International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 12, December-2020

Sign Language Recognition System using Convolutional Neural Network and Computer Vision

Mehreen Hurroo
Computer Science & Engineering Department
Delhi Technological University (DTU)
Delhi, India

Mohammad Elham Walizad
Computer Science & Engineering Department
Delhi Technological University (DTU)
Delhi, India

***Abstract* — Conversing to a person with hearing disability is always a major challenge. Sign language has indelibly become the ultimate panacea and is a very powerful tool for individuals with hearing and speech disability to communicate their feelings and opinions to the world. It makes the integration process between them and others smooth and less complex. However, the invention of sign language alone, is not enough . There are many strings attached to this boon.The sign gestures often get mixed and confused for someone who has never learnt it or knows it in a different language. However, this communication gap which has existed for years can now be narrowed with the introduction of various techniques to automate the detection of sign gestures . In this paper, we introduce a Sign Language recognition using American Sign Language.**
In this study, the user must be able to capture images of the hand gesture using web camera and the system shall predict and display the name of the captured image. We use the HSV colour algorithm to detect the hand gesture and set the background to black. The images undergo a series of processing steps which include various Computer vision techniques such as the conversion to grayscale, dilation and mask operation. And the region of interest which, in our case is the hand gesture is segmented. The features extracted are the binary pixels of the images. We make use of Convolutional Neural Network(CNN) for training and to classify the images. We are able to recognise 10 American Sign gesture alphabets with high accuracy. Our model has achieved a remarkable accuracy of above 90%.

Keywords: Sign Language, ASL, Hearing disability, Convolutional Neural Network(CNN), Computer Vision, Machine Learning, Gesture recognition, Sign language recognition, Hue Saturation Value algorithm.

I. INTRODUCTION

As well stipulated by Nelson Mandela[1], "*Talk to a man in a language he understands, that goes to his head. Talk to him in his own language, that goes to his heart*", language is undoubtedly essential to human interaction and has existed since human civilisation began. It is a medium humans use to communicate to express themselves and understand notions of the real world. Without it, no books, no cell phones and definitely not any word I am writing would have any meaning. It is so deeply embedded in our everyday routine that we often take it for granted and don't realise its importance. Sadly, in the fast changing society we live in, people with hearing impairment are usually forgotten and left out. They have to struggle to bring up their ideas, voice out their opinions and express themselves to people who are different to them. Sign language, although being a medium of communication to deaf people, still have no meaning when conveyed to a non-sign language user. Hence, broadening the communication gap.To prevent this from happening, we are putting forward a sign language recognition system. It will be an ultimate tool for people with hearing disability to communicate their thoughts as well as a very good interpretation for non sign language user to understand what the latter is saying. Many countries have their own standard and interpretation of sign gestures. For instance, an alphabet in Korean sign language will not mean the same thing as in Indian sign language. While this highlights diversity, it also pinpoints the complexity of sign languages. Deep learning must be well versed with the gestures so that we can get a decent accuracy. In our proposed system, American Sign Language is used to create our datasets. Figure 1 shows the American Sign Language (ASL) alphabets.
Identification of sign gesture is performed with either of the two methods. First is a glove based method whereby the signer wears a pair of data gloves during the capture of hand movements.Second is a vision based method, further classified into static and dynamic recognition[2]. Static deals with the 2dimensional representation of gestures while dynamic is a real time live capture of the gestures.

And despite having an accuracy of over 90%[3], wearing of gloves are uncomfortable and cannot be utilised in rainy weather. They are not easily carried around since their use require computer as well. In this case, we have decided to go with the static recognition of hand gestures because it increases accuracy as compared to when including dynamic hand gestures like for the alphabets J and Z. We are proposing this research so we can improve on accuracy using Convolution Neural Network(CNN).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 12, December-2020

Figure 1 American Sign Language alphabets

II. EXISTING LITERATURE

Literature review of our proposed system shows that there have been many explorations done to tackle the sign recognition in videos and images using several methods and algorithms.

Siming He[4] proposed a system having a dataset of 40 common words and 10,000 sign language images. To locate the hand regions in the video frame, Faster R-CNN with an embedded RPN module is used. It improves performance in terms of accuracy. Detection and template classification can be done at a higher speed as compared to single stage target detection algorithm such as YOLO. The detection accuracy of Faster R-CNN in the paper increases from 89.0% to 91.7% as compared to Fast-RCNN. A 3D CNN is used for feature extraction and a sign-language recognition framework consisting of long and short time memory (LSTM) coding and decoding network are built for the language image sequences.On the problem of RGB sign language image or video

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 12, December-2020

recognition in practical problems , the paper merges the hand locating network, 3D CNN feature extraction network and LSTM encoding and decoding to construct the algorithm for extraction. This paper has achieved a recognition of 99% in common vocabulary dataset.

Let's approach the research done by Rekha, J[5]. which made use of YCbCr skin model to detect and fragment the skin region of the hand gestures. Using Principal Curvature based Region Detector, the image features are extracted and classified with Multi class SVM, DTW and non-linear KNN. A dataset of 23 Indian Sign Language static alphabet signs were used for training and 25 videos for testing. The experimental result obtained were 94.4%for static and 86.4% for dynamic.

In [6] , a low cost approach has been used for image processing . The capture of images was done with a green background so that during processing, the green colour can be easily subtracted from the RGB colourspace and the image gets converted to black and white. The sign gestures were in Sinhala language. The method that thy have proposed in the study is to map the signs using centroid method. It can map the input gesture with a database irrespective of the hands size and position. The prototype has correctly recognised 92% of the sign gestures.

The paper by M. Geetha and U. C. Manjusha[7], make use of 50 specimens of every alphabets and digits in a vision based recognition of Indian Sign Language characters and numerals using B-Spine approximations. The region of interest of the sign gesture is analysed and the boundary is removed . The boundary obtained is further transformed to a B-spline curve by using the Maximum Curvature Points(MCPs) as the Control points. The B-spline curve undergoes a series of smoothening process so features can be extracted. Support vector machine is used to classify the images and the accuracy is 90.00%.

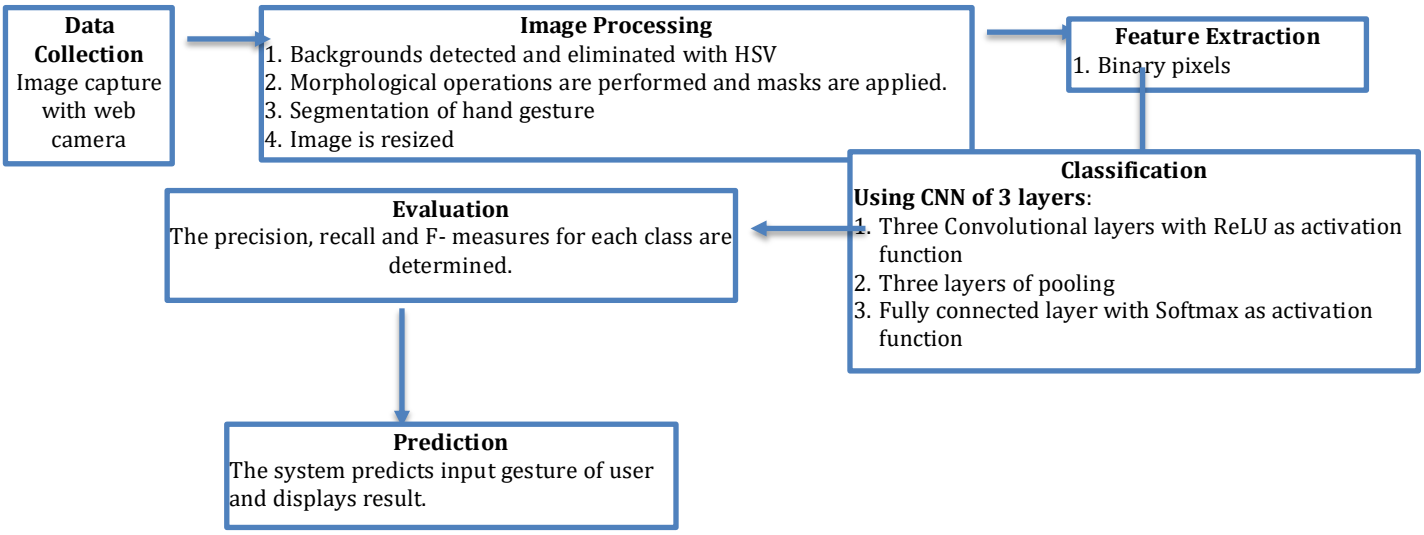
In [8], Pigou used CLAP14 as his dataset [9]. It consists of 20 Italian sign gestures.After preprocessing the images , he used a Convolutional Neural network model having 6 layers for training. It is to be noted that his model is not a 3D CNN and all the kernels are in 2D. He has used Rectified linear Units (ReLU) as activation functions. Feature extraction is performed by the CNN while classification uses ANN or fully connected layer. His work has achieved an accuracy of 91.70% with an error rate of 8.30%.

A similar work was done by J Huang [10].He created his own dataset using Kinect and got a total a total of 25 vocabularies which are used in everyday lives. He then applied a 3D CNN in which all kernels are also in 3D. The input of his model consisted of 5 important channels which are colour-r, colour-b, colour-g, depth and body skeleton. He got an average accuracy of 94.2%.

Another research paper on Action recognition topic by the author J.Carriera [11] shares some similarities to sign gesture recognition .He used a transfer learning method for his research As his pre-trained dataset, he used both ImageNet[12] and Kinetic Dataset [9] . After training the pertained models using another two datasets namely UCF-101 [13] and HMDB-51 [14], he then merged the RGB model, flow model, pre-trained Kinetic and pre-trained ImageNet.The accuracy he got on UCF-101 dataset is 98.0% and on HMDB-51 is 80.9%

III. METHODOLOGY

The first step of the proposed system is to collect data.Many researchers have used sensors or cameras to capture the hand movements. For our system, we make use of the web camera to shoot the hand gestures.The images undergo a series of processing operations whereby the backgrounds are detected and eliminated using the colour extraction algorithm HSV(Hue,Saturation,Value). Segmentation is then performed to detect the region of the skin tone. Using the morphological operations, a mask is applied on the images and a series of dilation and erosion using elliptical kernel are executed . With openCV, the images obtained are amended to the same size so there is no difference between images of different gestures . Our dataset has 2000 American sign gesture images out of which 1600 images are for training and the rest 400 are for testing purposes. It is in the ratio 80:20. Binary pixels are extracted from each frame, and Convolutional Neural Network is applied for training and classification. The model is then evaluated and the system would then be able to predict the alphabets.



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 12, December-2020

IV. DATA COLLECTION

Data collection is indelibly an essential part in this research as our result highly depends on it. We have therefore created our own dataset of ASL having 2000 images of 10 static alphabet signs.We have 10 classes of static alphabets which are A,B,C,D,K,N,O,T and Y. Two datasets have been made by 2 different signers. Each of them has performed one alphabetical gesture 200 times in alternate lighting conditions.The dataset folder of alphabetic sign gestures is further split into 2 more folders, one for training and the other for testing. Out of the 2000 images captured, 1600 images are used for training and the rest for testing. To get higher consistency, we have captured the photos in the same background with a webcam each time a command is given.The images obtained are saved in the png format .It is to be pinpointed that there is no loss in quality whenever an image in png format is opened ,closed and stored again.PNG is also good in handling high contrast and detailed image. The webcam will capture the images in the RGB colourspace.

V. DATA PROCESSING

A. HSV colourspace and background elimination

Since the images obtained are in RGB colourspaces, it becomes more difficult to segment the hand gesture based on the skin colour only. We therefore transform the images in HSV colourspace. It is a model which splits the colour of an image into 3 separate parts namely: Hue,Saturation and value. HSV is a powerful tool to improve stability of the images by setting apart brightness from the chromaticity [15]. The Hue element is unaffected by any kind of illumination, shadows and shadings[16] and can thus be considered for background removal. A track-bar having H ranging from 0 to 179, S ranging from 0-255 and V ranging from 0 to 255 is used to detect the hand gesture and set the background to black. The region of the hand gesture undergoes dilation and erosion operations with elliptical kernel. The first image is obtained after applying the 2 masks as shown in fig 3(b).

Figure 3 (a) Image captured from web-camera. (b) Image after background is set to black using HSV (first image).

B. Segmentation

The first image is then transformed to grayscale. As much as this process will result in the loss of colour in the region of the skin gesture, it will also enhance the robustness of our system to changes in lighting or illumination. Non-black pixels in the transformed image are binarised while the others remain unchanged, therefore black.The hand gesture is segmented firstly by taking out all the joined components in the image and secondly by letting only the part which is immensely connected, in our case is the hand gesture. The frame is resized to a size of 64 by 64 pixel. At the end of the segmentation process, binary images of size 64 by 64 are obtained where the area in white represents the hand gesture, and the black coloured area is the rest.

Figure 4 (a) Image after binarise. (b) Image after segmentation and resizing.

C. Feature Extraction

One of the most crucial part in image processing is to select and extract important features from an image. Images when captured and stored as a dataset usually take up a whole lot of space as they are comprised of a huge amount of data. Feature extraction helps us solve this problem by reducing the data after having extracted the important features automatically.It also contributes in maintaining the accuracy of the classifier and simplifies its complexity. In our case, the features found to be crucial are the binary pixels of the images. Scaling the images to 64 pixels has led us to get sufficient features to effectively classify the American Sign Language gestures . In total, we have 4096 number of features, obtained after multiplying 64 by 64 pixels.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 12, December-2020

VI. SYSTEM ARCHITECTURE

A CNN model is used to extract features from the frames and to predict hand gestures. It is a multilayered feedforward neural network mostly used in image recognition. The architecture of CNN consists of some convolution layers, each comprising of a pooling layer, activation function, and batch normalisation which is optional. It also has a set of fully connected layers. As one of the images moves across the network, it gets reduced in size. This happens as a result of max pooling. The last layer gives us the prediction of the class probabilities.

A. Classification

In our proposed system, we apply a 2D CNN model with a tensor flow library. The convolution layers scan the images with a filter of size 3 by 3. The dot product between the frame pixel and the weights of the filter are calculated. This particular step extracts important features from the input image to pass on further. The pooling layers are then applied after each convolution layer. One pooling layer decrements the activation map of the previous layer. It merges all the features that were learned in the previous layers' activation maps. This helps to reduce overfitting of the training data and generalises the features represented by the network. In our case, the input layer of the convolutional neural network has 32 feature maps of size 3 by 3, and the activation function is a Rectified Linear Unit. The max pool layer has a size of 2×2. The dropout is set to 50 percent and the layer is flattened. The last layer of the network is a fully connected output layer with ten units, and the activation function is Softmax. Then we compile the model by using category cross-entropy as the loss function and Adam as the optimiser.

VII. EVALUATION

The model is evaluated based on 10 alphabetic American sign language including : A, B, C, D, H, K, N,O,T and Y. We have used a total of 2000 images to train the Convolutional Neural Network. The dataset is split in the ratio of 80:20 for training and testing respectively. The results used in this paper gives us an accuracy of over 90.0%, which is better than any work mentioned in the paper. Table 1 shows detailed precision, recall and F- measures for each class.

TABLE 1 PRECISION, RECALL, F-MEASURE

Letter	Precision	Recall	F-Measure	Support
A	0.98	1.00	0.99	40
B	1.00	1.00	1.00	40
C	1.00	1.00	1.00	40
D	1.00	1.00	1.00	40
H	1.00	1.00	1.00	40
K	1.00	0.88	0.93	40
N	1.00	0.97	0.99	40
O	0.87	0.97	0.92	40
T	1.00	1.00	1.00	40
Y	1.00	1.00	1.00	40
Accuracy			0.98	400

VIII.CONCLUSION

Many breakthroughs have been made in the field of artificial intelligence, machine learning and computer vision. They have immensely contributed in how we perceive things around us and improve the way in which we apply their techniques in our everyday lives. Many researches have been conducted on sign gesture recognition using different techniques like ANN, LSTM and 3D CNN. However, most of them require extra computing power . On the other hand, our research paper requires low computing power and gives a remarkable accuracy of above 90%. In our research, we proposed to normalise and rescale our images to 64 pixels in order to extract features (binary pixels) and make the system more robust. We use CNN to classify the 10 alphabetical American sign gestures and successfully achieve an accuracy of 98% which is better than other related work stated in this paper.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 12, December-2020

IX.PROBLEMS

Sign languages are very broad and differ from country to country in terms of gestures, body language and face expressions. The grammars and structure of a sentence also varies a lot. In our study, learning and capturing the gestures was quite a challenge for us since the movement of hands had to be precise and on point . Some gestures are difficult to reproduce. And it was hard to keep our hands in exact same position when creating our dataset.

X. FUTURE WORK

We look forward to use more alphabets in our datasets and improve the model so that it recognises more alphabetical features while at the same time get a high accuracy. We would also like to enhance the system by adding speech recognition so that blind people can benefit as well.

REFERENCES

[1] https://peda.net/id/08f8c4a8511
[2] K. Bantupalli and Y. Xie, "American Sign Language Recognition using Deep Learning and Computer Vision," *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018, pp. 4896-4899, doi: 10.1109/BigData.2018.8622141.
[3] CABRERA, MARIA & BOGADO, JUAN & Fermín, Leonardo & Acuña, Raul & RALEV, DIMITAR. (2012). GLOVE-BASED GESTURE RECOGNITION SYSTEM. 10.1142/9789814415958_0095.
[4] He, Siming. (2019). Research of a Sign Language Translation System Based on Deep Learning. 392-396. 10.1109/AIAM48774.2019.00083.
[5] International Conference on Trendz in Information Sciences and Computing (TISC). : 30-35, 2012.
[6] Herath, H.C.M. & W.A.L.V.Kumari, & Senevirathne, W.A.P.B & Dissanayake, Maheshi. (2013). IMAGE BASED SIGN LANGUAGE RECOGNITION SYSTEM FOR SINHALA SIGN LANGUAGE
[7] M. Geetha and U. C. Manjusha, , "A Vision Based Recognition of Indian Sign Language Alphabets and Numerals Using B-Spline Approximation", International Journal on Computer Science and Engineering (IJCSE), vol. 4, no. 3, pp. 406-415. 2012.
[8] Pigou L., Dieleman S., Kindermans PJ., Schrauwen B. (2015) Sign Language Recognition Using Convolutional Neural Networks. In: Agapito L., Bronstein M., Rother C. (eds) Computer Vision - ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science, vol 8925. Springer, Cham. https://doi.org/10.1007/978-3-319-16178-5_40
[9] Escalera, S., Baró, X., Gonzàlez, J., Bautista, M., Madadi, M., Reyes, M., . . . Guyon, I. (2014). ChaLearn Looking at People Challenge 2014: Dataset and Results. Workshop at the European Conference on Computer Vision (pp. 459-473). Springer, . Cham.
[10] Huang, J., Zhou, W., & Li, H. (2015). Sign Language Recognition using 3D convolutional neural networks. IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). Turin: IEEE.
[11] Jaoa Carriera, A. Z. (2018). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (pp. 4724-4733). IEEE. Honolulu.
[12] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). IEEE. Miami, FL, USA .
[13] Soomro, K., Zamir , A. R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. Computer Vision and Pattern Recognition, arXiv:1212.0402v1, 1-7.
[14] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: a large video database for human motion recognition. Computer Vision (ICCV), 2011 IEEE International Conference on (pp. 2556-2563). IEEE
[15] Zhao, Ming & Bu, Jiajun & Chen, C.. (2002). Robust background subtraction in HSV color space. Proceedings of SPIE MSAV, vol. 1. 4861. 10.1117/12.456333.
[16] Chowdhury, A., Sang-jin Cho, & Ui-Pil Chong. (2011). A background subtraction method using color information in the frame averaging process. Proceedings of 2011 6th International Forum on Strategic Technology. doi:10.1109/ifost.2011.6021252