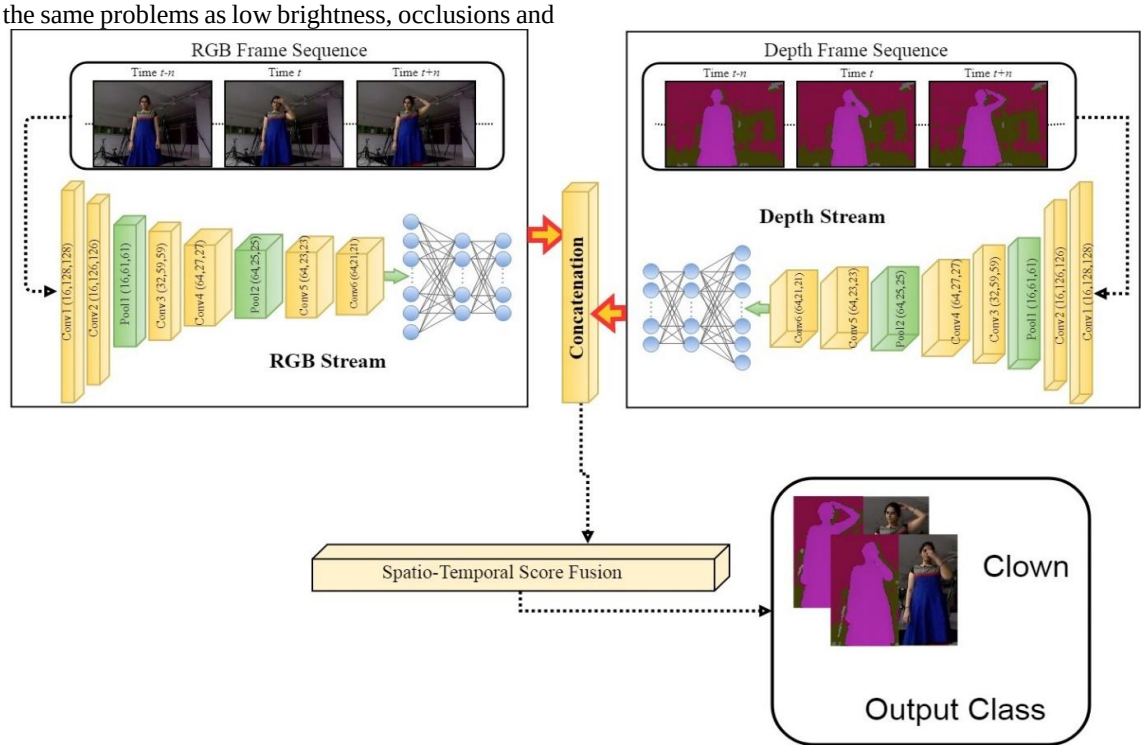

Fig.1 DSLR-Net Architecture

III. PROPOSED APPROACH

In this section we described about the creation of data from Kinect and Video camera. Secondly, we evaluate the data with the proposed model and claiming the novelty of our model by evaluating with the publicly available dataset.

A. Dataset

BVCSL3D is a self-created dataset comprises of 200 sign classes from Indian sign language. Each sign class is captured with Kinect and Video camera both are set to capture @30fps. Each sign class is performed by ten signers, 15 times with multiple views, scales and hand speeds which gives 150 video signs per class. The entire dataset is having 30000 videos in RGB, depth with 3000000 frames in each mode.

Fig 2. BVCSL3D sample data (a) RGB (b) Depth sign data of subject 3, (c) RGB (d) Depth sign data of subject 4.

To validate DSLR-Net, three benchmark datasets namely MSRDailyActivity3D [16], UT Kinect [17] and NTU RGB-D [18] were used for training and testing. MSRDailyActivity3D is constructed from 20 actions, 10 subjects and 567 action videos in RGB-D formats. UT Kinect has 10 actions, 10 subjects and 200 action videos. NTU RGB-D is a gaming action dataset with 16 actions, 16 subjects and 1280 videos. Inspired from these action videos, we developed our own dataset (BVCSL3D) of actions by adapting all the three dataset classes.

Fig 3. BVCSL3D sample data of subject 1 and subject 2 performing signs with different variations

B. DSLR-Net

DSLR-Net is shown in Fig.1 and is inspired by traditional CNN architecture. DSLR-Net is having two streams on which one stream is with RGB and another stream is with Depth as the inputs. Each RGB and Depth stream consists of six convolution layers followed by the Rectified Linear Unit (ReLU) as activation function and pooling layers. Finally, these features are given to two Dense layers among them first dense layer is followed by ReLU and other with Softmax activation layer which gives the output probability. Now the outputs are concatenated such that weights of both the streams will be shared. We validate our model against the regular multi stream CNN architectures with score fusion on two [19][20] and three stream CNNs [21][22].

The networks weights and bias parameters are set arbitrarily utilizing a zero-mean Gaussian distribution function with difference 0.01 toward the beginning of each preparation stage on all datasets. The system learns by updating weights and backpropagation gradient descent algorithm is used for bias. ReLU in convolutional layers and SoftMax in dense layers are hyperparameter activations connected in our CNN.

C. Learning

Our DSLR-Net trains on BVCSL3D Indian sign language dataset, MSRDailyActivity3D, UT Kinect and NTU RGB-D. we set the learning rate to 0.01 initially and then decreased by 0.1 for every 31k iterations for BVCSL3D. and 0.5 initially and decreased by 0.1 for every 41k iterations for remaining datasets MSRDailyActivity3D, UT Kinect and NTU RGB-D. The whole program for DSLR-Net has done in Python 3.7 by using libraries Keras and Tensorflow on a HPC (High Performance Computer) available in our university. HPC consists of two NVIDIA Tesla K20 graphics card with 6 nodes and Training takes almost 28 hours. We used stochastic gradient descendant (SGD) as optimizer and its expression is described in eqn.1 and categorical cross entropy (CCE) is a loss function in SGD the expression is shown in eqn.2.

$$\alpha_{sgd} = \alpha_{sgd} - \eta.\nabla_{\theta} J(\alpha_{sgd}, x, y) \quad (1)$$

Where α_{sgd} is the parameter of model vector, η is the learning rate and $\nabla_{\theta} J(\alpha_{sgd}, x_i, y_i)$ is the gradient parameter of x input data to y output class.

$$CCE = -\frac{1}{N}\sum_{j=0}^{M}(y_j.\log(y_j) + (1-y_j).\log(1-y_j)) \quad (2)$$

Where y_j is the predicted label and y_j is the Actual label and CCE finds the loss based on these two weight functions.

IV. RESULTS AND ANALYSIS

In this section, we evaluate our DSLR-Net with different datasets namely BVCSL3D Indian sign language dataset, MSRDailyActivity3D, UT Kinect and NTU RGB-D. Network parameters will be investigated during training and the detailed explanation will be shown. We also compare our proposed architecture with state-of-art model and performance metrics will be calculated.

Fig 4. BVCSL3D sample data of all 10 subject signs with different complex backgrounds

DSLR-Net is trained with BVCSL3D dataset which consists of 10 subjects with different views, scales and hand speeds these sample variations are shown in fig 4. These subjects are captured under different background variations and with different lighting conditions.

Fig 5 shows the performance metrics in terms of accuracy and loss of the DSLR-Net model. In accuracy plot the graph shows the gradual increase of training accuracy and validation accuracy for 50 epochs. In loss plot the graph shows the gradual decrease of training loss and validation loss for 50 epochs. These plots give the evidence that our model is free from overfitting without any dropouts in our model. Hence these parameters give the rise in classification rate and can be able to classify the signs even better.

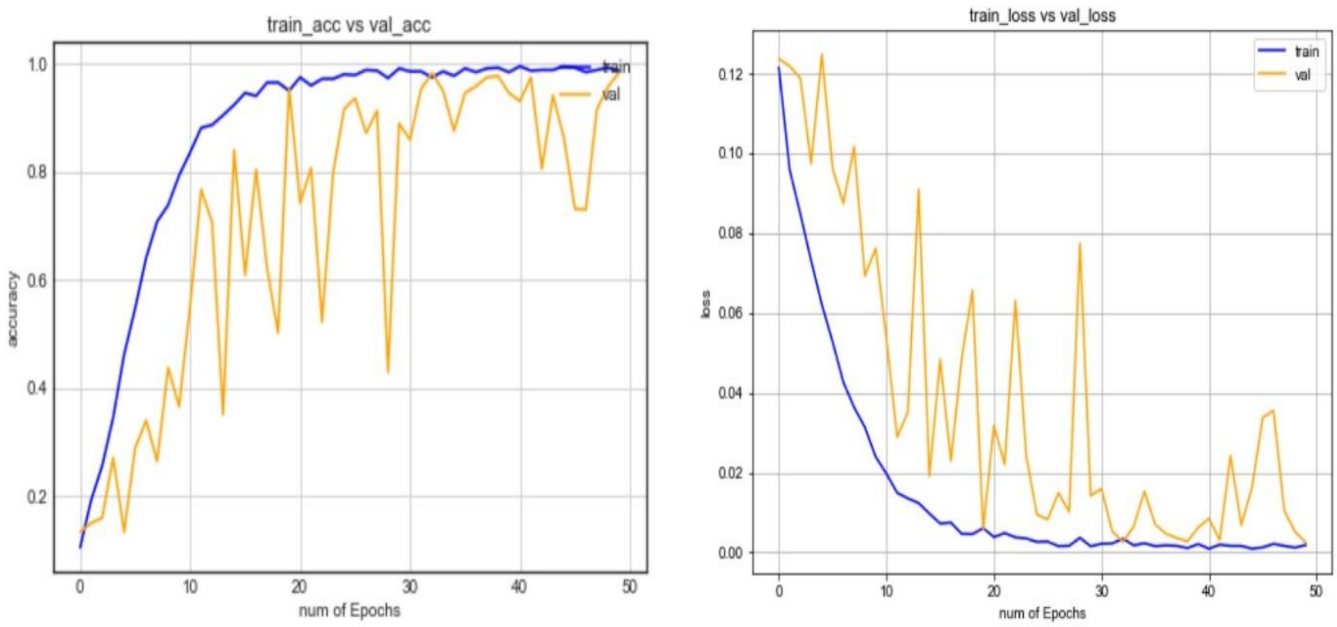

Fig 5. DSLR-Net Accuracy and Loss plots during Training

After the successful training we tested our model with the data other than trained data i.e., with different subjects here are some of predictions shown in fig 6. We gained the classification rate of 92 % for recognizing signs. The performance of the DSLR-Net is tabulated below in table-1 with cross subjects and cross views with different datasets. Table-2 gives the overall performance of the model compared to the different state-of-art techniques and with different datasets.

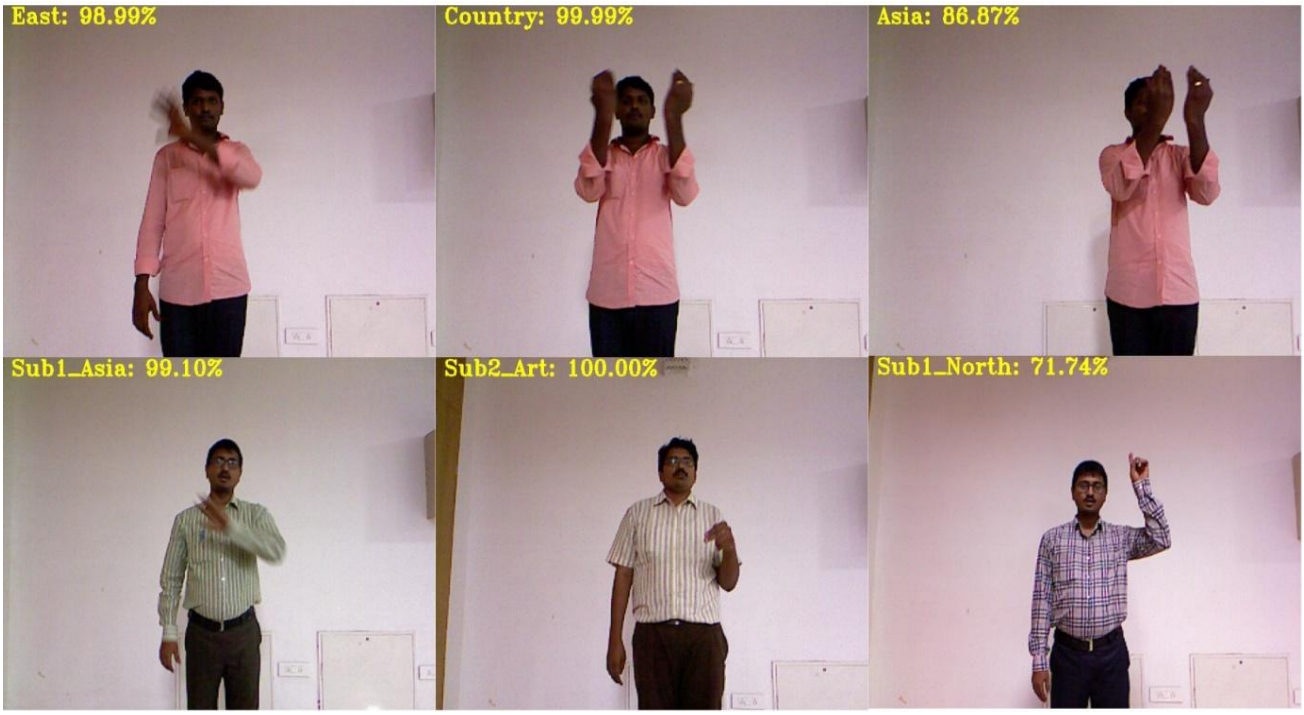

Fig 6. Exact Sign prediction when tested with trained model

Table-1. Recognition rates of different methods on publicly available datasets

Datasets		Recognition rates of different methods on publicly available datasets in %				
		X. Wang [19]	Y. Wang [20]	P. Wang [21]	D. Liu [22]	DSLR-Net
Cross Subject	BVCSL3D	85.81	88.32	88.14	82.19	94.58
	NTU RGB-D	82.34	89.64	87.25	83.14	88.24
	MSR Daily Activity 3D	79.24	81.26	80.35	78.24	87.52
	UT Kinect	81.35	84.57	82.67	81.92	85.68
Cross View	BVCSL3D	82.24	82.49	84.94	85.24	90.54
	NTU RGB-D	78.48	80.15	85.19	77.24	87.91
	MSR Daily Activity 3D	76.28	78.95	77.66	75.05	85.37
	UT Kinect	77.29	79.24	78.16	77.34	82.55
Cross Scale	BVCSL3D	76.42	77.51	76.31	77.39	88.35
	NTU RGB-D	73.24	72.29	73.21	72.28	76.14
	MSR Daily Activity 3D	69.27	71.54	70.21	68.34	73.33
	UT Kinect	69.91	72.62	72.51	68.31	73.64

Table-2. Overall Recognition rates of different methods on publicly available datasets

Overall Recognition rates of different methods on publicly available datasets in %				
Methods	Dataset			
	NTU RGB-D	MSR Daily Activity 3D	UT Kinect	BVCSL3D
X. Wang [19]	78.02	74.93	76.18	81.49
Y. Wang [20]	80.69	77.25	78.81	82.77
P. Wang [21]	81.88	76.04	77.78	83.13
D. Liu [22]	77.55	73.87	75.85	81.61
DSLR-Net	84.10	82.07	80.62	91.15

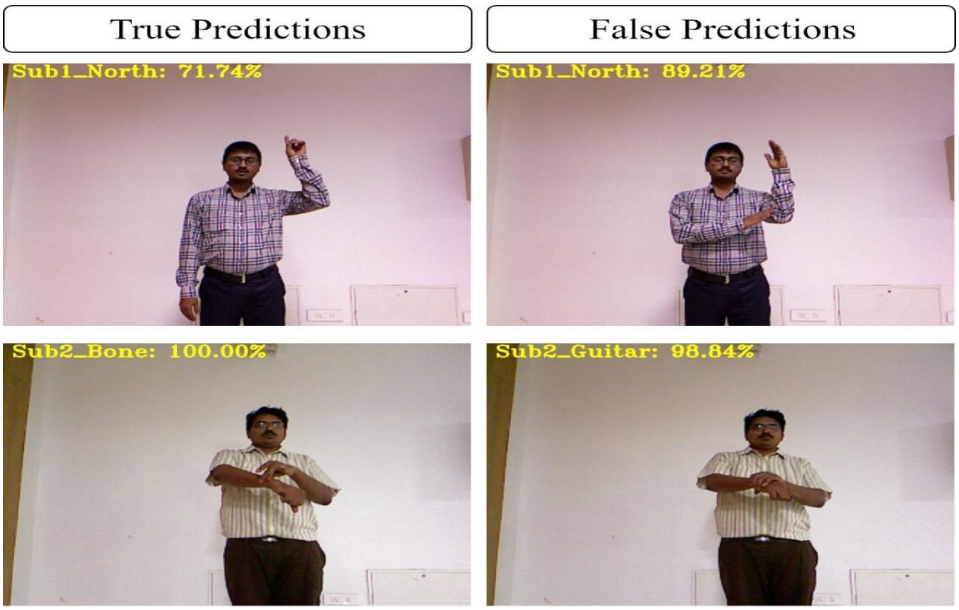

Fig 7. wrongly predicted signs when tested with trained model

To complete an action there will be one or more sub-actions these sub-actions may be a subset of different actions. Based on these sub-actions similarity between actions will be calculated. Due to these similar actions recognition rates are being affected in order to minimize this problem, we are converting the videos to frames and recognition of each frames are calculated. Average of these recognition rates gives the actual classified action. These similar actions are represented in fig 7.

Here in fig 7 show some of the signs which predicted wrongly because of the similarity in signs. In false prediction case the sign performed by subject 1 is east but shows the sign as north with 89.21% and second sign performed by subject 2 is Bite but shows the sign as guitar with 98.84%. Average of all the true predictions are increased to 92% and false predictions decreased to remaining 8%.

The confusion matrix in fig 8 is generated with RGB inputs tested on the cross subjects. The recognition rates in the fig 8 show that the network is not overfitting as we found no 100% matches with the same sign. The prediction threshold is set at 0.9 across all testing. To validate the trained DSLR-Net in cross view, we showed the remaining subjects as input in RGB frames captured in different views. Fig 9 shows the resulting confusion matrix for cross view testing on a few signs for two subjects. Most of the failures or misclassifications occurred in signs using hands with other upper body part such as head and double hands intersecting or occluding each other. The signs by subject in fig 9 shows higher misclassification rate due to the change in video background and overlapping of hands with other hands in cross view.

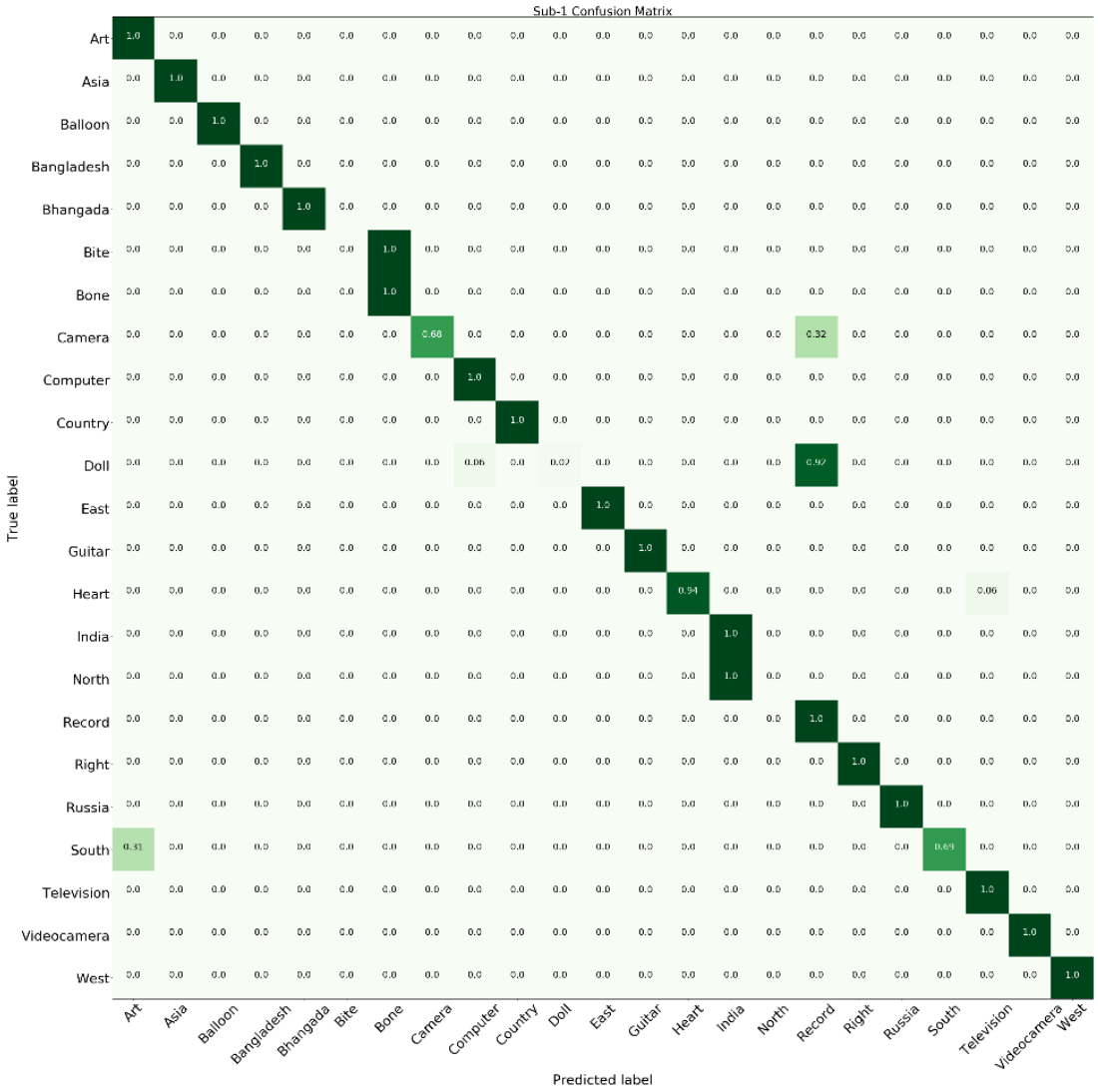

Fig 8. Confusion matrix of DSLR-Net when tested with Cross Subjects

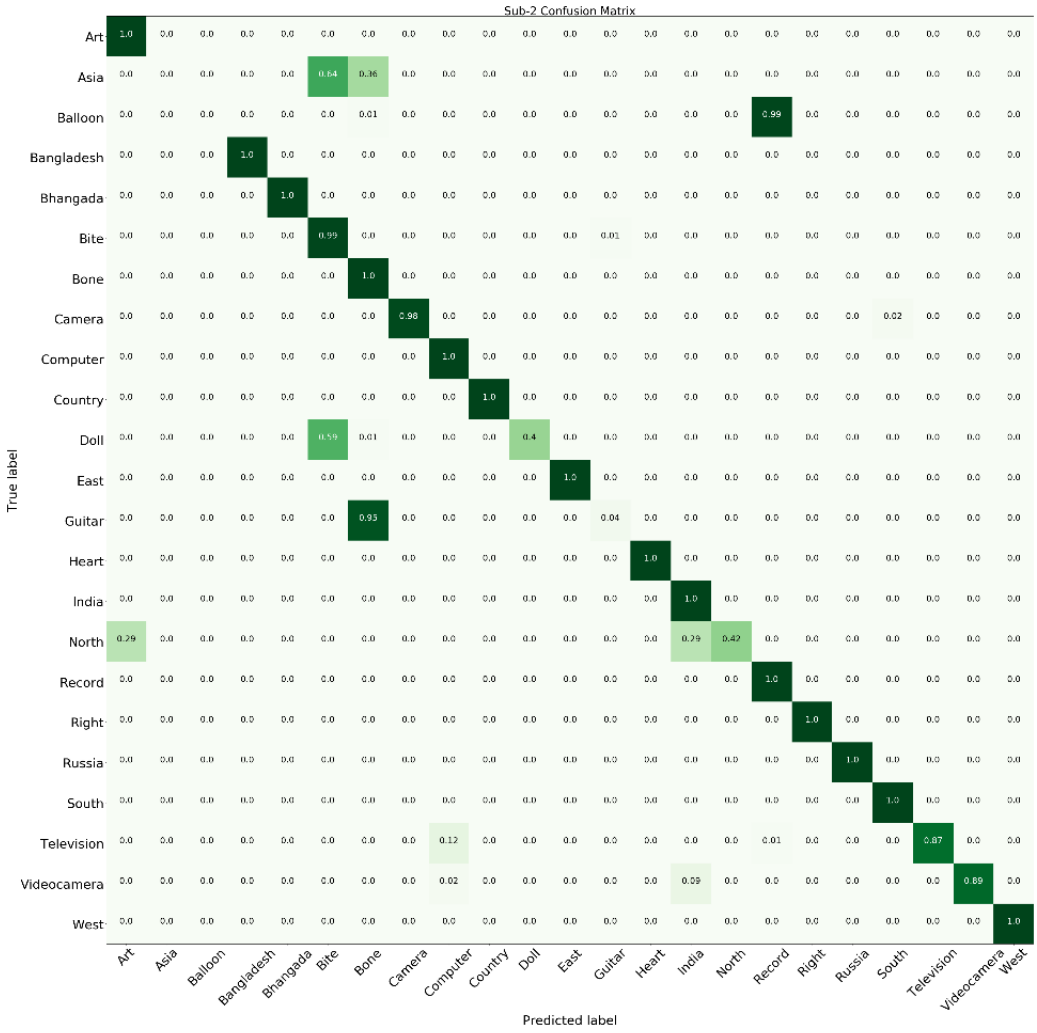

Fig 9. Confusion matrix of DSLR-Net when tested with Cross view

V. CONCLUSION

In this study, we propose to train a DSLR-Net architecture with RGB and Depth data and test the same with a RGB data for real time deployment of deep models for sign language recognition using RGB – D datasets. The CNN architecture is clustered into RGB and depth streams in which scores of each streams are concatenated to share weights of the model for better recognition rates. We test our proposed DSLR-Net model, on our own Indian sign language RGB – D dataset, BVCSL3D and three benchmark RGB – D action datasets namely MSRDailyActivity3D, UT Kinect and NTU RGB-D. The network performed better on all these datasets without initiating the dropout in any of the layers. For sign language dataset, the recognition rate is obtained as 92% on RGB input data on trained model. The results point towards a new class of deep learning architectures, which can be deployed in real time with missing trained depth modal data. The proposed DSLR-Net also performed better against some state – of – the – art CNN based RGB – D sign language recognition models.

REFERENCES

1. T. Y. Pan, L. Y. Lo, C. W. Yeh, J. W. Li, H. T. Liu and M. C. Hu, "Real-Time Sign Language Recognition in Complex Background Scene Based on a Hierarchical Clustering Classification Method," 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), Taipei, 2016, pp. 64-67.
2. Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. "A survey of depth and inertial sensor fusion for human action recognition." Multimedia Tools and Applications 76, no. 3 (2017): 4405-4425.
3. Zhang, Hao, and Lynne E. Parker. "Code4d: color-depth local spatio-temporal features for human activity recognition from rgb-d videos." IEEE Transactions on Circuits and Systems for Video Technology 26, no. 3 (2016): 541-555.
4. Maddala, T.K.K. et al., 2019. YogaNet: 3D Yoga Asana Recognition Using Joint Angular Displacement Maps with ConvNets. IEEE Transactions on Multimedia, pp.1–1.
5. Kushwah, Mukul Singh, Manish Sharma, Kunal Jain, and Anish Chopra. "Sign Language Interpretation Using Pseudo Glove." In Proceeding of International Conference on Intelligent Communication, Control and Devices, pp. 9-18. Springer Singapore, 2017.
6. Cooper, Helen, Brian Holt, and Richard Bowden. "Sign language recognition." In Visual Analysis of Humans, pp. 539-562. Springer, London, 2011.
7. Sako, S., Hatano, M. & Kitamura, T., 2016. Real-Time Japanese Sign Language Recognition Based on Three Phonological Elements of Sign. Communications in Computer and Information Science, pp.130–136. Available at: http://dx.doi.org/10.1007/978-3-319-40542-1_21.

8. Belgacem, Selma, Clément Chatelain, and Thierry Paquet. "Gesture sequence recognition with one shot learned CRF/HMM hybrid model." Image and Vision Computing 61(2017): 12-21.
9. Kishore, P. V. V., D. Anil Kumar, E. N. D. Goutham, and M. Manikanta. "Continuous sign language recognition from tracking and shape features using fuzzy inference engine." In Wireless Communications, Signal Processing and Networking (WiSPNET), International Conference on, pp. 2165-2170. IEEE, 2016.
10. Kishore, P. V. V., A. S. C. S. Sastry, and A. Kartheek. "Visual-verbal machine interpreter for sign language recognition under versatile video backgrounds." In Networks & Soft Computing (ICNSC), 2014 First International Conference on, pp. 135-140. IEEE, 2014.
11. Plouffe, G., Cretu, A.M.: 'Static and dynamic hand gesture recognition in depth data using dynamic time warping'. IEEE Transactions on Instrumentation and Measurement, 2016, 65, (2), pp. 305–316.
12. L. Pigou, S. Dieleman, P. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in Proc. Eur. Conf. Comput. Vis. Workshops, 2014, pp. 572–578.
13. Palacios, J.; Sagüés, C.; Montijano, E.; Llorente, S. Human-computer interaction based on hand gestures using RGB-D sensors. Sensors 2013, 13, 11842–11860.
14. Van den Bergh, M.; Carton, D.; De Nijs, R.; Mitsou, N.; Landsiedel, C.; Kuehnlenz, K.; Wollherr, D.; van Gool, L.; Buss, M. Real-time 3D hand gesture interaction with a robot for understanding directions from humans. In Proceedings of the IEEE RO-MAN, Atlanta, GA, USA, 31 July–3 August 2011; pp. 357–362.
15. Chai, X.; Li, G.; Lin, Y.; Xu, Z.; Tang, Y.; Chen, X.; Zhou, M. Sign Language Recognition and Translation with Kinect. In Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, Shanghai, China, 22–26 April 2013.
16. Wang, Jiang, Zicheng Liu, Ying Wu, and Junsong Yuan. "Mining actionlet ensemble for action recognition with depth cameras." In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1290-1297. IEEE, 2012.
17. L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Jun. 2012.
18. Shahroudy, A., Liu, J., Ng, T.T. and Wang, G., 2016. NTU RGB+ D: A large scale dataset for 3D human activity analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1010-1019).
19. Wang, X., Gao, L., Wang, P., Sun, X., & Liu, X. (2018). Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length. IEEE Transactions on Multimedia, 20(3), 634-644.
20. Wang, Y., Song, J., Wang, L., Van Gool, L., & Hilliges, O. (2016, September). Two-Stream SR-CNNs for Action Recognition in Videos. In BMVC.
21. Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., & Ogunbona, P. O. (2016). Action recognition from depth maps using deep convolutional neural networks. IEEE Transactions on Human-Machine Systems, 46(4), 498-509.
22. Liu, D., Wang, Y., & Kato, J. (2017, November). Evaluation of Triple-Stream Convolutional Networks for Action Recognition. In Digital Image Computing: Techniques and Applications (DICTA), 2017 International Conference on (pp. 1-6). IEEE.