International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-3, Issue-11 E-ISSN: 2347-2693

Interpretation of Indian Sign Language through Video Streaming

Juilee Rege[1], Ankita Naikdalal[2], Kaustubh Nagar[3*] and Prof. Ruhina Karani[4]

[1,2,3*,4] *Dept. of Computer Engineering, D.J Sanghvi College Of Engineering, India*

www.ijcseonline.org

Received: Oct/22/2015 Revised: Nov /04/2015 Accepted: Nov/17/2015 Published: Nov/30/2015

***Abstract*—** Sign Language is the language used by the deaf and dumb people to communicate. However, this language is rarely learnt by the general public. So it becomes difficult for these people to communicate with the general masses. Various such methods and techniques have been developed for the American Sign Language. This paper proposes an interpretation technique for the Indian Sign Language which is equally complex in nature and uses various parts of the body to convey messages such as hand orientations, palm movement, fingertips, etc. Our proposed technique will be able to take in a live video stream consisting of gestures and convert it into an equivalent sentence in English. The solution offered consists of steps like frame extraction, segmentation and refining of images, feature extraction, and training of neural network. The various methods will have different accuracy and efficiency levels and thus training of the network to perfectly guess each sign is of utmost importance.

***Keywords*—** Neural Networks, Video Processing, Indian Sign Language

I. INTRODUCTION

Normal people communicate their thoughts and ideas using their voice or speech. However, this is not the same for the dumb and deaf people. These people use sign language to communicate with each other. It consists of various gestures formed by hand, body, or facial expressions. This is used by deaf and dumb to express and communicate their thoughts. Such a language is not understood by the general public. This causes a big gap in communication between the deaf–dumb people and the normal people. Sign language interpreters are used for communication between these impaired people and normal people, which is not feasible all the time. This can cause a lot of problems. Such a system that automatically recognizes the sign language gestures and interprets it is necessary. This can help reduce the gap between these two groups in the society [1].

There are numerous sign languages which are used across the world. It depends on the location and culture of the place. Examples include American Sign Language (ASL) for United States of America, British Sign Language (BSL) for Britain, and Brazilian Sign Language for Brazil etc. Indian Sign Language (ISL) is the sign language which is used in India. It is a language in itself. Like ASL, it has its own grammar. This is very different from the grammar which the spoken languages use. ASL consists of ASL Phonology, ASL Morphology and ASL Tenses. Similarly, ISL consists of word level gestures and fingerspelling. Each of these has a significant importance and recognizing these is very important.

This paper proposes a system which would convert a video feed consisting of Indian Sign Language gestures to the corresponding sentence in English. Although projects have been developed for conversion of American Sign Language images to text, very few of them have been developed to convert Indian Sign Language images to its corresponding text [2]. The system aims to convert a video feed of the gestures to the corresponding text. The paper is arranged in the following manner. Section II will be a literature survey discussing previous projects and techniques used therein. Section III will discuss the proposed solution and the steps and techniques which will be utilized. Section IV will be the conclusion and talk about the future scope of the project. References will be mentioned at the end.

II. LITERATURE REVIEW

Various methods are used for conversion of Indian sign language to its corresponding text. Various methods are deployed, each having its advantages and disadvantages alike. Many such employed are for the conversion of ASL to text. The same methods could be applied to interpret the ISL as well.

The first method uses a video based approach for translating the sign to the equivalent sentence in English. It consists of 3 modules. The first module is the video processing module. This maps the signs to static images. Frames are extracted from the videos which form the dictionary. These videos from which frames are extracted consist of person enacting the ASL. The frames are also extracted from the input video which are then mapped with these frames in the dictionary. Folders are created to store images related to particular words. The Natural Sign Language module is used to generate a simple English sentence which conveys the corresponding meaning. The third module is used to convert this sentence to speech [3].

The second method uses an Artificial Neural Network to interpret the ISL. It consists of getting the image and Pre-processing it to refine it, segmentation of the hand area, extraction of features, and final classification. Image acquisition consists of capturing the images which represent the hand gestures of the ISL. This step is used to form the database for further processing. Hand segmentation is the next step which consists of extracting only the hand depicting the sign from the entire image. Various color segmentation models like RGB model, and YCbCr model is used for this purpose. Feature extraction is used for extracting the various features of the image. This is used to classify the image. The feature vector obtained from the feature extraction step is used as the input of the classifier that recognizes the sign. Artificial neural network is used as the classification tool [4].

A phoneme based method is also used to recognize the sign language. There are 44 phonemes in English language; therefore 44 gestures can be formed. There are 11 categories of gestures which are formed in ASL. The right hand shows categories, left hand shows sign in category. For prepossessing and filtering the image, RGB color space is used. RGB algorithm shows the head and hands region. The next step consists of using the vertical interleaving method for image compression. The features of the image are extracted using the 2-D moment invariant which are then fed to the neural network which recognizes the equivalent text [5].

Component of each of these methods have been used for the proposed system being developed.

A. Segmentation of Image:

Pixels make up the image which are formed. Image segmentation is the activity in which pixels are clustered on which various operations are performed. The main aim is to simplify the image for further processing. Labels are used to group pixels which share the same characteristics. This is done to find object boundaries in the images.

Various colour based systems are used:

- RGB Model: It is an additive colour model in which red, green and blue colours are added together in various ways to produce a large range of colours. The rules to design this are very simple and time required to compute is very less when it is compared to all other models. This is why this is a preferred method and is widely used [6]. The skin region rule and the cluster boundary are specified are defined by this. Using these three colours, the skin colour is detected [7]. The algorithm for the same is as follows:

1. Acquire RGB image frame
2. From each pixel the R, G and B components are separated.
3. do Step 4 if R>95 and G>40 and B>20, else it is not a skin region
4. If Max{R, G, B}-Min{R, G, B}>15, do Step 5 else it is not a skin region
5. If IR-GI>15 and R>G and R>B it is a skin region else it is not a skin region

This algorithm is applied to each of the frames and they are segmented [5,7].

- HSV Model: HSV is a common cylindrical coordinate representation of points in an RGB Colour model. It basically tries to rearrange the geometry of RGB. The full form of HSV is hue, saturation and value. This can be used for segmentation. For this purpose, the luminance is disregarded and indices are provided for the 2D histograms by the Hue and saturation factors. An NxN matrix is formed in which each of the indexes are associated with the hue and the saturation. At each wanted segmentation, the location is incremented in the matrix while it is decremented when a background colours are identified. This helps to form the colour predicate. Any colour locations in the matrix that have values greater than a threshold values are left in the image and all others are replaced with black pixels. In such a way, hand images can be segmented from the rest of the unneeded picture.

- YCbCr Model: YCbCr is a family of colour spaces used as part of the colour image pipeline in videos. Y is the luma component i.e. light intensity is non-linear RGB primaries, Cb and Cr are the blue-difference and red-difference chroma components. It is a way of encoding RGB information. RGB values are transformed to YCbCr colour space using the following equations [8]:

$$Y = 0.299R + 0.587G + 0.114B \quad (1)$$
$$Cr = 128 + 0.5R - 0.418G - 0.081B \quad (2)$$
$$Cb = 128 - 0.168R - 0.331G + 0.5B. \quad (3)$$

The pixels in the input image which are skin coloured are identified by a threshold value which is decided on the skin colour distribution in this model. Each pixel is transformed to black or white according to the values. If all three components are within the specified range then the pixel is set to white otherwise it is changed to black. This gives us a binary image. Thus, compared to the RGB model, YCbCr colour space is luminance independent and has separate luminance and chrominance components which is why it gives a better performance.

- CIELab Model: It is a uniform color space. Perceptual uniformity defines the difference between the two colors when seen by humans.

This uniformity in these color spaces is obtained due to heavy computational transformations. In these color spaces, the computation of the luminance (L) and the chroma (uv) is found through a nonlinear mapping of the XYZ coordinates. In CIELab, the three components represent luma component that is illumination information and ab represent the chroma information. L*=0 gives the black color and L*=100 gives the white color .The a* values a*<0 that indicate green while the values a*>0 indicate magenta. The b* values b*<0 that indicate blue and values b*>0 indicate yellow.

The CIElab model is better than the YCbCr model as it provides a better range of colours to deal with than all the models [9].

Types of segmentations are:

- Thresholding: It is the simplest method. It is based on a method to convert the grayscale image to a binary image. The main operation is to choose an appropriate threshold to separate the wanted and unwanted details.

- Clustering methods: The K-means algorithm is an iterative process used to partition an image. The algorithm is given as:

1. Pick K cluster centers on some basis.
2. Each pixel is assigned to one of the clusters. This is done based on proximity.
3. Calculate the cluster centers by calculating mean of all pixels in the cluster.
4. Repeat above steps until convergence happens.

- Histogram-based method: This method is more efficient as all the information is acquired after one pass through the pixels. The clusters are located using the peaks and valleys while the pixels are used to computed a histogram. Colour and intensity can also be used as a measure. One drawback for this method is that it fails to recognize the peaks and valleys which are significant for use.

B. Feature Extraction

Once the image has been cleaned and segmented, important features need to be extracted from it to feed into the network in order for them to be classified. It is needed to reduce the amount of resources to describe a large set of data. Shape is an important visual feature of the image. Many different methods are used to represent shapes.

- Method 1 - Distance Transformation: It is a derived representation of a binary image. This transformation gives a new image in which each pixel value is replaced by the least distance between that pixel and the background pixel. This method gives us an image which represents the distance from the closest boundary pixel for each of the grayscale intensity.

1. Due to the invariant nature of Euclidean distances to the rotation of the image, these distances are considered. The Euclidean distance transform is obtained by applying a square root operation over the squared Euclidean distance transform matrix.

$$d_e(P,Q) = \sqrt{(x-u)^2 + (y-v)^2} \quad (4)$$

2. Projections of distance transform coefficients: It calculates the row and column projection vectors which are 1D functions of the above result. It works as follows: Find sum of pixel values in rows and columns of above result returning two vectors. One is row vector R where each element is sum of non-zero pixel values of the corresponding row and column vector C where each element is sum of non-zero pixel values of the corresponding column.

3. Fourier Descriptors: The Fourier descriptors are formed using the coefficients of the shape descriptors. Frequency domain is represented by these. They have strong discrimination power and noise sensitivity is overruled present in the shape representation. They are also information preserving and can be normalized easily.

For the two vectors the DFT are:

$$u_n = \frac{1}{N}\sum_{t=0}^{N-1} R(t)\, exp\left(\frac{-j2\pi nt}{N}\right) \quad (5)$$

$$where\ n = 0,1,2 \ldots N-1$$
$$and\ N\ is\ the\ size\ of\ R$$

and

$$v_n = \frac{1}{N}\sum_{t=0}^{N-1} C(t)\, exp\left(\frac{-j2\pi nt}{N}\right) \quad (6)$$

$$where\ n = 0,1,2 \ldots N-1$$
$$and\ N\ is\ the\ size\ of\ C$$

4. Feature Vector: The Fourier descriptors of the row and column projection vectors are used to form the feature values by considering only the magnitude of the Fourier coefficients and ignoring the phase information. These are normalized by dividing the magnitude of the fourier coefficients by the value of the first coefficient known as the dc component.

- Method 2 - 2D Moment Invariants: It is an effective method to extract features from an object for classification. The algorithm derives a number of self-characteristic properties from a

binary image of an object. These are invariant to rotation, translation and scaling [10].
The 2-D moment of order (p+q) of a digital image f(x, y) of size M x N is defined as

$$m_{pq} = \sum_{x=1}^{M-1} \sum_{y=1}^{N-1} x^p y^q f(x,y) \quad (7)$$

where p = 0, I, 2.... and q = 0, I, 2.... are integers.

The related central moment of order (p+q) is defined as

$$\mu_{pq} = \sum_{x=1}^{M-1} \sum_{y=1}^{N-1} (x-x')^p (y-y')^q f(x,y) \quad (8)$$

for p = 0,1,2, ... and q = 0,1,2, …
where $x' = \frac{m_{10}}{m_{00}}$ and $y' = \frac{m_{01}}{m_{00}}$

The normalized central moments, denoted η_{pq} are defined as

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \quad (9)$$

where $\gamma = \frac{p+q}{2} + 1 \; for \; p+q = 2,3 \ldots$

Moment invariants derived from second order moments:

$$\emptyset_1 = \eta_{20} + \eta_{02} \quad (10)$$

In the same way, the Hu set of invariant moments are derived.

- Method 3 - Discrete Cosine Transform (DCT) Transformation: This method transforms an image from the spatial to the frequency domain, where the image is decomposed into the combination of various frequency components [11]. Let image f(x, y) be represented as f(m,n) of size MxN. The 2D DCT of an image f (m, n) is given as:

$$B(p,q) = \alpha(p)\alpha(q) \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m,n) \cos\frac{\pi(2m+1)p}{2M} \times \cos\frac{\pi(2n+1)q}{2N} \quad (11)$$

p = 0….M-1 and q = 0….N-1
where p and q denote the frequencies.

$$\alpha(p) = \{\frac{1}{\sqrt{M}}; p = 0 \; \sqrt{\frac{2}{M}}; p = 1 \ldots M-1 \quad (12)$$

$$\alpha(q) = \{\frac{1}{\sqrt{N}}; p = 0 \; \sqrt{\frac{2}{N}}; p = 1 \ldots N-1 \quad (13)$$

where M and N denote the row and column size of f(m,n). The transform coefficient B(0,0) represents the average value of the input sequence and is denoted by DC coefficient while all other transforms are denoted by AC. The DCT is thus applied on grayscale images and features are extracted.

III. PROPOSED SOLUTION

Initially, the user of this system will enter his details. Since this system uses a video based approach, the user will record a video in ISL using a camera provided by the system or any other hardware such as webcams. The application will then convert the video into a sentence in simple English. This system will enable him to communicate with others without the need of a human interpreter.

Researching all the above techniques to carry out the various stages of the project, we have proposed the techniques which we will use in our solution.

1. Frame extraction: The video which we have recorded will be saved in a pre-determined folder. The video will be taken from here and frames will be extracted. An interval is defined to extract the required frames.

2. Segmentation: This step attempts to simplify an image by partitioning it into multiple segments of sets of pixels. We will use the CIELab color model because it carries out fast segmentation and it is an absolute color space so it defines the colors exactly without depending on input devices. It also includes more colors than other color spaces.

3. Feature Extraction: This step will extract the important features from the images to feed them to the neural network. Once the image has been cleaned and segmented, important features need to be extracted from it to feed into the network for them to be classified. It is needed to reduce the amount of memory and speed required to classify the data.

4. Neural Network: This step uses the extracted features as an input. We are using the Error Backpropagation Training Algorithm to train the network. This is because it will take a lot of training for the network to classify the images into the correct signs and the errors in each step have to be continuously reduced to get a good success

rate. The errors are propagated backwards from the classifier so that the network can learn.

Our output is expected to be a sentence in English, in which the words correspond to the interpreted signs classified by the machine.

IV. CONCLUSIONS AND FUTURE SCOPE

A neural network based method for automatically recognizing the Indian sign language gestures is proposed in this paper. The signs will be identified by the features extracted from the hand shapes. We propose to use skin colour based segmentation to extract the hand region from the image. The features that will be extracted from the sign image will be used to train a feedforward neural network that recognizes the sign. The method will be implemented completely by making use of digital image processing techniques so the user does not have to wear any special hardware device to get the features of the hand shape. Our proposed method aims to achieve low computational complexity and very high accuracy when compared to the existing methods. This project has a broad scope. Initially we are developing this project as a desktop application. This application can be installed on laptops, PC's and all-in-one computers and thus can be used in schools where deaf and dumb children can express themselves to normal people. It can also be used at public places such as malls, offices and railway stations. At such locations, this software can be installed on public computers which would assist disabled people to communicate with others with ease. In the future, mobile applications can be designed to utilize this facility on the move.

References

[1] David M. Perlmutter, "The Language of Deaf", The New York Review of Books, March **1991**.

[2] Geetha M and Manjusha U. C, "A Vision Based Recognition of Indian Sign Language Alphabets and Numerals Using B-Spline Approximation", International Journal on Computer Science and Engineering (IJCSE), March **2012**.

[3] Aradhana Kar and Pinaki Sankar Chatterjee, "A Video-based Approach for Translating Sign Language to Simple Sentence in English", Proc. of Int. Conf. on Advances in Computer Science, AETACS, **2013**.

[4] Adithya V, Vinod P.R and Usha Gopalakrishnan, "Artificial Neural Network Based Method for Indian Sign Language Recognition", Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT), **2013**.

[5] Paulraj M P, Sazali Yaacob, Mohd Shuhanaz Zanar Azalan and Rajkumar Palaniappan, "A Phoneme Based Sign Language Recognition System using 2D Moment Invariant Interleaving feature and Neural Network", IEEE Student Conference on Research and Development, **2011**.

[6] J. Yang, W. Lu and A. Waibel, "Skin-color modeling and adaptation", ACCV98, **1998**.

[7] Yona Falinie bte Abdul Gaus, Farrah Wong and Kenneth teo, "Malaysian Sign Language Recognition Using Neural Network", Proceedings of 2009 Conference on Research and development (SCOReD), **2009**.

[8] Amit Kumar Mandal and Dilip Kumar Baruah, "Image Segmentation Using Local Thresholding And Ycbcr Color Space", Int. Journal of Engineering Research and Applications, Vol. **3**, Issue **6**, Nov-Dec **2013**.

[9] Amanpreet Kaur and B.V Kranthi, "Comparison between YCbCr Color Space and CIELab Color Space for Skin Color Segmentation", International Journal of Applied Information Systems (IJAIS) – ISSN : **2249-0868** Volume **3**, No.**4**, July **2012**.

[10] William K.Pratt, "Digital Image Processing-PIKS Scientific Inside lh ed ", A Wiley-Interscience publication.

[11] Khadidja Sadeddine, Fatma Zohra Chelali and Rachida Djeradi, "Sign Language Recognition using PCA, Wavelet and Neural Network", Control, Engineering & Information Technology (CEIT), 2015