Mobile Application Based Translation of Sign Language to Text Description in Kannada Language

https://doi.org/10.3991/ijim.v12i2.8071

Ramesh M. Kagalkar(✉)
Visvesvaraya Technological University (VTU), Belgaum, Karnataka, India
rameshvtu10@gmail.com

Shyamrao V Gumaste
MET League of College, Nashik, Maharashtra, India

Abstract—Sign language is a main mode of communication for vocally disabled. This language use set of representation which is finger sign, expression or mixture of both to express their information among others. This system presents a novel approach for mobile application based translation of sign action analysis, recognition and generating a text description in Kannada language. Where it uses two important steps training and testing. In training set of 50 different domains of video samples are collected, each domain contains 5 samples and assign a class of words to each video sample and it will be store in database. Where in testing test sample under goes preprocessing using median filter, canny operator for edge detection, HOG for feature extraction. SVM takes input as a HOG features and predict the class label based on trained SVM model. Finally the text description will be generated in Kannada language. The average computation time is minimum and with acceptable recognition rate and validate the performance efficiency over the conventional model.

Keywords—Gesture recognition, Image processing, Sign language, Video processing.

1 Introduction

The activity recognition aims to recognize the actions and goals of one or more agents from a series of observations on the agents' actions and the environmental conditions. Since the 1980s, this research field has captured the attention of several computer science communities. Due to its strength in providing personalized support for many different applications and its connection to many different fields of study such as medicine, human-computer interaction, or sociology. In image processing, input is taken as an image later perform processing on that image based on the requirement. Many types of input are to be taken in image processing such as a video, image or collect frames from video and after output is produced in the form of an image or set of parameter related to image. The use of image processing for improve the image quality and gather useful information in the image this process is called as

feature extraction. This image processing techniques can be used for detecting the hand gesture or analyzing actions or many other purposes in various fields. In the system developed these image processing techniques are used for communication purpose in such community where people are vocally impaired and hearing impaired.

The communication by the body language is understood as the fundamental trademark include that characterizes a hard of hearing group. The critical piece of communication via gestures acknowledgment plots in general society human progress is to guarantee that hard of hearing people have correspondence of chance and full commitment in the public arena. Communication via gestures is spoken to fundamentally by ceaselessly changing diverse hand shapes and development by an underwriter. Sign based correspondence is a physical action by using arms, hands, fingers and eye with which we can talk with idiotic and in need of a hearing aide people. Seeing human action from picture game plans is a champion among the most troublesome issues in computer vision with various imperative applications, for instance, tunning video perception, content-based video recuperation, human-robot collaboration, and splendid home. The errand is troublesome not in light of between class assortments, camera advancements, establishment confusing and fragmentary obstacle, also to some between class spreads and similarities, for instance, running as opposed to running or walking. Earlier tackles human movement affirmation in video frequently used overall representations.

The dynamic signals affirmation applications require the picking up of a high data rate of hand positions by and large gave using development taking after gloves that are set up to do unequivocally recording finger joint developments through flex sensors in an immovably fitting glove. Hand flag gives a trademark and characteristic correspondence philosophy for human–computer participation. Successful human computer interactions (HCIs) must be created to allow computer to ostensibly see persistently hand movements. In any case, vision-based hand taking after and movement affirmation is a trying issue due to the capriciousness of hand signs, which are rich in diversities on account of high Degrees of adaptability (DOF) required by the human hand. In order to viably fulfill their part, the hand movement HCIs need to meet the requirements with respect to progressing execution, affirmation precision, and vigor against changes and jumbled foundation.

The gesture based communication correspondence understanding includes semantic examination of hands taking after, hands shapes, hands presentations, sign verbalization moreover with basic etymological information talked with head advancements and outward appearances. Motion based correspondence is from different points of view assorted structure talked tongue, for instance, facial and hand stating, references in virtual checking space, and syntactic complexities as illuminated. The huge inconvenience in signal based correspondence affirmation stood out from talk affirmation is to see at the same time assorted correspondence properties of a guarantor, for instance, hands and body advancement, external appearances and body act. These property must be viewed as at the same time for a better than average affirmation structure. The second huge issue went up against by motion based correspondence affirmation structure engineers is taking after the underwriter in the confuse of other information by researchers to describe a model for spatial information containing the

components made in tail able in the video. This is studied by many experts as checking space. An imperative test stood midst of the correspondence through marking talk. This paper addresses the problem of action recognition that is how to determine the type of action that is happening in a video. Here the problem of video representation is considered that means how to encode videos in a robust way? Which type of representation is suitable for a wide variety of action classes, tasks and video types? This paper shows the system which is used for recognition of hand gesture of sign language and translate it into corresponding in Kannada language. Hence the proposed system has wide scope for vocally disable individual to express the feelings and talent on paper, who is residing in rural areas of Karnataka state of India.

The rest of the paper is organized as in section 2 gives the detail survey discussion of related papers work carried out so far in this area. In section 3 illustrate the overview of system architecture and description of its phase. In the section 4 implementation of system is outlined where training and testing algorithms steps are discussed. The results and discussion of proposed system is discussed in section 5. Finally the section 6 conclusion of the work is discussed.

2 Literature Outline

In the literature survey section, the history of the earlier work done in this area and there issues are discussed. It contains a record of all the research going in this area. In this section detailed study of the earlier work done on the sign language recognition is discussed. M. R. Abid et al [1] describes in this paper, the state-of-the art Dynamic sign language recognition (DSLR) system for smart home interactive applications. The novel DSLR system comprises two main subsystems: an image processing (IP) module and a Stochastic linear formal grammar (SLFG) module. IP Module used the Bag-of-features (BOFs) and a local part model approach for bare hand dynamic gesture recognition from a video. The SLFG module analyzes the sentences of the sign language (i.e., Sequences of gestures) and determines whether or not they are syntactically valid. The DSLR system is not only able to rule out ungrammatical sentences, but it can also make predictions about missing gestures, which, in turn, increases the accuracy of our recognition task. And this module makes the aggregate performance of the DSLR system as accurate as 98.65%. Houssem Lahiani et al [2] proposed a system based on SVM for recognizing various hand gesture. The system consist of four steps: hand segmentation, smoothing, feature extraction & classification. With this system all steps can be done by the smartphone. In this paper, for image acquisition, frontal camera of the smartphone is used. After that frames are getting from the video, the color sampling is done which is followed by making binary representation of the hand, and then contours representing the hand were described with convex polygons to get information about fingertips and finally the input gesture was recognized using proper classifier.

Rishabh Agrawal et al [3] represent the system to recognize hand gesture for human computer interaction, using computer vision and image processing techniques. The proposed system can successfully replace such devices needed for interacting

with a personal computer and it uses the commercial depth +rgb camera called Senz3D, which is cheap and easy to buy as compared to other depth cameras. The proposed method works by analyzing 3D data in real time and uses a set of classification rules to classify the number of convexity defects into gesture classes. This results in real time performance and negates the requirement of any training data. Jian Wu et al [4] proposed, fusing information from an inertial sensor and SEMG sensors. An information gain-based feature selection scheme is used to select the best subset of features from a broad range of well-established features. Four popular classification algorithms are evaluated for 80 commonly used ASL signs on four subjects. The experimental results show 96.16% and 85.24% average accuracies for intra-subject and intra-subject cross session evaluation, respectively, with the selected feature subset and a support vector machine classifier. The significance of adding sEMG for ASl recognition is explored and the best channel of sEMG is highlighted. Md. Mohiminul Islam et al.[5] come with this system to present a real time HGR System based on ASL, recognition with greater accuracy. This system acquires gesture images of ASl with black background from mobile video camera for feature extraction. For feature extraction " K Convex Hull" algorithm is used which can detect fingertip with high accuracy. In this system, Artificial neural network (ANN) is used with feed forward, back propagation algorithm for training a network using 30 feature vectors to recognize 37 signs of American alphabets and numbers properly which is helpful for HCI system. The total gesture recognition rate of this system is 94.32% in real time environment. Deniz Ekiz et al [6] present a smartwatch application that recognizes important sign sentences. This method represents a smart watch app that collects 3d accelerometer and 3d gyroscope data from the watch and recorded 8 questions and 13 sentences from 5 people who are fluent in sign language. This system use dynamic time warping to compute the distances between the gestures and templates in all data dimensions. The resulting distances serve as input for discriminating the gestures. We evaluated the discriminative ability of logistic regression. Ohene-Djan et al. [7] locate another arrangement, he utilize Mak-Messenger that has manual decision strategy for signs from a catch board. This approach of physically picking pictures wasn't triple-crown on account of the deficiency of ease of use.

3 System Overview

For the developed approach for translation of sign language to text description where a single significant transformation is carried out for a Kannada text description performed. To represent the processing efficiency, a set of sign action consider in training for formulating a text description. This sign action frames are then processed to evaluate the performance for sign language detection. Word processing is carried out as a recursive process of a sign action symbol representation, where each frame data are processed for a HOG features. The frame data are extracted based on the frame reading rate and multiple frames are processed in successive format to extract the region of interest. A system outline to process the real time sign action and to give an optimal frame processing for sign recognition a word level process is performed.

To perform the word processing, the basic approach of the developed system is shown in figure 1.

The scope of the system is to provide a platform for dumb individuals to share their views among every one. Analyze human motion from images and video and developing application like Facebook where dumb people communicate with each other. Video is collection of video frames where frames are collected from video later and after that this frames are used for processing. There is a need to analyze the frame and based on that frame action is identified. Preprocessing purpose the blur from images is removing to improve result. Feature extraction can be done using HOG algorithm and we train the support vector machine by using feature collected from algorithm. In real time web camera to get image and that image to extract HOG feature later that feature used for test the support vector machine. Based on the training support vector machine predict the class label as output. In above chapter the architecture of the system is studied in detail. The overall systems perform preprocessing, Feature extraction, Classification, Detecting the hand gesture and finally generating text description.

The proposed system consists of two major phases training and testing. Where the training process is carried out for a developed database as outlined in the next section [8].

3.1 Training Phase

In training module the images extracted from the captured video and are trained by using SVM after that stored in the database by assigning class label. Figure 1 shows the overview of the system in the training section. All the trained images are used to extract features and which are further used for testing. Firstly, through live video the different frames are captured since a video is nothing but a set of images. Then the training is performed on that captured frames. After that, every Image is processed by filtering technique (noise removal, edge detection or shape detection) and applying Histogram oriented gradient (HOG) algorithm is used for feature Extraction. HOG algorithm defines the objects (hand) and motion shapes in the images by describing the intensity gradient and edge detection. After that a gray scale image is generated. This gray image used as input. The output is a list of points on the image each associated to a vector of low-level descriptors. These points are said key points and their descriptors are invariant by rescaling, in-plane rotating, and noise addition and in some cases by changes of illuminant. The gesture captured from the images are used to generate exact meaning and are ranked in English language. Thus whatever done in training is based on hand gesture and motion used to create exact meaning. Thus in training section, meaning of each gesture are insert into database [9].

3.2 Testing Phase

This module test live video and gets the result in terms of segmentation of frames. In this phase, a video is processed and divided into frames and these frames are further processed by applying the purifying algorithm to remove noise from images.

Median blur technique is used to filter image. The lower part of figure shows the testing phase. After elimination of noise, the features of images are extracted and these features are linking with training videos to recognize text. The system undergo following step to yield the desired result. Concept that focuses on the components or elements of a structure or system and unifies them into a coherent and functional whole according to a particular approach in achieving the objective(s) under the given constraints or limitations. A block diagram is a specialized, structure provides a high-level overview of major system components, key process participants, and important working relationships. In figure1 block diagram functions used for implementation are represented. Video processing, image processing, feature extraction and action detection using SVM classifier are functions used for application implementation.

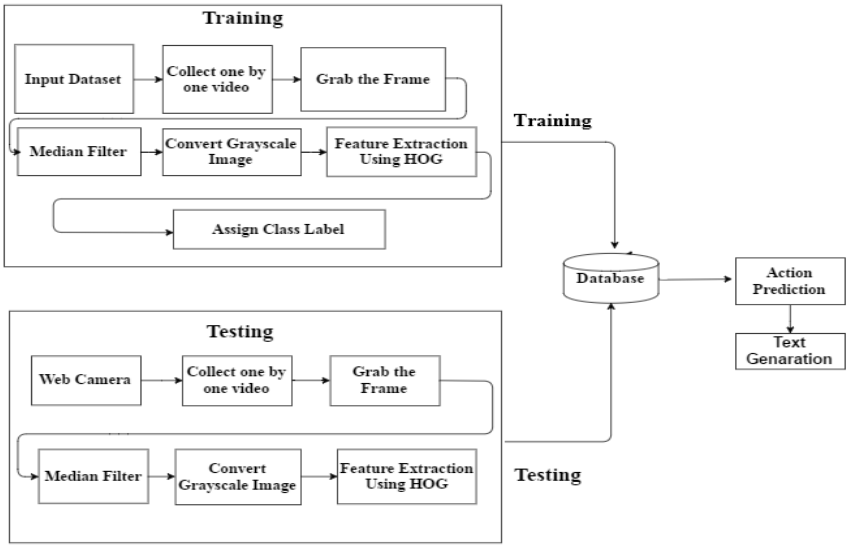

Fig. 1. Overview of the system.

4 System Implementation

For the developed approach of sign language detection, where a single significant transformation is carried out, a Kannada text detection is then performed. To represent the processing efficiency, a set of cue symbols is used for formulating a word. This word symbols are then processed to evaluate the performance for sign language detection. Text processing is carried out as a recursive process of a single cue symbol representation, where each frame data are processed for a shape feature. The frame data are extracted based on the frame reading rate and multiple frames are processed in successive format to extract the region of interest. A system outline to process the sign video data and to give an optimal frame processing for sign recognition a text level process is performed. To perform the text processing, the basic approach of the developed system is shown in figure 1.

The text symbols extract from a given video sample, where the video data is processed frame wise manner and the recurrent frame information are eliminated as redundant bits. To perform the frame coding the video frame under process, is processed using a joint adjacent matching and a singleton region matching algorithm is used for frame processing. In the join adjacent region processing, in this frame processing, each characters are extracted as a set of image data and processed in a recurrent manner to extract the feature described. On the process of the text recognition process the video frames are extracted based on the frame rate of the video sample. The video data are processed as an energy correlation, where the video sample is processed in time frame slices. Energy interpolation for video sign recognition gives the advantage of information retrieval based on energy mapping, where energy details are used as an informative parameter for sign language detection. The extracted HOG features are passed to classifier to make a final decision. The classifier logic performs a classification based on searching the best match feature using SVM approach. The recognized character is processed for mapping into the class level [11]. These processing systems have a variant output format, such as text output. In this section all the methods and techniques that are used for the system implementation are discussed below. For training and testing follows below steps,

1. **Video Acquisition:** It acquires the video from the user and performs translation of videos into its frames (Multiple images). Then each frame will undergoes preprocessing before extracting features and the preprocessing step will be discussed in the next section. Meanwhile video is continue to capture from web camera and is divided into multiple frames. Every individual image called as a frame. Video framing is the process of extracting frames in giving video using video attributes like frame rate. For example the video duration is the 2 min and 29 seconds and frame rate will be 1000 FPS (Frame per second) then extracted frames will be 14900 frames.
2. **Preprocessing:** This section consists preprocessing on video like noise and blur elimination. Video contain huge amount of frames in that frame contain visual distortion like a video shoot in low light area, voice distortion, light conditions etc. The preprocessing is a common stage in every image processing area. The principal motivation behind preprocessing is to diminish commotion on the edge and improve the picture highlight for further handling. The median filter uses the nonlinear filtering techniques to remove blur from input frame. By using filtering techniques to improve image quality and output result. The main idea behind median filtering is replacing every entry with neighboring entry[12].
3. **Feature Extraction:** HOG is widely used to extract the feature from the input image. HOG is an image processing algorithm. This will detect the object in the image by using feature vector. Whenever extract the feature from the image it produced output in the form of feature vectors. Feature contains the local shape information this can use for many tasks such as classification, detection of objects, track the object.
4. **Classification:** SVM characterization is basically a double i.e. binary (two-class) order system, which must be altered to deal with the multiclass undertakings in cer-

tifiable circumstances. SVM order utilizes elements of picture to the group[13]. The characterization utilizes prepared video and group testing video with specific depiction and gives yield.

5. **Text Generation:** At last when the classification process is completed the equivalent grammatical text description will be generated with the help of the class labels which are assigned during the training phase.

4.1 Algorithm for Implementation

The proposed system uses two phases training and testing to perform sign to text description. In the both phase uses logical operation to perform the activity such as training and testing algorithm.

Training algorithm. Training algorithm is used to train the system with the help of the available data set. The training data set involve different algorithms to train the system [14]. The steps of training algorithm is as follows,

Algorithm: Training

Input: Consider the different set of sign action video from the signers of different domain.

Output: Set of features extracted from each real time sign action signers and is stored in database.

Steps:

Begin

Step 1. Consider the real time sign action from the signer with static background with high camera resolutions.

Step 2. Perform frame extraction where the real time sign action data is processed frame wise manner and the recurrent frame information are eliminated as redundant bits.

Step 3. Perform preprocessing to remove noise and blurriness effect using median filter algorithm from the real time sign action.

Step 4. Convert the gray scale image form extracted frames because it require less processing time than that of the colored image.

Step 5. Detect the edge from this sample with the canny edge detection technique. Mathematically Canny edge technique is expressed as,

$$f(x) = \frac{I_r - I_l}{2}\left(\operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma}\right) + 1\right) + I_l.$$

Where,

$$I_l = \lim_{x \to -\infty} f(x)$$

$$I_r = \lim_{x \to \infty} f(x).$$

Step 6. Compute the features with the HOG technique. These features are stored into the database in .csv file along with the associated class labels and also mapped features are buffered into an array to formulate the database. Apply HOG feature extraction algorithm to extract the feature from a given image. This algorithm implementation follows 4 steps as,

1. **Gradient Computation:** The initial step of estimation in many component locators in picture pre-handling is to guarantee standardized shading and gamma values.
2. **Orientation Binning:** The second step of estimating is making the cell histograms. Every pixel in the cell makes a choice for an introduction construct histogram channel depends with respect to the qualities found in the inclination calculation.
3. **Descriptor Block:** To represent changes in light and complexity, the angle qualities must be privately standardized, which requires gathering the phones together into bigger, spatially associated squares.
4. **Block Normalization:** Dalal and Triggs investigated four unique techniques for piece standardization. They provide equation for a frame at that point the standardization element can be one of the accompanying: L2-standard which is,

$$f = {}^{v}\Big/_{\sqrt{\|v\|_2^2 \quad + e^2}} \tag{4}$$

Where,
f: A frame
v: video sample
e: detected object/element from the video

Step 7. Similarly repeat the step 1 to 6 for all sign video's from the data base and extract all features and will be stored in data base one by one.

End

Testing algorithm. e In this testing of sign video sample from the data base is considered and before testing it is compulsory to train the data base samples [15-16]. The testing algorithm follows the same procedure as that of the training algorithm with the exception that there is no need to capture the video explicitly the system can capture the video with the web camera installed. The steps of training algorithm is discussed below,

Algorithm: Testing
Input: Sign action video from the data base.
Output: The text description in Kannada language.
Steps:
Begin
Step 1. Consider the sign action from the signer with static background with high camera resolutions.

Step 2. Perform frame extraction where the real time sign action data is processed frame wise manner and the recurrent frame information are eliminated as redundant bits.

Step 3. Perform preprocessing to remove noise and blurriness effect using median filter algorithm from the real time sign action.

Step 4. Convert the gray scale image form extracted frames because it requires less processing time than that of the colored image.

Step 5. Detect the edge from this sample with the canny edge detection technique. Mathematically Canny edge technique is expressed as,

$$f(x) = \frac{I_r - I_l}{2}\left(\operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma}\right) + 1\right) + I_l.$$

Where,

$$I_l = \lim_{x \to -\infty} f(x)$$

$$I_r = \lim_{x \to \infty} f(x).$$

Step 6. Compute the features with the HOG technique. These features are stored into the database in .csv file along with the associated class labels has to assign and also mapped features are buffered into an array to formulate the database.

Apply HOG feature extraction algorithm to extract the feature from a given image. The algorithm implementation follows 4 steps as,

1. **Gradient Computation:** The initial step of estimation in many component locators in picture pre-handling is to guarantee standardized shading and gamma values.
2. **Orientation Binning:** The second step of estimating is making the cell histograms. Every pixel in the cell makes a choice for an introduction construct histogram channel depends with respect to the qualities found in the inclination calculation.
3. **Descriptor Block:** To represent changes in light and complexity, the angle qualities must be privately standardized, which requires gathering the phones together into bigger, spatially associated squares.
4. **Block Normalization:** Dalal and Triggs investigated four unique techniques for piece standardization. They provide equation for a frame at that point the standardization element can be one of the accompanying: L2-standard which is

$$f = {}^{v}\Big/\sqrt{\|v\|_2^2} \quad + e^2$$

Where,
f: A frame
v: video sample
e: detected object/element from the video

Step 7. Apply SVM classifier to match the features to make a final decision. The classifiers also predict the meaning of action. The recognized character is processed for mapping into an equivalent text description.

Step 8. The equivalent grammatically correct text description is generated.

5 Results and Discussion

For the realization of the proposed system, the need of Kannada sign dataset (KSD) is developed using all the constraints to sign and character representation. The Kannada (ಕನ್ನಡ) is the official local language of the southern Indian state of Karnataka. It is a Dravidian language spoken by about 44 million people in the Indian states of Karnataka, Andhra Pradesh, Tamil Nadu and Maharashtra. The Kannada script is widely used for writing Sanskrit texts in Karnataka. Several minor languages, such as Tulu, Konkani, Kodava, Sanketi and Beary, also use alphabets based on the Kannada script. For the sign action recognition there would be a huge amount of database is generated. As the actions has no limit for development different type of actions with different views and with different perspective are recorded. In this system by studying the signs of the deaf community we group these actions into 50 different domains and each domain has 5 examples [17-18]. Out of these action videos 5 videos samples are trained and tested by the system because of the available machine configuration. That means the system is able to be run only these 5 samples trained video s and the detail of these video is available in the table1.

In the Meeting_Video1 is actually a sign action performed by the deaf and captured by the system through the web camera. The first example in the table meeting _Video1 has a size 7.25 MB the meaning of that sign action is **Nice to meet you**. The system will take total of 10 seconds time to process this sign action to translate and it generate text description in kannada meaning like "**ಸಿಕ್ಕಿದಕೆ ಖುಷಿಯಾಯಿತು ನೀನು**" as the expected result.

Similarly the Time_Video1 sign action video having meaning is **What is todays date** for this system will take near about 12 second to process and produce the result in kannada as "**ಇವತ್ತಿನ ದಿನಾಂಕ ಏನು?**" is as per expectation hence the remark for that is successful. The processing time is 2 seconds more than the previous example because in this example the action part of the user is more [19].

In the table 7.2 the count of expected words actual generated words and expected words are same hence accuracy rate for all the 5 examples are is 100% because we have selected testing samples from Database which are already trained. In this section, time require to process video is shown. The system in turn uses time in milliseconds which is required for the processing of each frame is shown in figure. Three different graphs for each video are drawn. First graph is for preprocessing time, second graph is for edge detection time and third graph is for the feature extraction. As the number of frames increases the time required for processing is also increases [20].

Table 1. Over all summary of the implemented system.

Sr. No	Video Sample	Time Duration (Sec)	Processing Time (Sec)	Expected Output	Actual Output	Recognition Rate in %
1.	Meeting _Video1	20	10	ಸಿಕ್ಕಿದಕೆ ಖುಷಿಯಾಯಿತು ನೀನು. Nice to meet you.	ಸಿಕ್ಕಿದಕೆ ಖುಷಿಯಾಯಿತು ನೀನು. Nice to meet you.	EWC= 4 AWC= 4 Accuracy=100%
2.	Time _Video1	25	12	ಇವತ್ತಿನ ದಿನಾಂಕ ಏನು? What is todays date?	ಇವತ್ತಿನ ದಿನಾಂಕ ಏನು? What is todays date?	EWC= 4 AWC= 4 Accuracy= 100%
3.	Place_Video1	30	13	ಇದು ಅಪಾಯಕಾರಿ ಸ್ಥಳ. This is dangerous place.	ಇದು ಅಪಾಯಕಾರಿ ಸ್ಥಳ. This is dangerous Place.	EWC= 4 AWC=4 Accuracy= 100%
4.	Gossips _Video1	50	20	ಹಲೋ ಏನು ನಿಮ್ಮ ಹೆಸರು? Hello What is your Name?	ಹಲೋ ಏನು ನಿಮ್ಮ ಹೆಸರು? Hello what is your Name?	EWC= 5 AWC= 5 Accuracy= 100%
5.	Theater _Video1	26	10	ನಾನು ಚಲನಚಿತ್ರ ರಂಗ ಭೂಮಿಗೆ ಹೋಗುತ್ತೇನೆ. I go to movie theater.	ನಾನು ಚಲನಚಿತ್ರ ರಂಗ ಭೂಮಿಗೆ ಹೋಗುತ್ತೇನೆ. I go to move theater.	EWC= 4 AWC= 4 Accuracy= 100%

5.1 Result Analysis for Video Samples

The time taken by each Median filter, Canny and HOG algorithm are discussed and processing time required for video sample 1(Meeting_Video1). The following table and graph shows the time required for preprocessing, Canny Edge detection and HOG algorithm.

Preprocessing (Median filter). The table 2 gives the details about the frames extracted and processing time in milliseconds of a video sample and its corresponding graph is shown in figure 2 where the preprocessing time Vs frame extracted is plotted.

Table 2. Preprocessing time

Sr. No.	Video sample	Frames Extracted	Processing Time (millisecond)
1.	Meeting_Video1	Frame 1	400
		Frame 2	450
		Frame 3	400
		Frame 4	450
		Frame 5	400
		Frame 6	450
		Frame 7	450

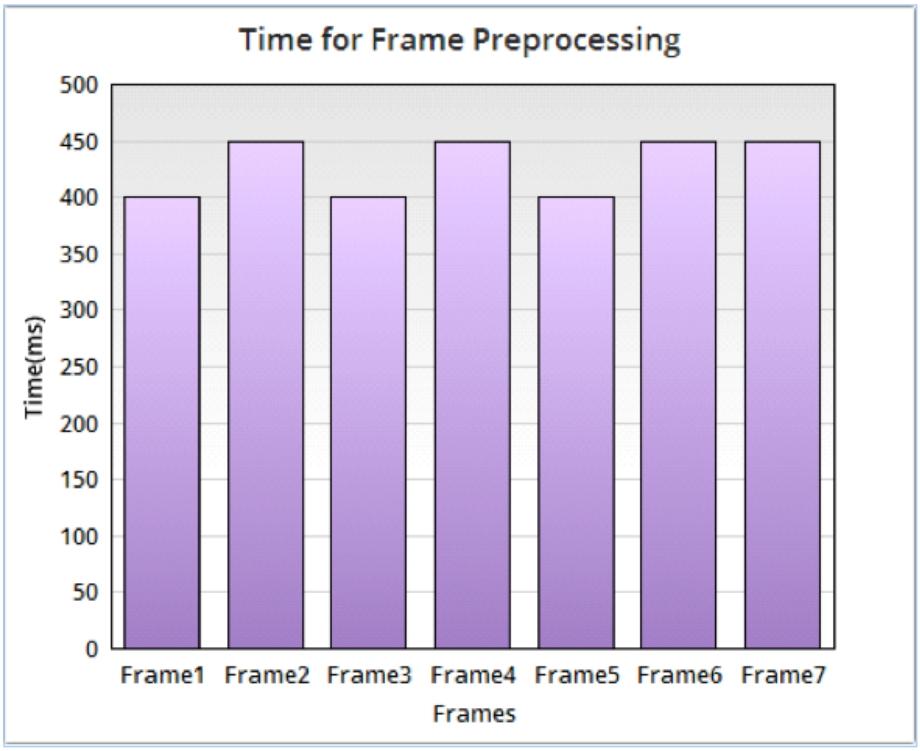

Fig. 2. Graph for preprocessing of each frame.

Canny Edge Detection (Canny Edge Detector). The table 3 gives the details of edge detection processing time and its feature extraction of sample video and its corresponding graph is shown in figure 3.

Table 3. Edge detection

Sr. No.	Video	Frames Extracted	Processing Time (millisecond)
1.	Meeting_Video1	Frame 1	450
		Frame 2	400
		Frame 3	400
		Frame 4	400
		Frame 5	500
		Frame 6	450
		Frame 7	500

Time in milliseconds require for canny edge detection technique for each frame of videos is shown. As size of frames increase, the time requires processing that frames also get increases.

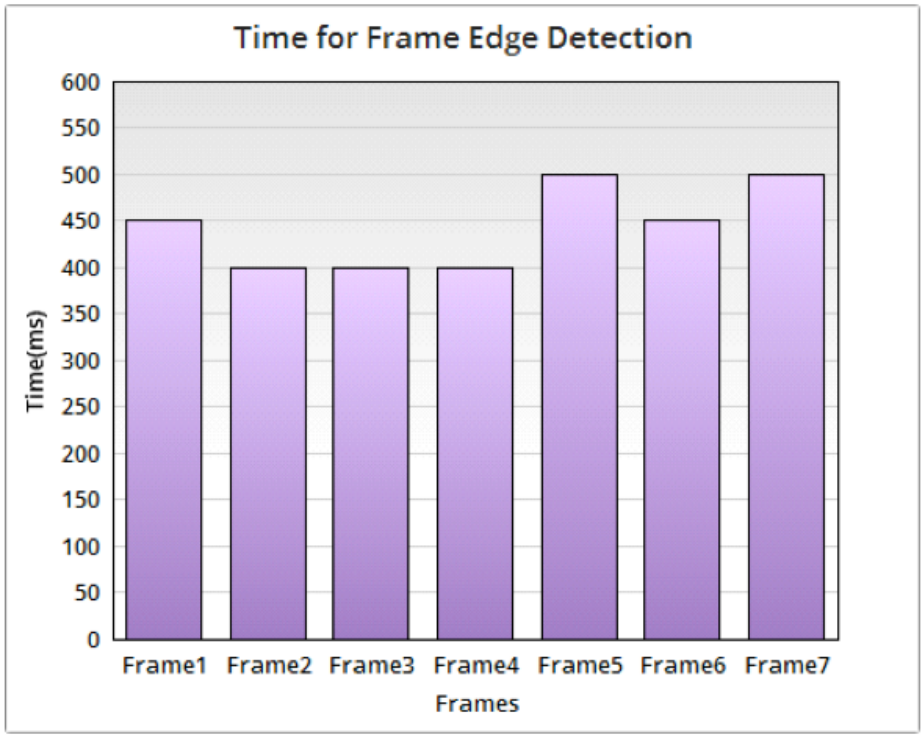

Fig. 3. Graph for edge detection time.

Feature Extraction (HOG). Time required for feature extraction from each frame is shown in following table 4 and its graph is also shown in figure 4. The HOG algorithm require more time as compared to edge detection and preprocessing.

Table 4. Feature extraction time.

Sr. No.	Video	Frames Extracted	Processing Time (millisecond)
1	Meeting_Video1	Frame 1	3000
		Frame 2	4500
		Frame 3	4500
		Frame 4	3500
		Frame 5	5000
		Frame 6	5000
		Frame 7	4500

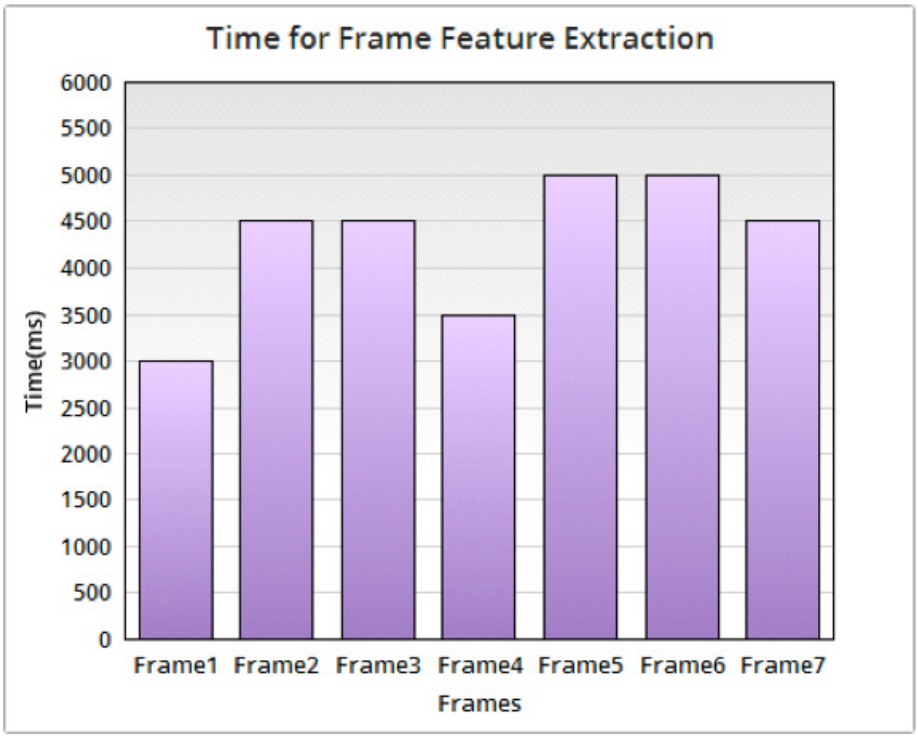

Fig. 4. Graph for feature extraction.

5.2 Analysis of Overall Videos

In this section, overall analyses of video samples are discussed here. In the table 5 video samples of different domain are taken as an input and it shows the number of frames extracted for each video. Dissimilar frames are considered and the similar frames are discarded. In figure 5 shown a graph for comparison of frames.

Table 5. Number of frames comparison.

Sr. No.	Domain	Number of Frames Extracted
1.	Meeting	7
2.	Theater	5
3.	Gossips	6
4.	Place	8
5.	Gratitude	7

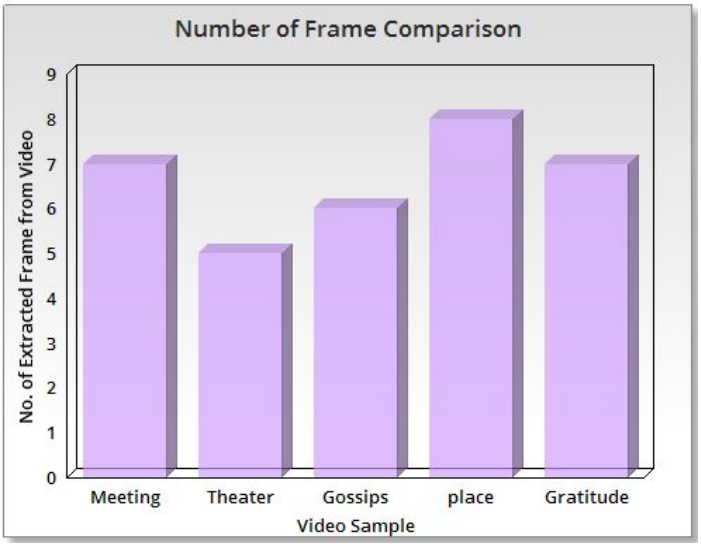

Fig. 5. Graph for comparison of frames.

From this analysis the average number of frames and the overall accuracy of extraction of the frame can be calculated in percentage as,

$$\text{Average number of frames} = \frac{\sum of\ all\ the\ extracted\ frames\ from\ each\ domain}{\sum number\ of\ domain\ taken\ for\ testing}$$

$$\text{Therefore, Average number of frames} = \frac{7+5+6+8+7}{5} = 6.5 \sim 7 \text{ frames}$$

So finally 7 frames are to be extracted from the video to get the expected output in text description from that video.

5.3 Analysis of Processing Time

The following table 6 shows total processing time taken by each video. It is the average time taken by each algorithm i.e. Median Filter, Canny and HOG. Processing time is directly related to the video size and the number of objects present in each frame. More the video size, processing time will be more. In figure 6 shows a graph for time comparison.

Table 6. Time comparison.

Sr. No.	Domain	Time in Milliseconds
1.	Meeting	10000
2.	Theater	21000
3.	Gossips	47000
4.	Place	10000
5.	Gratitude	12000

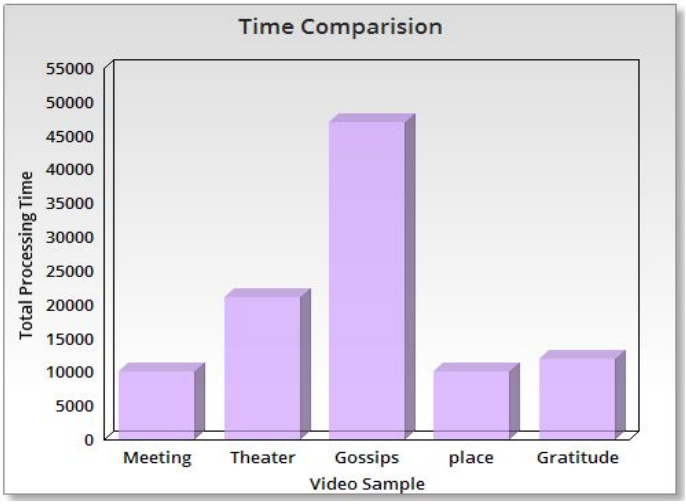

Fig. 6. Graph for time comparison.

From this analysis the average number of time and the overall accuracy come up with the processing time can be calculated in percentage as,

Average processing time =

$$\frac{\sum Indivisual\ processing\ time\ taken\ by\ each\ video\ from\ different\ domain}{\sum number\ of\ domain\ taken\ for\ testing}$$

Therefore,

Average number of frames = $\frac{10000+21000+47000+10000+12000}{5}$ = 20000 milliseconds

That means an average there is 20000 milliseconds (20 seconds) are required to process each video. If any video are taken and trained into the system then definitely it will take minimum 20 seconds to gives an output. This processing time is an average so it may be varies.

5.4 Analysis of Recognition Rate for Trained Video

For the understanding of the working capability of the system the analysis system result is necessary. In this section the recognition rate analysis of trained video sample is done. The recognition rate of the system can be calculated based on the number of words in a generated text description to the expected words which must have been generate by the system as per its design.

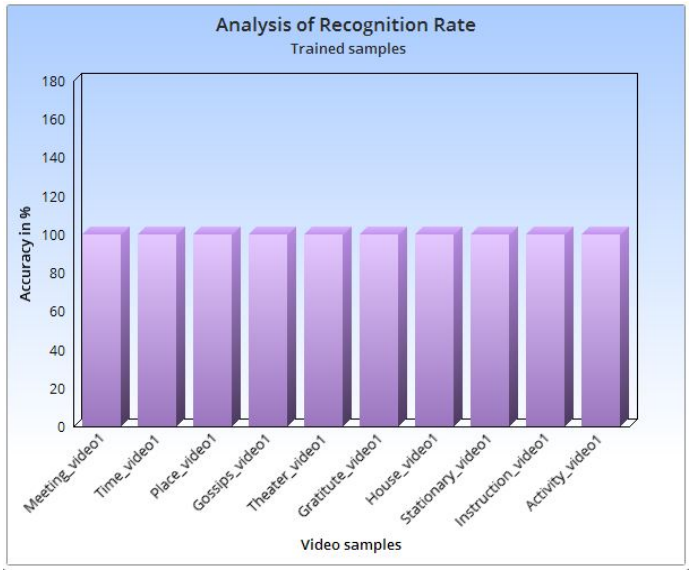

Fig. 7. Analysis of recognition rate for trained samples.

The trained sample are such videos having features are extracted in training stage are stored into the database. Hence when these examples are used for the training it generates the exact text description as expected by the signer. Therefore the accuracy of all trained samples is 100. The figure 7 shows all the trained example explained in the table 1 with its respective accuracy rate. The accuracy of the recognition rate is calculated as,

$$\text{Accuracy} = \frac{number\ of\ actual\ words\ in\ the\ text\ description}{number\ of\ Expected\ \ words\ in\ the\ text\ description} * 100$$

Here the Meeting_video1 sample has expected meaning is "Nice to meet you." And the actual generated text is "**Nice to meet you**" The expected and actual output is same here hence the word count is also same therefore by putting it in above formula it gives its recognition accuracy rate is,

$$\text{Accuracy rate for meeting\_video1} = \frac{4}{4} * 100 = 100\%.$$

Likewise accuracy rate for all the trained example in table 7.2 is calculated and the graph is generated. The average accuracy for trained video sample is calculated as,

$$\text{Average accuracy} = \frac{\sum sum\ of\ accuracy\ of\ individual\ video\ sample}{\sum number\ of\ the\ video\ sample}$$

$$\text{Average accuracy of trained video} = \frac{100+100+100+100+100+100+100+100+100+100}{10} = 100$$

Hence the average accuracy of trained video is 100 % because the video samples are already trained during training.

6 Conclusion

The developed system is processed for real time sign action to yield grammatically correct words. Where different samples are considered for the feature extraction, the class label assigned to extracted features and finally generate text. The purposed approach results in higher retrieval accuracy as compare to conventional processing system. This system result in lower descriptive feature with a minimum processing frames which hence achieved objective of the higher accuracy and lower processing overhead. The system can be further continue to minimize the processing time and high recognition rate a different technique can be applied for future work. In future we are looking at development of system which is signer independent and will generate summery.

7 References

[1] Muhammad Rizwan Abid, Emil M. Petriu, Fellow, IEEE, and Ehsan Amjadian, "Dynamic Sign Language Recognition for Smart Home Interactive Application using Stochastic Linear Formal Grammer ", IEEE Transactions On Instrumentation And Measurement, Vol. 64, No. 3, March 2015.

[2] Houssem Lahiani, Mohamed Elleuch and Monji Kherallah, "Real Time Hand Gesture Recognition System for Android Devices", 2015, 15th International Conference on Intelligent Systems DeSign and Applications (ISDA). https://doi.org/10.1109/ISDA.2015.7489184

[3] Rishabh Agrawal and Nikita Gupta, "Real Time Hand Gesture Recognition for Human Computer Interaction", 2016 IEEE 6th International Conference on Advanced Computing.

[4] Jian Wu, Student Member, IEEE, Lu Sun, and Roozbeh Jafari, Senior Member, IEEE, "A Wearable System for Recognizing American Sign Language in Real Time Using IMU and Surface EMG Sensors", IEEE Journal of Biomedical and Health Informatics, Vol. 20, No. 5, September 2016.

[5] Md. Mohiminul Islam, Sarah Siddiqua, and Jawata Afnan, "Real Time Hand Gesture Recognition using Different Algorithm Based on American Sign Language", ISBN.978-1-5090-6004-7/17/ ©2017 IEEE.

[6] Deniz Ekiz, Gamze Ege Kaya, Serkan Buğur, Sıla Güler, Buse Buz, Bilgin Kosucu and Bert Arnrich, "Sign Sentence Recognition with Smart Watches", ISBN.978-1-5090-6004-6/17/ ©2017 IEEE.

[7] Ohene-Djan, J., Zimmer, R., Bassett-Cross, J., Mould, A. and Cosh, B., Mak- Messenger and Finger-Chat, Communications Technologies to Assist in Teaching of Signed Lan-

guages to the Deaf and Hhearing. In: IEEE International Conference on Advanced Learning Technologies, 2004. pp. 744 – 746.

[8] M. R. Abid, L. B. S. Melo, and E. M. Petriu, "Dynamic Sign Language and Voice Recognition for Smart Home Interactive Application", in Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA), May 2013, pp. 139–144.

[9] X. Sun, M. Chen, and A. Hauptmann, "Action Recognition via Local Descriptors and Holistic Features," in IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL,USA, 2009, pp. 58–65.

[10] T. Wenjun, W. Chengdong, Z. Shuying, and J. Li, "Dynamic Hand Gesture Recognition using Motion Trajectories and Key Frames," in Proc. second International Conference. Advance Computing Control (ICACC), March 2010, pp. 163–167.

[11] Ramesh M. Kagalkar and Dr. S.V. Gumaste, "Curvilinear Tracing Approach for Extracting Kannada Word Sign Symbol from Sign Video", International Journal of Image, Graphics and Signal Processing, Volume 9, pp.18-27, Published Online September 2017 in MECS (http://www.mecs-press.org/) https://doi.org/10.5815/ijigsp.2017.09.03

[12] Ramesh M. Kagalkar and Dr. S.V. Gumaste, "ANFIS Based Methodology for Sign Language Recognition and Translating to Number in Kannada Language", International Journal of Recent Contributions from Engineering, Science & IT (DBLP indexed Journal), Volume 5, Issue No. 1, pp. 54-66, 2017.

[13] Ramesh M. Kagalkar and Dr. S.V.Gumaste, "Gradient Based Key Frame Extraction for Continuous Indian Sign Language Gesture Recognition and Sentence Formation in Kannada Language: A Comparative Study of Classifiers", International Journal of Computer Sciences and Engineering, Volume 4, Issue 9, 2016.

[14] Ramesh M. Kagalkar and Dr. S.V.Gumaste, "Review Paper: Detail Study for Sign Language Recognition Techniques", CiiT International Journal of Digital Image Processing, Volume 8, No 3, 2016.

[15] Rashmi Hiremath and Ramesh M. Kagalkar, "Methodology for Sign Language Video Interpretation in Hindi Text Language", International Journal of Innovative Research in Computer and Communication Engineering, Volume. 4, Issue 5, May 2016.

[16] Rashmi Hiremath and Ramesh M. Kagalkar, "Methodology for Sign Language Video Analysis into Text in Hindi Language", CiiT International Journal of Fuzzy Systems, Volume 8, No 5, 2016.

[17] Ramesh M Kagalkar and Nagaraj H.N., "New Methodology for Translation of Static Sign Symbol to Words in Kannada Language", International Journal of Computer Applications 121(20):25-30, July 2015.

[18] Ramesh M. Kagalkar, and Nagaraj H.N., and Dr. S.V.Gumaste, "International Journal of Advanced Research in Computer and Communication Engineering", Vol. 4, Issue 7, July 2015.

[19] "A Novel Technical Approach for Implementing Static Hand Gesture Recognition", International Journal of Advanced Research in Computer and Communication Engineering, Volume. 4, Issue 7, July 2015.

[20] Amitkumar and Ramesh M. Kagalkar, "Sign Language Recognition for Deaf Sign User", International Journal for Research in Applied Science and Engineering Technology (IJRASET), Volume 2, Issue 12, December 2014.

8 Authors

Ramesh M. Kagalkar is Research Scholar, VTU-RRC, Visvesvaraya Technological University (VTU), Belgaum, Karnataka, India.

Shyamrao V Gumaste is Professor, Dept. Computer Engineering, MET League of College, Nashik, Maharashtra, India.

Article submitted 04 December 2017. Resubmitted 04 January 2018. Final acceptance 05 March 201^8. Final version published as submitted by the authors.