Hand Gesture Recognition System using Convolutional Neural Networks

1st Raj Patel
Information Technology
Dwarkadas J. Sanghvi College of Engg
Mumbai, India
raj1471997@gmail.com

1st Jash Dhakad
Information Technology
Dwarkadas J. Sanghvi College of Engg
Mumbai, India
jashdhakad10@gmail.com

2nd Kashish Desai
Information Technology
Dwarkadas J. Sanghvi College of Engg
Mumbai, India
kashishdesai1997@gmail.com

2nd Tanay Gupta
Information Technology
Dwarkadas J. Sanghvi College of Engg
Mumbai, India
tanaygupta3010@gmail.com

3rd Prof. Stevina Correia
Information Technology
Dwarkadas J. Sanghvi College of Engg
Mumbai, India
stevina.dias@djsce.ac.in

***Abstract*—Gesture recognition plays an important role in communication through sign language. It is a fast growing domain within computer vision and has attracted significant research due to its widespread social impact. To tackle the difficulties faced by hearing impairment, it is the need of the hour to develop a system which translates the sign language into text which can easily be recognized by the impaired people. In this paper, a static hand gesture recognition system is developed for American Sign Language using deep Convolutional Neural Network. The system architecture is light weight to make the system easily deployable and mobile. In order to achieve high accuracy on live scenarios we employ, a number of image processing techniques which assist in appropriate background subtraction and frame segmentation. Our approach focuses on mobility,cost-free and easy deployment in low computational environment. Our system achieved a testing accuracy of 96%.**

***Index Terms*—Sign Recognition, Gesture Recognition, Computer Vision, Convolutional Neural Networks.**

I. Introduction

Communication is imparting, sharing and conveying of information, news, ideas and feelings. Of them, sign language is one of the way of non-verbal communication which is gaining impetus and strong foothold due to its applications in a large number of fields. The most prominent application of this method is its usage by differently disabled persons like deaf and mute people. They can communicate with non-signing people without the help of a translator or interpreter by this method. Some other applications are in the automotive sector, transit sector, gaming sector and also while unlocking a smart phone [1]. The sign gesture recognition can be done in two ways: static gesture and dynamic gesture [2]. While communicating, the static gesture makes use of hand shapes while the dynamic gesture makes use of the movements of the hand [2]. Our paper focus on static gestures. Hand gesture recognition is a way of understanding and then classifying the movements by the hands. But the human hands have very complex articulations with the human body and therefore a lot of errors can arise [3]. Thus it is tough to recognize the hand gestures. Our paper focuses on detecting and recognizing the hand gestures using different methods and finding out the accuracy by those methods. Also we see the performance, convenience and issues related with each method. Currently a lot of methods and technologies are being used for sign and gesture recognition. Among them the most common ones used are Hand Glove Based Analysis, Microsoft Kinect Based Analysis, Support Vector Machines and Convolutional Neural Networks. One of the objective of these methods is to bridge the communication gap between speech and hearing impaired people with the normal people and also successful and smooth integration of these differently abled people in our society. In our research paper we build a real time communication system using the advancements in Machine Learning. Currently the systems in existence either work on a small dataset and achieve stable accuracy or work on a large dataset with unstable accuracy. We try to resolve this problem by applying Convolutional Neural Network (CNN) on a fairly large dataset to achieve a good and stable accuracy.

II. Literature Survey

In order to bridge communication gap between hearing and speech impaired members, different approaches have been used by researchers for recognition of various hand gestures. These approaches can be broadly divided into three categories - Hand Segmentation Approach, Gesture Recognition Approach and Feature Extraction Approach.

Two categories of visual-based hand gesture recognition can be used.The first one is a 3-D hand gesture model that works by comparing input frames which makes use of sensors like gloves, helmet, etc[4].The other one is Microsoft Kinect based analysis which makes use of Kinect camera. Kinect hardware gives accurate tracking of several user joints. So a huge dataset is required for the 3-D hand gesture model since it requires a huge data set and also has a higher hardware cost due to

978-1-5386-6947-1/18/$31.00 ©2018 IEEE

sensors on the gloves. This glove based model for American Sign Language was proposed by Starner and Pentland [5]. It is not practically possible for the user to wear gloves continuously.

The 2-D hand gesture model make use of image dataset for feature extraction and detection. There are many other approaches used for image based gesture recognition like ANN (Artificial Neural Network), HMM (Hidden Markov Model), Eigenvalue based and Perceptual colour based. The feature vector extracted from the image are inputted into HMM [6]. For classification, particle filtering and segmentation methods like Support Vector Machine (SVM) is used where image frame is converted into HSV colour space as it is less sensitive to light effects[7]. Feature extraction can be employed using various methods. One of the most used method for feature extraction is by Contour Shape Technique which extracts the boundary information of the sign.

III. Currently used Methodologies

A. Feature Extraction

A feature is a function of one or more measurements computed so that it quantifies some significant characteristic of the object [9]. Feature extraction is a special form of dimensionality reduction. In pattern recognition and also in image processing,if the input given is quite large for processing,then it is suspected to be redundant and eventually the input data which is given will transform into a reduced representation set of features [8]. Feature extraction can be defined as a process of transforming input data in set of features. The general expectation is that the features set will extract the information which is relevant from the input data if we extract the features carefully in order to perform the desired task using this reduced representation instead of the full size input.

Some issues with feature extraction are, firstly ,the features should carry enough information about the image and should not require any domain-specific knowledge for their extraction [9]. Secondly, the features should be easy to compute in order to make feature extraction more feasible for a large image collection and rapid retrieval. Also, they should relate well to the human perceptual characteristics since users finally determine the suitability of the images retrieved.

B. Hand Segmentation Approach

Hand tracking and Segmentation should be always done in an efficient manner as they are the keys of success towards any gesture recognition, because of the challenges vision based methods pose such as intensity of the continuous variation in lightning, many objects in the background(complex) and detection of the skin color. Color is very powerful descriptor for object detection. Thus, color information was used for the segmentation purpose , which is invariant to rotation and geometric variation of the hand [10]. Human sees color component's features such as saturation, hue and the brightness component more than the percentage of primary colors which are red, green and blue [10]. These color models represent the standardized way of a particular color. It is a space-coordinated system in which any color which is speified is represented by single point. Here, using different color spaces for robust hand detection and segmentation, three techniques were introduced. Hand tracking and segmentation (HTS) technique using HSV color space is identified for the pre-processing of HGR system.

Some issues with hand segmentation are,firstly some objects, which are irrelevant,might overlap with the hand.Also,performance of the hand segmentation algorithm is degraded when the distance between the user and the camera is more than 1.5 meters [11]. Lastly,hand segmentation restricts the user to make some gestures in a particular manner,like gestures must be made with the right hand only, the arm should be vertical, the palm should face the camera and the background should be clear and uniform.

C. Glove based hand gesture recognition

Glove based approaches make use of gesture or capacitive touch sensors embedded into gloves to recognize hand gesture.The widely used methods make use of hand motion to convey hand signs and the motion is tracked and translated to text. Hand motions are categorized using clustering techniques such as k-means. Other approaches use charge-transfer touch sensors for translation by using On / Off binary signals. These approaches achieve high accuracy but incur high cost due to the necessary hardware.

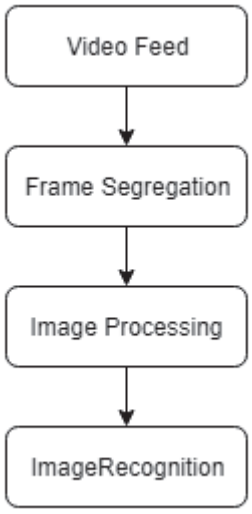


Fig. 1: System Block Diagram

IV. Proposed Methodology

In the view of the limitations posed by the approaches mentioned above, our system would focus on mitigating those incompetencies. The system to be built has to be capable to be able to be deployed on a mobile or web application for far reach and easy accessibility so, it has to be lightweight and computationally competent enough to recognize the signs appropriately.

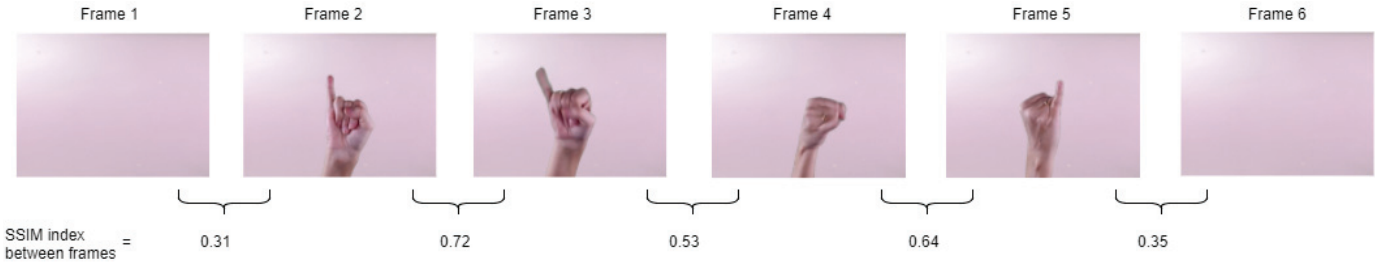


Fig. 2: SSIM between intermediate frames

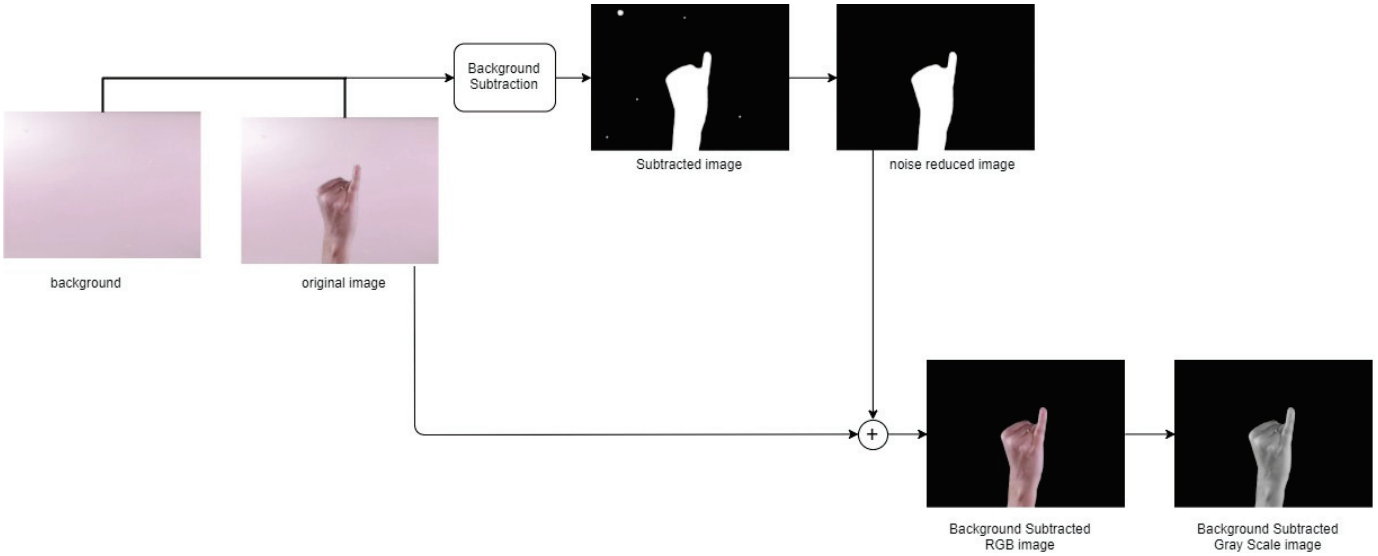


Fig. 3: Image Processing Steps

We propose a computer vision based approach to recognize static hand gestures. The system would analyze a video feed and recognize the hand gesture and then output the correct class label. For each sign performed by the user, the system will output 1 among 36 class labels comprising of ASL Gestures for alphabets and numbers. The system would take in a video feed which could be pre-recorded or coming live from an input device The system would segregate each sign and output the sign label accordingly. The major challenges identified were performing precise background subtraction and achieving high accuracy in order for the system to be used in formulating sentences. Background Subtraction required tackling changing illumination and foreground noise in input images. Operations such as Gaussian Mixture based segmentation and Image Morphology are performed in order to reduce background noise. Henceforth, our system architecture employs three phases viz. Frame segregation, Image Processing and Image recognition.Fig 1. illustrates the system block diagram

V. Implementation

A. Frame Segregation

Frame segregation is the first stage which involves identifying the frames which contain the sign gesture and segregating those frames for further processing and recognition. In order to extract each individual frame from the video feed, we perform frame by frame comparison by computing the Structural Similarity Index(SSIM) between two adjacent frames and based upon a threshold we factor in distinct frame selection. The video feed taken into consideration is captured from a webcam at a resolution of 900 * 900 and 23fps. The video feed is reduced down to 12 fps and the user is given an ROI to perform the sign gesture. Thus the final images after the cropping are of size 300 * 300. For two images I (i, k) and K(i, j) SSIM is calculated according to equation 1.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (1)$$

where, u_x is the average of x
u_y is the average of y
σx^2 is the variance of x
σy^2 is the variance of y
σxy is the co variance of x and y
$(c_1 - k_1L)^2$ and $(c_2 - k_2L)^2$ are two variables to stabilize the division with weak denominator
L represents the dynamic range of the pixel values and k_1=0.01 and k_2=0.03 by default [12].

An SSIM of 1 indicates perfect similarity. By testing for two images to be considered distinct the threshold value

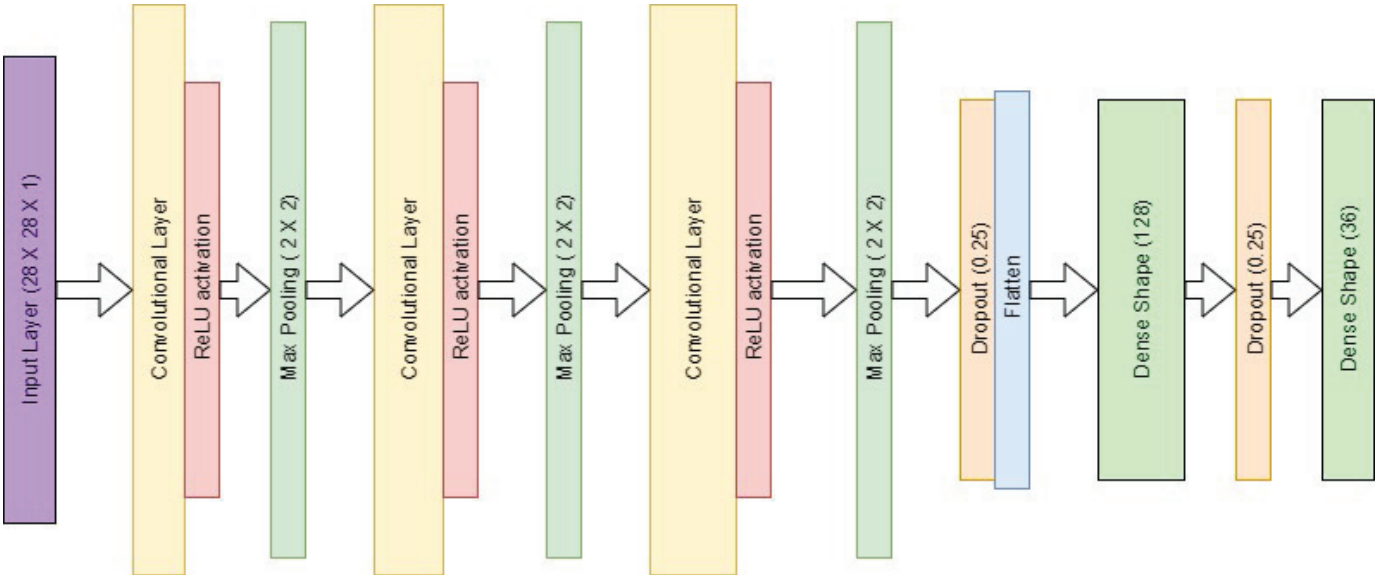


Fig. 4: CNN Architecture

for SSIM was identified to be 0.45. Fig 2 illustrates the SSIM between intermediate frames. As evident from the calculated values, the SSIM between two similar images is greater than the threshold value. For dissimilar frames such as frame 1 and frame 2, the SSIM calculated is less than the threshold value but since it transitions from background to foreground, frame 2 is not considered as distinct. To the contrary frame 5 is considered as a distinct frame moving from foreground to the background.

B. Image Processing

After the image segregation phase, the next phase is processing the output frames. The first step is to perform background subtraction. Since our system is mobile, the input images will vary a lot in terms of the background and the lighting conditions. Background subtraction is performed using Gaussian Mixture based background segmentation algorithm. It uses a mixture of K Gaussian distributions to model each background picture [13][14]. The output frames are compared against the background image and a resulting image is obtained after the background subtraction. Normal background subtraction was ruled out due to its low tolerance to dynamic conditions. Noise reduction was performed since some observable noise was present on the resulting image. The two main types of noise observed were spatial noise due to motion and salt and pepper noise due to change in lighting conditions. In order to remove spatial noise, low pass spatial filtering was used with a kernel of size 3 and for reducing other kinds of noise, morphological opening was performed on the subtracted image, with a structuring element of size 5. Morphological opening performs erosion followed by dilation which is useful in removing noise.

The resulting binary image after performing the operations is illustrated in figure 3. This resultant image wont yield high accuracy in image recognition tasks because of important positional features of hand and fingers being lost when background subtraction is performed. Thus, in order to retain those positional features, AND operation is performed with the original image and the noise reduced subtracted image. This results in the white pixels of binary image acting as a filter for the RGB image. After which, the resulting RGB image after addition is converted to gray scale. This is done to eliminate any bias due to the user skin tone or foreground lighting during recognition.

C. Image Recognition

The last phase in the system is the image recognition phase. In order to achieve higher accuracy as compared to existing systems and to keep the system computationally lightweight, we make use of Convolutional neural network (CNN) for image recognition. Convolutional neural network are a class of feed-forward artificial neural networks commonly used for visual analysis tasks. They comprise of neurons which act as learnable parameters having their own weight and biases. The entire neural network learns with the help of a loss function, a learning rate is used to fine tune the learning. The input layer takes in the data which gets propagated through the various layers and a output is generated, the generated output is compared with the actual output and the system updates its weights and biases to correct itself, this step is crucial and is known as backpropagation and this process done iteratively is called as training. The training duration of CNN is decided according to the size, the number of layers and also the learning rate. The CNN was trained using the ASL sign language image dataset consisting of around 35K images with each class having a minimum of 800 images. The dataset consisted of gray scale static sign images concerning alphabets and numbers. Fig 5. shows some images from the dataset. The character labels associated with each image were converted into binary vectors

using one hot encoding, thus converting categorical values into numbers.The proposed architecture of our CNN is illustrated in figure 4. It consists of three convolutional layers with 32, 64, 128 number of filters, having intermediate max-pooling layers and Relu activations. A kernel of size 3 and pool size of 2 was used accordingly. The last three layers consisted of a

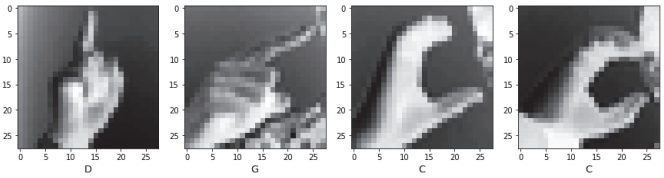


Fig. 5: Images from dataset

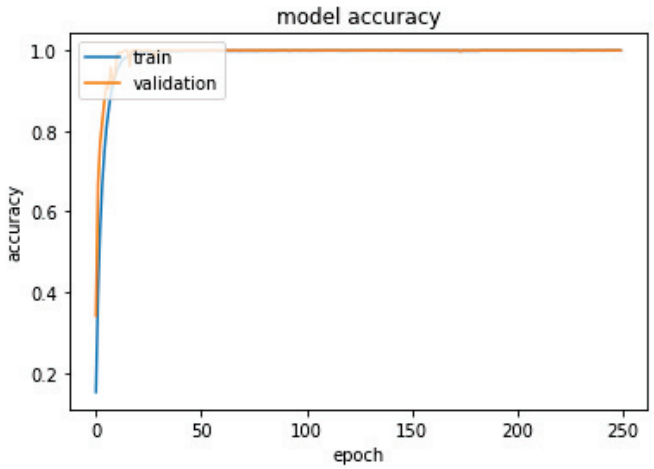
flattening layer and fully connected layers with dropout layers in between in order to avoid overfitting. The final dense layer is of size 36 corresponding to the number of class labels with softmax activation. The input to the CNN will be the gray scale processed image resized to 28 * 28 as per the dataset and the output of the CNN would be a probability distribution to classify the image into probabilistic values between 0 and 1. The loss function used for training was categorical cross entropy and the optimizer used was rmsprop. The training was conducted for 250 epochs with a batch size of 512. For any image feed into the CNN, it outputs a probability distribution. The node containing the highest probability value is considered as the output node and the correct label against that node is outputted. In this way the system determines what sign the user performed.

VI. Results

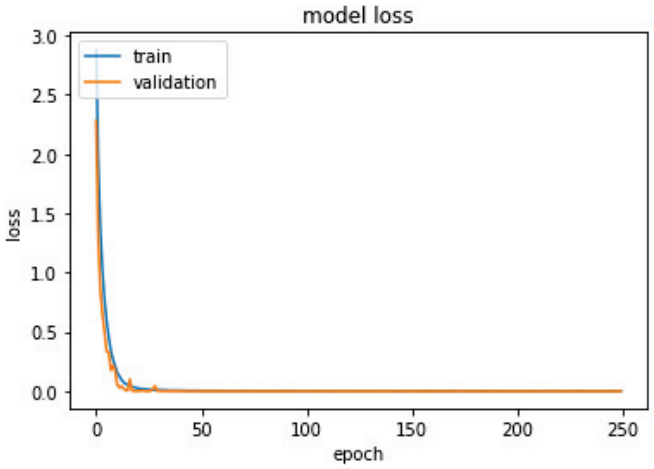
The recognition accuracy of the CNN obtained on the test set was 96.36% . Figure 6 illustrates the model accuracy and loss during the training and validation. From the figure it is evident that the model loss converges to almost zero, henceforth eliminating the case of under-fitting i.e. the model is capable enough to generalize to the dataset and also the test accuracy obtained rules out the suspicion of model being over-fitted i.e. the system is able to guess unseen data correctly.
Testing over a live setting, our system yielded 38 correct predictions out of a set of 40 trials, gaining a accuracy of 95%. Image segregation was able to pick out the distinct frames correctly and the system performed well even when the lighting conditions were changed. Depending on the constraints posed in the scenario of static gesture recognition, our system produced the highest results as compared to the existing systems out there which rely on feature extraction or employment of costly gloves. It can be deployed as an application on a minimal system and comes at a zero-cost.

VII. Conclusion and Future Work

In this study a system to classify static gestures was identified and implemented using Convolutional Neural Network. Our system is adaptive and performs robustly under varied lighting and background conditions. The proposed system is



(a) Model Accuracy



(b) Model Loss

Fig. 6: CNN performance graphs

a computationally low cost and can be deployed in an mobile setting while makes it suitable for real time applications. The research related to vision based gesture recognition is still in progress and our future research would be based on further improving the accuracy, expanding the classification dictionary and employing dynamic gestures for recognition.

References

[1] Gesture Recognition (2018, October, 4) Wikipedia [Online] Available:
[2] Priyanka C Pankajakshan, Thilagavathi B, Sign Language Recognition System, IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS15
[3] Jobin Francis and Anoop B K. Article: Significance of Hand Gesture Recognition Systems in Vehicular Automation-A Survey. International Journal of Computer Applications 99(7):50-55, August 2014.
[4] T. Starner and A. Pentland, "Real-time American sign language recognition from video using hidden Markov models", Technical Report, M.I.T Media Laboratory Perceptual Computing Section, Technical Report No. 375, 1995.
[5] Camastra, Francesco, and Domenico De Felice. "L VQ-based hand gesture recognition using a data glove." Neural Nets and Surroundings. Springer Berlin Heidelberg, 2013. 159-168.

[6] Lang, S., B. Marco and R. Raul. Sign Language Recognition Using Kinect. In: L.K. Rutkowski, Marcin and R. T. Scherer, Ryszard Zadeh, Lotji Zurada, Jacek (Eds.), Springer Berlin / Heidelberg,pp:394-402,20 11.
[7] V. K. Verma, S. Srivastava, and N. Kumar, "A comprehensive review on automation of Indian sign language," IEEE Int. Conf, Adv. Comput. Eng. Appl. Mar 2015 pp. 138-142
[8] Sanaa Khudayer Jadwaa, Feature Extraction for Hand Gesture Recognition: A Review, International Journal of Scientific Engineering Research, Volume 6, Issue 7, July-2015
[9] George Karidakis et al , Feature Extraction-Shodhganga
[10] Archana Ghotkar, Gajanan K.Kharate, Hand Segmentation Techniques to Hand Gesture Recognition for Natural Human Computer Interaction, International Journal of Human Computer Interaction
[11] Rafiqul Zaman Khan, Noor Adnan Ibraheem, Comparative Study of Hand Gesture Recognition System, SIPM, FCST, ITCA, WSE, ACSIT, CS IT 06, pp. 203213, 2012.
[12] Structural Similarity (2018, August, 27) Wikipedia [Online] Available:
[13] P.KaewTraKulPong, R.Bowden, An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection , In Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems, AVBS01. Sept 2001. VIDEO BASED SURVEILLANCE SYSTEMS: Computer Vision and Distributed Processing
[14] Background Subtraction, Open Source Computer Vision