MAT JOURNALS

Journal of Image Processing and Artificial Intelligence
Volume 4 Issue 1

Hand Gesture Recognition using Gradient based Key Frame Extraction

Vibhav Joshi, Rushikesh Hiray, Sakshi Kale, Shweta Mantode
Department of Computer Engineering, KKWIEER, Nashik
Email: joshivibhav1@gmail.com, rhiray1996@gmail.com, sakshikale871@gmail.com, samantode@gmail.com

Abstract

Sign language is a mainstream method of communication by people suffering from hearing and speech impairment and other disabilities. Around 360 million people globally are suffering from disabled hearing loss. Minimizing the communication gap between differently-abled and normal people becomes a necessity to ensure effective communication among all. In this paper, a method for developing a system which recognizes cue symbols of sign languages to act as a communication medium between healthy and differently-abled people has been proposed. Two standard sign languages are considered for reference, as, cue symbols vary for each region. The system includes a hand gesture recognition system, where gestures are given as input and the corresponding output is in the form of text as well as speech. We have used gradient value for splitting two or more gestures from a continuous video sequence and extracting key frames. Features of pre-processed key frames are extracted using OH (Orientation Histogram) and dimensions of these features are reduced using PCA (Principal Component Analysis). SVM (Support Vector machine) classifier is used for classification of the isolated pre-processed gestures. After the classification process, corresponding connotation of gestures given as input is obtained.

Keywords: *Gesture, gradient, key-frame, American Sign Language (ASL), Indian Sign language (ISL), Orientation Histogram (OH), Principal Component Analysis (PCA), Support Vector Machine (SVM).*

INTRODUCTION

Movement of any part of the body, particularly hand or head, to convey a feeling or any idea is called as gesture. Sign language is a way of communication with the help of gestures. It involves use of facial expressions, orientations and movement of hand, finger spellings, body languages, head movements and head gazes. Communicating with the world using sign language is a widely accepted way by differently-abled people. There is no awareness about sign language in the society because of which a communication barrier exists between differently-abled people and normal people. We have developed this system with an aim of bridging the existing barrier and creating equal opportunities for the deprived ones. This system stands on the pillars of Image Processing and Computer Vision. This approach eliminates the need of human translator for explaining the connotation of sign languages as it enables computer to understand what human wants to convey through gestures. It enables human to interact with computer naturally, without using any mechanical device.

Here, computer is trained to recognize static hand gestures (one handed or two handed) from Indian Sign Language (ISL) and American Sign Language (ASL). Working of the system is explained sequentially using 5 steps. First step is Pre-processing in which frames are extracted from a continuous video and further tasks are performed to distil only hand portions

in them. Second step is key-frame extraction in which key frames are extracted by calculating the gradient value in both X and Y direction. Sobel operator is used for performing these calculations. Uninformative frames are also removed in this step. Third step explains feature extraction using Orientation Histogram (OH) algorithm by using histogram of oriented gradients. The extracted features are of large dimensions. Principal Component Analysis (PCA) is used to reduce these dimensions. Fourth step is classification in which the machine is trained to recognize gestures using the features of reduced dimensions using Support Vector Machine (SVM) algorithm. Connotation of the gesture is obtained in text form. Fifth step involves converting the output obtained in text form into speech.

Literature Survey

There are some other methodologies developed for meeting the same goals as the method proposed in this paper.

Vidur S. Bhatnagar et al. [2] have proposed a method which uses a glove with embedded sensors in it along with statistical template matching technique that allows gesture recognition and accordingly the corresponding output is generated in speech form. It requires the person performing gestures to wear the gloves continuously and stay near the system.

Author Jung-Bae Kim et al. [3] have proposed a system that uses Fuzzy Logic and Hidden Markov Model for Korean Sign Language recognition. Using these methods, they have obtained 94 percent accuracy for 15 KSL sentences. They have rejected meaning less gesture motions such as preparatory motion and useless movement between sign words, using fuzzy partitioning and state automata. They have mainly concentrated on two features like speed and velocity of hand motion. But this system recognizes a whole sentence and not individual gestures in a sentence. Proposed method is complicated and the system has high computational burden.

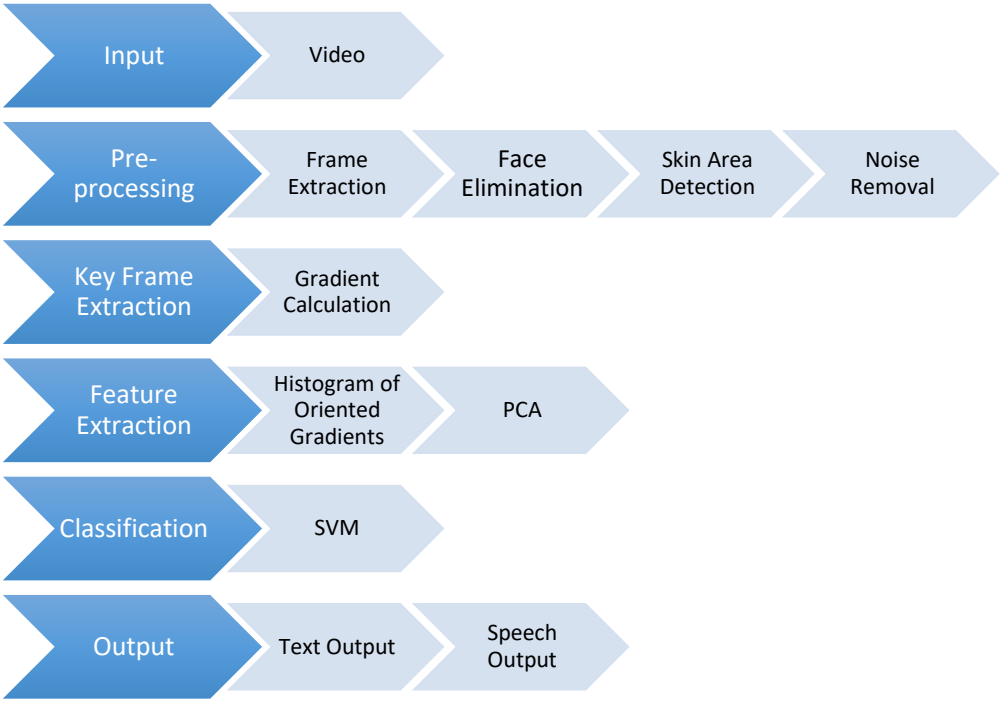
Lalit Kane et al. [4] uses Sign animation generated by the given input, to facilitate and reduce the communication gap between a healthy and an impaired person.

Matheesha Fernando et al. [5] has used a method involving feature detection based on contours. It is identified with the help of template matching. The system matches each symbol with their corresponding template. There is overhead for storing matching template of each gesture.

The system proposed in this paper focuses on removing limitations of translator dependency, ease of implementation, use of hardware and cost effectiveness. Detailed explanation for each step involved in recognizing gesture using gradient based key frame extraction is given in further sections.

Proposed Methodology

Dataset consists of pre-recorded videos of one handed or two-handed gestures from ISL and ASL. The videos can be recorded from mobile camera, webcam or any other camera device. Video for each sentence is recorded multiple times for better accuracy. The classifier is trained by processing these videos after giving as input. Dataset obtained from Robita, IIIT Allahabad [6] containing important 23 gestures of ISL is also used for training system for classification purpose.



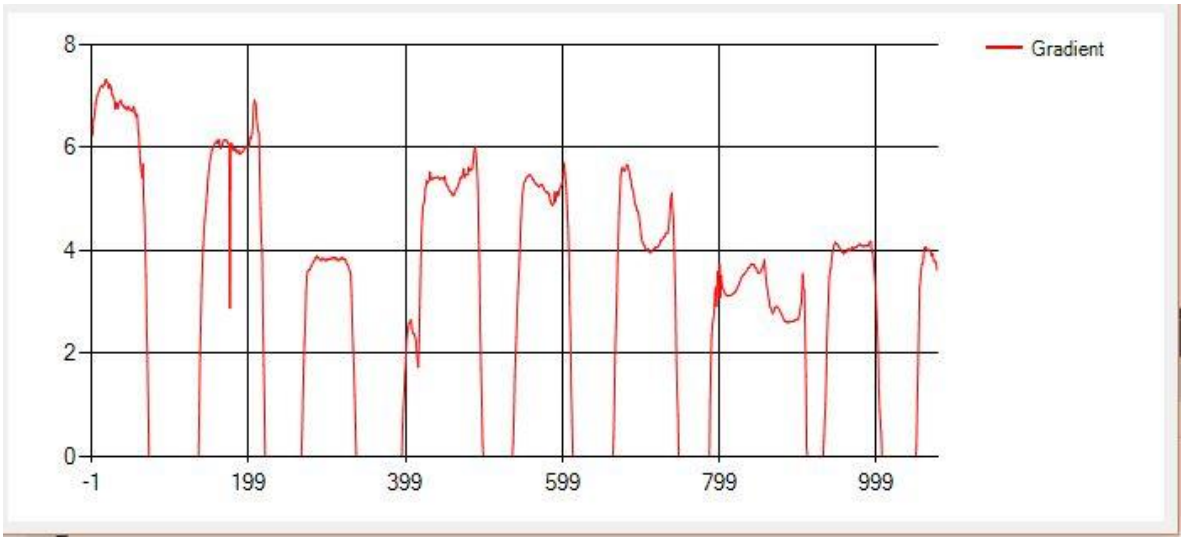
Block Diagram

Step 1: *Pre-processing*
Pre-processing is used for converting raw input into meaningful frames that can be used to identify the key frames.

In this step, each video given as input is converted into sequence of RGB frames. The face region in the frames is detected and eliminated as we only require the hand region showing gestures. The skin area and non-skin area is distinguished from these frames by using hue, saturation, and RGB thresholds. The values for thresholds are as mentioned below:
0.0<hue<50.0
0.1<saturation<1.0
R>95 & G>40 & B>20
The region satisfying the above threshold values is marked as gray and remaining area is marked as zero. Median Blur technique is applied which preserves the edges (outer boundary) of the hand gestures. It also helps in reducing unwanted noise. Thus, we obtain segmented region from frames containing only hand gestures with negligible noise present in it.

Step 2: *Key Frame Extraction*
A single gesture is formed by a sequence of large number of frames. Not all of these frames are useful, in fact most of these frames are non-informative frames containing preparatory gestures or post gesture frames. Thus, it is vital to extract the key frames for each gesture. We use the gradient value of each frame to identify the key gestures.

Gradient value is found using Sobel operator in both, X and Y direction for all the frames obtained from pre-processing step. Recognizing difference between two or more gestures in a continuous video sequence is a challenging task. We made it feasible using the values of gradient.



You can see that for frame number 73 to 130, the value of gradient is 0. This indicates end of one gesture (1st gesture in this case). Value of gradient for frame number 130 and onwards is non-zero till frame number 210. Here, frame number 130 indicates start of another gesture (2nd gesture in this case). As frame number 1 to frame number 72 contains one gesture, we consider middle 15 frames for extracting features as they contain most of the information about that gesture among all 72 frames. Same procedure is applied further for all gestures present in the input and key frames are extracted.

Step 3: *Feature Extraction*

All cue gestures have a unique set of features that can be used to differentiate between different gestures. These features can be used to classify the different gestures.

Features of the key-frames are extracted using Orientation Histogram algorithm. Here, we concentrate more on gradient because it is an important feature which helps distinguishing two gestures. Change in intensity of pixels is maximum at edges and corners. We get the features required for classification by calculating histogram of oriented gradients.

The steps followed are as given below:

1. The key frames are subsampled into dimensions 64*128. Thus, size of each input frame is 64*128*3 (3 channels).
2. Gradient value of each key frame is found using Sobel operator in X and Y direction.
   Magnitude = $\sqrt{gx^2 + gy^2}$ Angle= $\tan^{-1}\frac{gy}{gx}$
   ('gx' and 'gy' are gradient values in X and Y direction respectively)
3. The frame is divided into 8*8 blocks and histogram of oriented gradients is calculated for each of the block by writing 8*8*2 (2 because magnitude and angle of gradient) i.e. 128 numbers in 9-bins.
4. The frame is then divided into 16*16 blocks and feature vector of length 3780 (for image size 64*128) is obtained.
5. As dimensions of the obtained feature vector are large, Principal Component Analysis (PCA) is used to reduce them.

Step 4: *Classification*

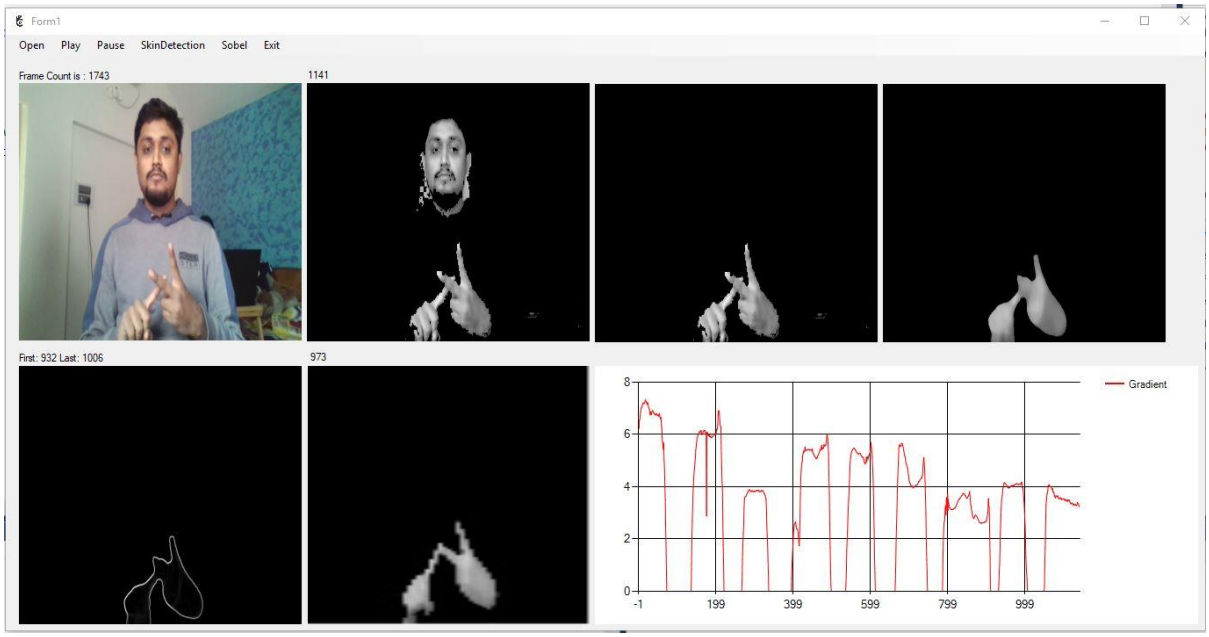
We use the features extracted in the previous step to classify the gestures based on the test data sets. These gestures are then converted to their respective text equivalent.

Support Vector Machine algorithm is used for classification process. Features (with reduced dimensions) obtained from the

previous step are given to the classifier for training. Further, the system provides correct connotation of the gesture given as input. Output obtained in this step is in text form.

Step 5: *Text to Speech conversion.*
Here, the text output received from classification step is given as input. The text is analysed, processed, and output is given in the form of speech.

RESULT



The first picture box shows the original video given as input. The total number of frames in the video is 1743 as shown in the image. The second box shows the skin region of the current frame. The current frame being processed is 1141 as shown. The third box eliminates the face. We use median blur operation in the fourth box for noise removal and to preserve the edges of the hands. The fifth box shows the edges of the gesture using Sobel operator. The last box shows the key frames extracted by calculating the gradient value of each frame. The Graph plots the gradient value of each frame with time or frame number. We use SVM for training and classification to recognize the gestures and get their text equivalent.

CONCLUSION

This system successfully recognizes the gestures given as input (video) and their connotation is provided in text and speech form. It helps Difficult task of differentiating between two or more gestures in a continuous video sequence has become feasible by calculating gradient values of the video frames. Orientation Histogram (OH) helps extract exact required features for classification. Principal Component Analysis (PCA) has helped reduced the overhead of classifier by reducing the dimensions of extracted features. Support Vector Machine (SVM) algorithm trains the system with extracted features and provides appropriate output to user.

Performance of the system can be amplified by providing input data with variance in background conditions. Adaptive nature and accuracy of the system can be enhanced by using Artificial Neural Network (ANN) for classification and gesture recognition. The system can be used in institutions to learn sign languages

(ISL and ASL). It can be used to create equal opportunities for all in the society.

ACKNOWLEDGEMENT

We are grateful to our guide, Prof. Priti Vaidya for her valuable guidance for this research work. We would also like to show gratitude towards our institution K. K. Wagh Institute of Engineering Education and Research, Nasik and Head of Computer Engineering Department Prof. Dr S. S. Sane for providing us the platform required for the work. We would also like to thank our colleagues for their useful insights towards this work.

REFERENCES

1. Kumud Tripathi∗, Neha Baranwal and G. C. Nandi, Continuous Indian Sign Language Gesture Recognition and Sentence Formation, *Robotics and Artificial Intelligence Lab, Indian Institute of Information Technology, Allahabad,* Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015).
2. Vidur S. Bhatnagar, Rachit Magon, Rashi Srivastava, Manish K Thakur "A Cost Effective Sign Language to Voice Emulation System" in Proc. 8th International Conference, 2015.
3. Jung-Bae Kim, Kwang-Hyun Park, Won-Chul Bang and Z. Z. Bien, Continuous Gesture Recognition System for Korean Sign Language based on Fuzzy Logic and Hidden Markov Model, *IEEE International Conference on Fuzzy Systems, 2002, FUZZ-IEEE'02, Proceedings of the 2002*, vol. 2, pp. 1574, 1579 doi: 10.1109/FUZZ.2002.1006741, (2002).
4. Lalit Kane, Pritee Khanna "Towards Establishing a Mute Communication An Indian Sign Language Perspective" in IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction, Kharagpur, India, December 27-29, 2012.
5. Matheesha Fernando, Janaka Wijayanayaka "Low-cost approach for Real-time Sign Language Recognition" in 2013 IEEE 8th International Conference on Industrial and Information Systems, ICIIS 2013, Aug. 18-20, 2013, Sri Lanka.
6. Robita, IIIT Allahabad.