ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

Real-Time American Sign Language Recognition with Faster Regional Convolutional Neuralnetworks

S. Dinesh[1], S. Sivaprakash[2], M. Keshav[3], K. Ramya[4]

Student, Department of Computer Science Engineering, Velammal Institute of Technology, Chennai, Tamil Nadu, India [1,2,3]

Assistant Professor, Department of Computer Science Engineering, Velammal Institute of Technology, Chennai, Tamil Nadu, India [4]

ABSTRACT: A real-time sign language translator is an important milestone in facilitating communication between the deaf community and the general public. This paper presents the development and implementation ofanAmerican Sign Language (ASL) fingerspelling translator based on convolutional neural network. This paper utilizes a pre-trained GoogLeNet architecture trained on the ILSVRC2012 dataset, as well as the Surrey University and Massey University ASL datasets in order to apply Transfer-Learning to this task. This paper proposes a robust model that consistently classifies letters a-e correctly with first-time users and another that correctly classifies letters a-k in a majority of cases. Given the limitations of the datasets and the encouraging results achieved.

KEYWORDS: Faster R-CNN, ASL, RNN, ILSVRC2012 dataset, Softmax-based loss function.

I.INTRODUCTION

American Sign Language (ASL) substantially facilitates communication in the deaf community. However, there are only ~250,000-500,000 speakers which significantly limits the number of people that they can easily communicate with [1]. The alternative of written communication is cumbersome, impersonal and even impractical when an emergency occurs. In order to diminish this obstacle and to enable dynamic communication, we present an ASL recognition system that uses Convolutional Neural Networks (CNN) in real time to translate a video of a user's ASL signs into text. Our problem consists of three tasks to be done in real time:
1. Obtaining video of the user signing (input)
2. Classifying each frame in the video to a letter
3. Reconstructing and displaying the most likely word from classification scores (output).

From a computer vision perspective, this problem represents a significant challenge due to a number of considerations, including:

- Environmental concerns (e.g. lighting sensitivity, background, and camera position)
- Occlusion (e.g. some or all fingers, or an entire hand can be out of the field of view)
- Sign boundary detection (when a sign ends and the next begins)
- Co-articulation (when a sign is affected by the preceding or succeeding sign)

While Neural Networks have been applied to ASL letter recognition (Appendix A) in the past with accuracies that are consistently over 90% [2-11], many of them require a 3-D capture element with motion-tracking gloves or a

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

Microsoft Kinect, and only one of them provides real-time classifications. The constraints imposed by the extra requirements reduce the scalability and feasibility of these solutions.

II. LITERATURE SURVEY

ASL recognition is not a new computer vision problem. Over the past two decades, researchers have used classifiers from a variety of categories that we can group roughly into linear classifiers, neural networks and Bayesian networks [2-11].

While linear classifiers are easy to work with because they are relatively simple models, they require sophisticated feature extraction and preprocessing methods to be successful [2, 3, 4]. Singha and Das obtained accuracy of 96% on 10 classes for images of gestures of one hand using Karhunen-Loeve Transforms [2]. These translate and rotate the axes to establish a newcoordinate system based on the variance of the data. This transformation is applied after using a skin filter, hand cropping and edge detection on the images. They use a linear classifier to distinguish between hand gestures including thumbs up, index finger pointing left and right, and numbers (no ASL). Sharma et al. use piece-wise classifiers (Support Vector Machines and k-Nearest Neighbours) to characterize each color channel after background subtraction and noise removal [4]. Their innovation comes from using a contour trace, which is an efficient representation of hand contours. They attain an accuracy of 62.3% using an SVM on the segmented color channel model.

Bayesian networks like Hidden Markov Models have also achieved high accuracies [5, 6, 7]. These are particularly good at capturing temporal patterns, but they require clearly defined models that are defined prior tolearning. Starner and Pentland used a Hidden Markov Model (HMM) and a 3-D glove that tracks hand movement [5]. Since the glove is able to obtain 3-D information from the hand regardless of spatial orientation, they were able to achieve an impressive accuracy of 99.2% on the test set. Their HMM uses time-series data to track hand movements and classify based on where the hand has been in recent frames.

Suk et al. propose a method for recognizing handgestures in a continuous video stream using a *dynamicBayesian network* or DBN model [6]. They attempt toclassify moving hand gestures, such as making a circlearound the body or waving. They achieve an accuracy of over 99%, but it is worth noting that all gestures are markedly different from each other and that they are not American Sign Language. However, the motion-tracking feature would be relevant for classifying the dynamic letters of ASL: *j* and *z*. Some neural networks have been used to tackle ASLtranslation [8, 9, 10, 11]. Arguably, the most significant advantage of neural networks is that they learn the most important classification features. However, they require considerably more time and data to train. To date, most have been relatively shallow. Mekala et al. classified video of ASL letters into text using advanced feature extraction and a 3-layer Neural Network [8]. They extracted features in two categories: hand position and movement. Prior to ASL classification, they identify the presence and location of 6 “points of interest” in the hand: each of the fingertips and the center of the palm. Mekala et al. also take Fourier Transforms of the images and identify what section of the frame the hand is located in. While they claim to be able to correctly classify 100% of images with this framework, there is no mention of whether this result was achieved in the training, validation or test set. Admasu and Raimond classified Ethiopian Sign Language correctly in 98.5% of cases using a feedforwardNeural Network [9]. They use a significantamount of image preprocessing, including image sizenormalization, image background subtraction, contrastadjustment, and image segmentation. AdmasuandRaimond extracted features with a Gabor Filter andPrincipal Component Analysis. The most relevant work to date is L. Pigou et al’sapplication of CNN’s to classify 20 Italian gestures from the ChaLearn 2014 Looking at People gesture spotting competition [11]. They use a Microsoft Kinect on fullbodyimages of people performing the gestures and achieve a cross-validation accuracy of 91.7%. As in thecase with the aforementioned 3-D glove, the Kinectallowscapture of depth features, which aids significantly inclassifying ASL signs.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

III.PROPOSED METHODOLOGY AND DISCUSSION

3.1 Mathematical Equation for Transfer Learning

Our ASL letter classification is done using a convolutional neural network (CNN or ConvNet). CNNs are machine learning algorithms that have seen incredible success in handling a variety of tasks related to processing videos and images. Since 2012, the field has experienced an explosion of growth and applications in image classification, object localization, and object detection.

A primary advantage of utilizing such techniques stems from CNNs abilities to learn features as well as the weights corresponding to each feature. Like other machine learning algorithms, CNNs seek to optimize some objective function, specifically the loss function. We utilized a softmax-based loss function:

$$Loss = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{f_{i,y_i}}}{\sum_{j=1}^{C} e^{f_{i,j}}}\right)$$

$$f_j(z) = \frac{e^{z_j}}{\sum_{k=1}^{C} e^{z_k}}$$

N = total number of training examples
C = total number of classes

Equation (2) is the softmax function. It takes a feature vector *z* for a given training example, and squashes its values to a vector of [0,1]-valued real numbers summing to 1. Equation (1) takes the mean loss for each training example, x_i, to produce the full softmax loss.

Using a softmax-based classification head allows us to output values akin to probabilities for each ASL letter. This differs from another popular choice: the SVM loss.Using an SVM classification head would result in scores for each ASL letter that would not directly map to probabilities. These probabilities afforded to us by the softmax loss allow us to more intuitively interpret our results and prove useful when running our classifications through a language model.

3.1.1 Transfer Learning

Transfer-Learning is a machine learning technique where models are trained on (usually) larger data sets and refactored to fit more specific or niche data. This is done by recycling a portion of the weights from the pre-trained model and reinitializing or otherwise altering weights at shallower layers. The most basic example of this would be a fully trained network whose final classification layer weights have been reinitialized to be able to classify some new set of data. The primary benefits of such a technique are its less demanding time and data requirements. However, the challenge in transfer learning stems from the differences between the original data used to train and the new data being classified. Larger differences in these data sets often require re-initializing or increasing learning rates for deeper layers in the net.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

3.1.2 Caffe and GoogLeNet

We employed Caffe, a deep learning framework [12], in order to develop, test, and run our CNNs. Specifically, we used Berkeley Vision and Learning Center'sGoogLeNet pre-trained on the 2012 ILSVRC dataset. This net output three different losses at various depths of the net and combines these losses prior to computing a gradient at training time.

3.2. General Technique

Our overarching approach was to fine-tune a pre-trained GoogLeNet. Our data is composed exclusively of hands in 24 different orientations, while the ILSVRC data is composed of 1000 uniquely different objects or classes. Even though the ILSVRC data has some classes that are quite similar to each other (e.g. various breeds of dogs), our data was comprised of the same object merely positioned and oriented in different ways. Considering the stark differences between our data and the data that GoogLeNet was pre-trained on, we decided to test the effectiveness of altering a variety of the pre-trained weights at different depths. We amplified learning rates by a factor of ten and completely reset weights in the first one, two or three layers of the net using Xavier initialization.

3.3. Developing our pipeline

Recognizing a sign language gestures from continuous gestures is a very challenging research issue. The researchers solved this problem using gradient based key frame extraction method. These key frames were helpful in splitting continuous sign language gestures into sequence of signs as well as for removing uninformative frames. After splitting of gestures each sign has been treated as an isolated gesture. Then features of pre-processed gestures were extracted using Orientation Histogram (OH) with Principal Component Analysis (PCA) is applied for reducing dimension of features obtained after OH. Experiments were performed on their own continuous ISL dataset which was created using canon EOS camera. Probes were tested using various types of classifiers like

Euclidean distance, Correlation, Manhattan distance, city block distance etc. Comparative analysis of their proposed scheme was performed with various types of distance classifiers. From the above analysis they found that the results obtained from Correlation and Euclidean distance gives better accuracy than other classifiers.

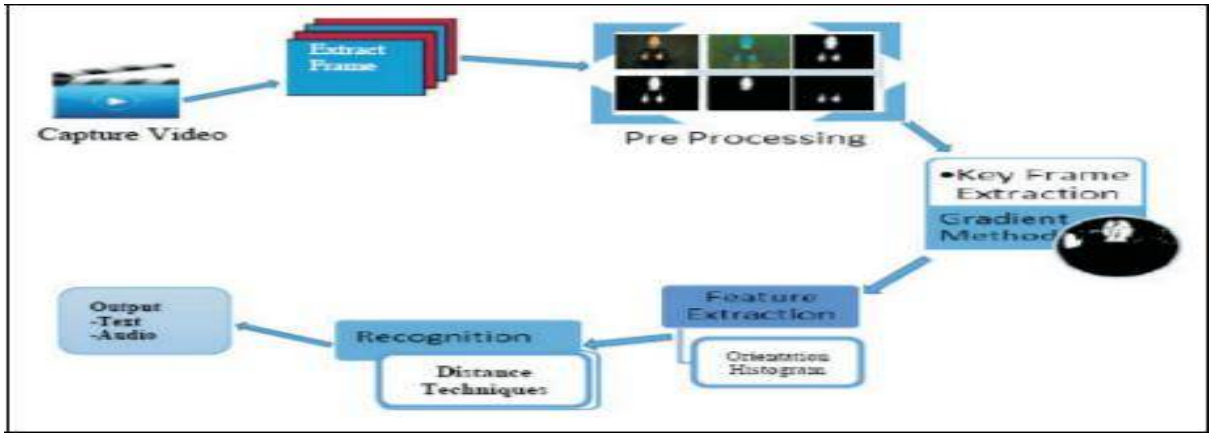

Fig 1. Data Flow for image classification using faster r-cnn

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

3.4. Datasets and Features

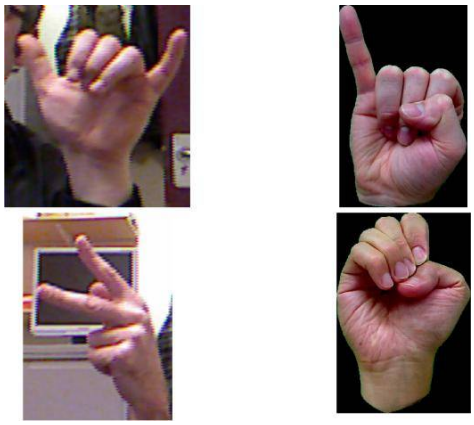

Fig 2. Dataset examples

3.5Extracting Spatial Features

In this approach spatial features are extracted for individual frames using inception model (CNN) and temporal features using RNN. Each video (a sequence of frames) was then represented by a sequence of predictions made by CNN for each of the individual frames. This sequence of predictions were given as input to the RNN.

3.5.1 Methodology

- First, we will extract the frames from the multiple video sequences of each gesture.
- After the first step, noise from the frames i.e background, body parts other than hand are removed to extract more relevant features from the frame.
- Frames of the train data are given to the CNN model for training on the spatial features. We have used inception model for this purpose which is a deep neural net.
- Store the train and test frame predictions. We'll use the model obtained in the above step for the prediction of frames.
- The predictions of the train data are now given to the RNN model for training on the temporal features. We have used LSTM model for this purpose.

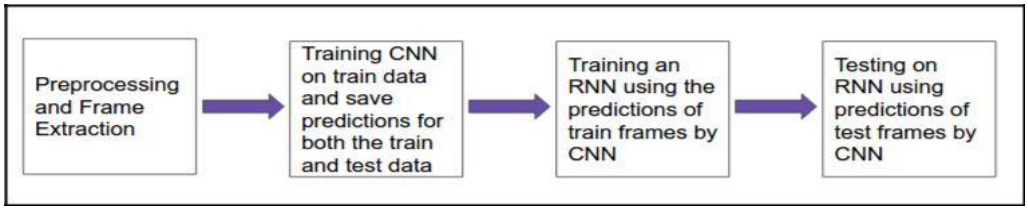

Fig 3. Common Flow for training images using faster r-cnn.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

3.5.2 Train CNN (Spatial Features) and Prediction

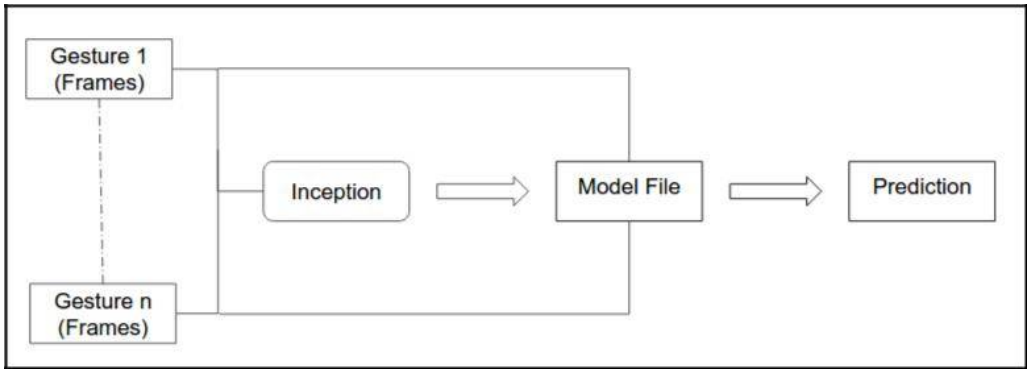

Fig 4. Frame after extracting hands (Background Removal)

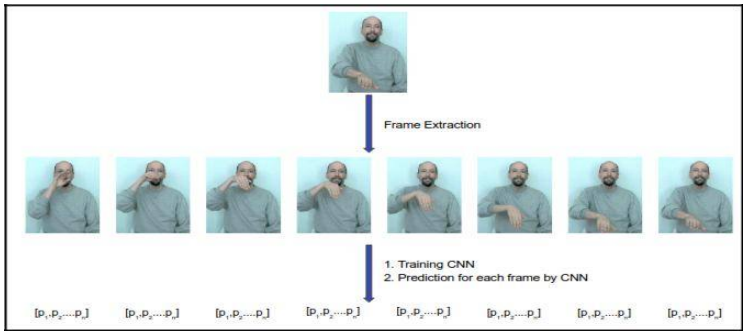

Fig. 5. Frame Extraction and prediction using faster r-cnn.

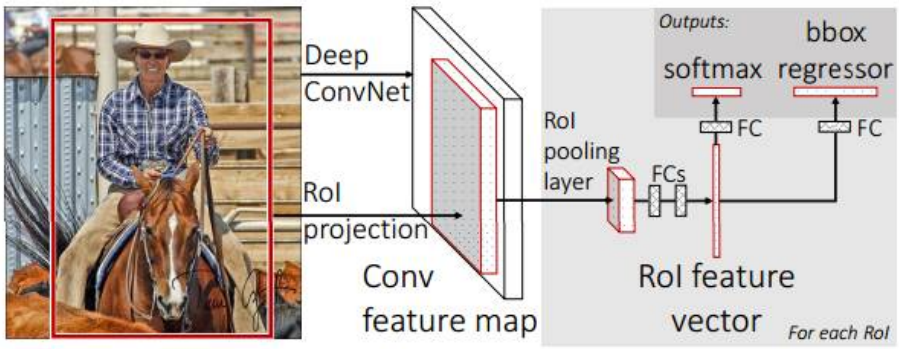

Fig 6. Object deduction using faster R-CNN

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

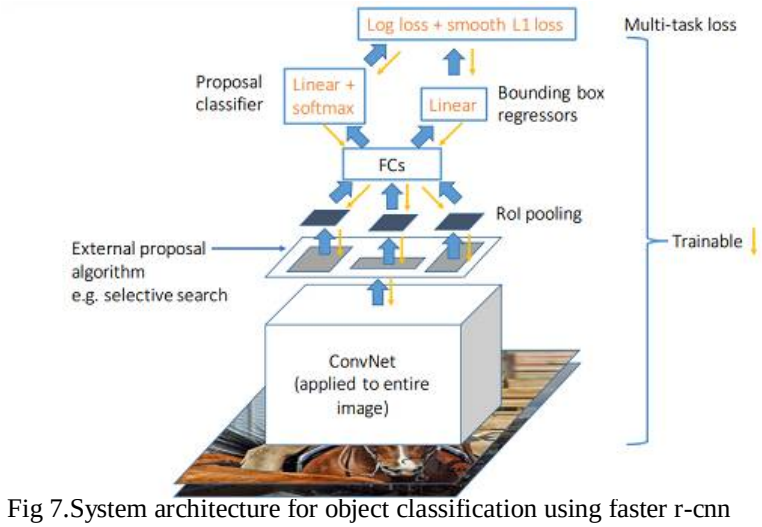

Fig 7.System architecture for object classification using faster r-cnn

IV. RESULTS

4.1 RESIZING THE IMAGES TO A COMMON SHAPE:

- Every images we randomly get on Internet or images we get from a dataset managing site are most probably the images of different shapes.

- In order to classify or play with images, we need those images to be in a common shape i.e., having common dimensions(width and height).
- Here in this module I'm converting the images in my folder to a pixel of 28*28.

Fig 8 - Alphabet A in the image dataset

4.2 RENAMING IMAGES IN THE FOLDER

- In order to classify any image's you need those images to be of common format.
- Images which I have downloaded from internet have different names right!
- Then we should process those data to get an ordered labelling of names such as A0.jpg, A1.jpg etc.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

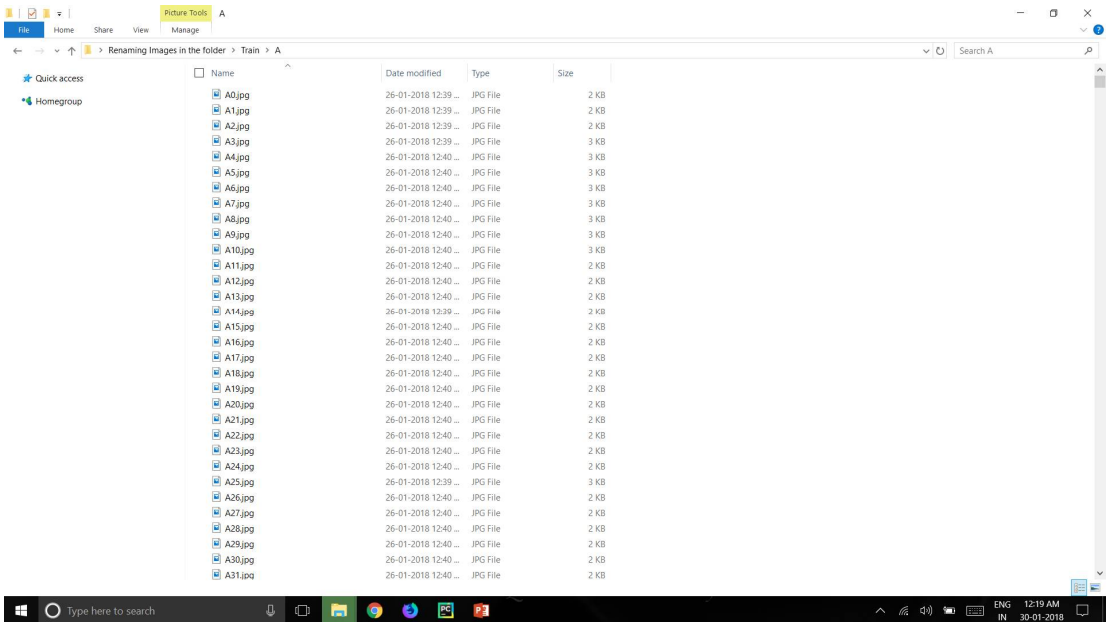

Fig 9- Renaming images to a common name

4.3PERCENTAGE OF ACCURACY USING FASTER R-CNN

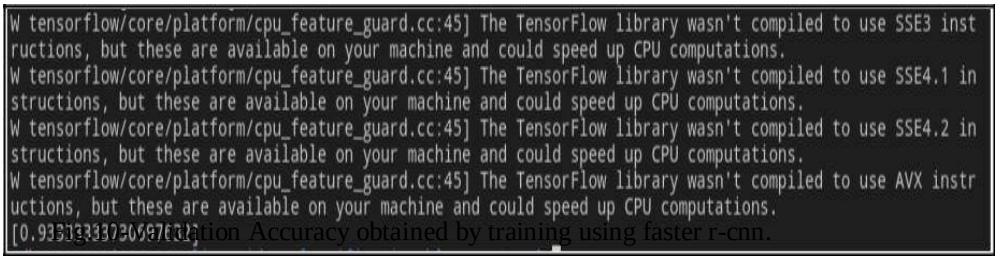

Fig 10- Classification Accuracy obtained by training using faster r-cnn.

Average accuracy obtained using this approach is 93.3333%.

V. CONCLUSION

Hand gestures are a powerful way for human communication, with lots of potential applications in the area of human computer interaction. Vision based hand gesture recognition techniques have many proven advantages compared with traditional devices. However, hand gesture recognition is a difficult problem and the current work is only a small contribution towards achieving the results needed in the field of sign language gesture recognition. This report presented a vision based system able to interpret isolated hand gestures from the Argentinian Sign Language (LSA).

Videos are difficult to classify because they contain boththe temporal as well as the spatial features. This paper's proposed model used two different models to classify on the spatial and temporal features. CNN was used to classify on the spatial features. Faster r-cnn model has obtained an accuracy of 93.33 %. This shows that CNN can be successfully used to learn spatial and temporal features and classify Sign Language Gestures.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

This method for individual gestures can also be extended for sentence level sign language. Also the current process uses two different models, training inception (CNN) followed by training RNN. For future work one can focus on combining the two models into a single model.

REFERENCES

[1] Mitchell, Ross; Young, Travas; Bachleda, Bellamie; Karchmer, Michael (2006). "How Many People Use ASL in the United States?: Why Estimates Need Updating" (PDF). Sign Language Studies (Gallaudet University Press.) 6 (3). ISSN 0302-1475. Retrieved November 27, 2012.
[2] Singha, J. and Das, K. "Hand Gesture Recognition Based on Karhunen-Loeve Transform", Mobile and Embedded
[3] Technology International Conference (MECON), January7-18, 2013, India. 365-371.
[4] D. Aryanie, Y. Heryadi. American Sign Language-Based Finger-spelling Recognition using k-Nearest Neighbors Classifier. 3rd International Conference on Information and Communication Technology (2015) 533-536.
[5] R. Sharma et al. Recognition of Single Handed Sign Language Gestures using Contour Tracing descriptor. Proceedings of the World Congress on Engineering 2013 Vol. II, WCE 2013, July 3 - 5, 2013, London, U.K.
[6] T.Starner and A. Pentland. Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. Computational Imaging and Vision, 9(1);227-243, 1997.
[7] M. Jeballi et al. Extension of Hidden Markov Model for Recognizing Large Vocabulary of Sign Language. International Journal of Artificial Intelligence & Applications 4(2); 35-42, 2013
[8] H. Suk et al. Hand gesture recognition based on dynamic Bayesian network framework. Patter Recognition 43 (9); 3059-3072, 2010.
[9] P. Mekala et al. Real-time Sign Language Recognition based on Neural Network Architecture. System Theory (SSST), 2011 IEEE 43rd Southeastern Symposium 14-16 March 2011.
[10] Y.F. Admasu, and K. Raimond, Ethiopian Sign Language Recognition Using Artificial Neural Network. 10th International Conference on Intelligent Systems Design and Applications, 2010. 995-1000.
[11] J. Atwood, M. Eicholtz, and J. Farrell. American Sign Language Recognition System. Artificial Intelligence and Machine Learning for Engineering Design. Dept. of Mechanical Engineering, Carnegie Mellon University, 2012.
[12] L. Pigou et al. Sign Language Recognition Using Convolutional Neural Networks. European Conference on Computer Vision 6-12 September 2014
[13] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2014.Lifeprint.com. American Sign Language (ASL) Manual Alphabet (fingerspelling) 2007.