Received: February 26, 2021. Revised: March 31, 2021. 123

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Static and Dynamic Hand Gesture Recognition Using GIST and Linear Discriminant Analysis

Phyu Myo The[1]* **May The` Yu[2]**

[1]*Image Processing Lab, University of Computer Studies, Mandalay, Myanmar*
[2]*Faculty of Information Science, University of Computer Studies, Mandalay, Myanmar*
* Corresponding author's Email: phyumyothwe@ucsm.edu.mm

Abstract: Hand gesture detection and recognition is a way to communicate between deaf-mute person and rest of society. In this study, the static sign and the dynamic sign recognition system based on hand gesture is presented. The dynamic sign gesture is more challenging than static sign gesture which contains the movement. The vision based sign language recognition using hand gesture consists of the following main steps: acquisition the sign data, detection the hand region, extraction the features and recognition. The data acquisition stage, the data acquired by using web camera without using any data gloves and special markers. In hand region detection process, this study proposed hybrid hand region detection process by combining CbCr channel from YCbCr colour space and motion detection by using mean filter background subtraction to definitely segment the hand region from the background. In hand feature extraction, the scale, shape, texture and orientation features extracted by using GIST (Generalized Search Tree), HOG (histogram of gradient) and HOG-LBP (Local Binary Pattern) methods. Finally, Quadratic Support Vector Machine (QSVM), Cubic Support Vector Machine (CSVM) and Linear Discriminant recognized the static and dynamic hand gesture. The recognition accuracy achieved 86.7% and 99.22 % acceptable accuracy on self-construct dynamic Myanmar sign word dataset and MUDB dataset.

Keywords: Hand region detection, CbCr channel, Motion detection, Generalized search tree, Histogram of gradient and hand gesture recognition.

1. Introduction

The sign language-based hand gestures are used as important characteristics in several applications. The detection and recognition of gestures are applied in sign language recognition; remote control access; smart home system control; robot control and many others. The sign language recognition processes are mainly of two types namely vision based and sensor based. Sensor based captures the gestures by using instrumented gloves equipped and attempt to recognize using suitable machine learning methods and image-based sign language recognition does not require any extra devices [1]. The human-machine interaction interface (HMI) has become a popular area of research that employs the concept of gesture detection and recognition. This has become major trend due to the role of hand gestures in electronic device [2]. The two processes of hand gesture are static based hand gesture and dynamic based hand gesture. The static gestures contain single image and not contain movement. The dynamic gestures present the gesture with the sequence of images or video and contain the movement when performing the gesture [3].

The proposed system for dynamic Myanmar Sign words recognition system that could of significant help to persona of hearing impaired. The different countries and regions have their own sign language. The sign language (SL) is not universal sign language. The sign language recognition helps the social and communication gap between deaf-mute person and rest of people. In sign language, we have generally three basic components: fingerspelling, word level sign and non-manual sign. The first

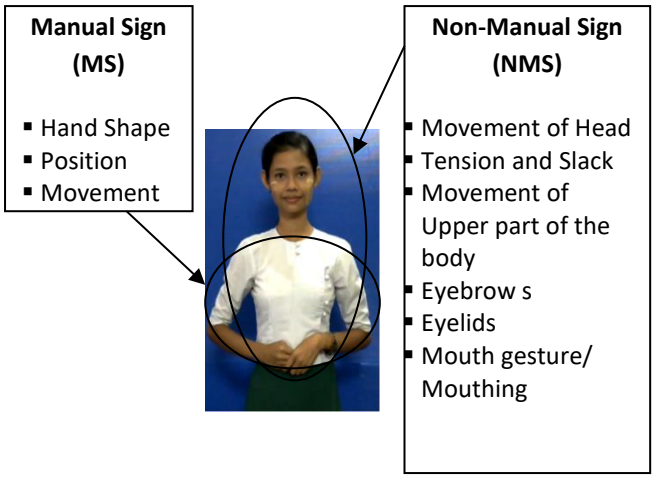

Figure. 1 Structure of Myanmar sign language

component is fingerspelling which is used to communicate with disable people of hearing and speaking, the second component is word level sign vocabulary which is used to describes the sign word meaning in the sign language; and the last component is non-manual sign communication that describes the meaning of signs such as movement of head, body, facial expression, eyebrows, eyelid and mouth. Myanmar fingerspelling is prevalent as the basic language for deaf signer and hearing-impaired people. The second component is commonly used by combining facial expression for deaf-signer, where the word level dynamic Myanmar signs are used for word recognition. Myanmar has four schools for deaf learners: (1) Mary Chapman School for the Deaf in Yangon; (2) School for the Deaf, Mandalay (3) School for the Deaf, Tamwe, Yangon and (4) Immanuel School for the Deaf in Kalay. The structure of Myanmar sign language is shown in Fig. 1 which depicts the manual sign, non-manual sign using hand and other movements. Different Myanmar sign language is used in different schools of Myanmar based on various regions of the nation. The government project was planned to implement the national sign language with the aid of the Japanese federation of the deaf [4] and work is in study and planning stages. The current literature suggests that the research on Myanmar sign language is very limited and could not be found much.

The aims of this paper through our work are as follows:

- to extract the hand region with face and hand overlap condition when performing the gesture.
- to analysis the suitable feature descriptor and classifier for hand gesture recognition.
- to help the communication gap between deaf mute person and other persons.

The major contribution of this paper is to find the most appropriate features descriptor and classifier for static and dynamic hand gesture recognition system. This section is an introduction of the dynamic Myanmar sign word recognition system and the remaining parts of the paper are described subsequently. The literature of current work in the relevant areas is described in section 2. The methodology of static and dynamic sign based hand gesture recognition and the performance evaluation of the method are described in section 3 and section 4. In last section, the conclusions are reported.

2. Related work

Hand gesture detection and recognition play an important role in the development of Human-Computer Interaction (HCI) system. We have studied a number of research papers for various applications, especially for sign language analysis and biometrics. The following section describes a brief discussion about the literature on visual based hand region detection and gesture recognition system.

Sahoo et al. (2018) proposed gesture recognition system based on hand which employed American sign characters implemented for American sign language (ASL) recognition by using discrete wavelet transform (DWT) for feature extraction and F-ratio (Fisher) method for selection of the best DWT coefficient among all coefficients. The recognition accuracy was achieved 98.64%, 95.42% and 98.08% while tested on Massey University Dataset (MUDB), Jochen_Triesch Dataset (JTD) Dataset and Complex Dataset respectively. The main issue in this paper was misclassification when the rotation noise above 15 degrees [5]. Rahim et al. (2019) implemented recognition of dynamic sign words convolutional neural network (CNN) which was tested on self-constructed dataset. The authors also studied and presented hybrid hand region segmentation on two colour model such as YCbCr and HSV. The recognition accuracy of this work was achieved as 97.28% [6]. Bao and partners studied a hand gesture recognition using CNN and without any help of localization method for tiny hand which is generally used in any deep learning methods. The main challenges of this work include slow response in recognition process that is slower speed the overall computation time become very large. Another issue which was reported in this work was performance degradation of the system when subjected to the complex background [7]. Lee et al. (2018) proposed recognition of the static hand gestures with the use of wristband-based contour features (WBCFs) for complex static hand gestures recognition. The accuracy in the recognition process was noted as 99.31% while tested on 29 Turkish fingerspelling

signs. The main disadvantage of the work was detection of the wristband when the background colour is similar to the black colour because the black shadows were observed while performing the hand gesture recognition [8]. Alejo et al (2019) implemented dynamic gesture recognition employing (HSV and CIELab) model-based hybrid segmentation for detection of hand gestures. The authors suggested and implemented PCA (principal component analysis) method for the feature extraction; the classification of the features was done by KNN (k-nearest neighbor) classifier. The recognition accuracy was found as 94.74% while tested on self-constructed dataset [9]. Lahiani et al (2018) [10] implemented static hand gesture recognition system by using hybrid method as combination of local binary pattern (LBP) and histogram oriented gradients (HOG) as feature extraction methods. The recognition accuracy was achieved 92 % while tested on enhanced NUS hand pose dataset I. This combination as HOG-LBP feature extraction method was used and it was compared with the original LBP feature extraction method and HOG feature extraction method. But, this combination was found more time consuming than other two methods. Moni et al., 2009 implemented the continuous sequence of sign gesture using Hidden Markov Models (HMM) [11]. Tubaiz et al [12] implemented Arabic sign language recognition system by wearing glove and using modified K-nearest neighbour method. The recognition accuracy was achieved as 98.9% tested on own constructed dataset. Zheng et al., 2018 [13] presented a static gesture recognition based on hand by employing a CbCr colour space based Gaussian mixture model and deep convolution neural network (DCNN). In the pre-processing stage of this work, three main steps were used such as colour adjusting, extraction the region of hand gesture area. Finally, the reconstruction of hand gesture area was done and the feature extraction and gesture recognition were performed by using self-constructed deep convolutional neural network on NUS dataset and own created dataset. The recognition accuracy was found as 99% while tested on both two datasets.

In state of art research methods, the YCbCr colour space model is generally used for the detection of human skin colour segmentation and pre-processing of hand region detection. The threshold values of YCbCr colour space are not fixed because of different skin colours; lighting conditions; and input device shadow effects. The different threshold values of YCbCr colour space were considered in the literature [14-16]. Thakur et al. (2011) proposed skin colour model as combination of three colour spaces such as RGB, HSV and YCbCr called RGB_HS_CbCr [17]. In the literature, the other colour space model was used to detect the human skin colour. Dariusz et al (2015) suggested CMYK colour space model for human skin colour detection [18]. Reshan et al. 2017 [19] presented Indian sign language recognition using YCbCr colour channel in segmentation of the hand region in the pre-processing stage. The YCbCr colour channel is efficient for the recognition of image pixels in colour image. Tang and partners (2019) [20] implemented dynamic hand gesture recognition system which was based on key frame extraction and fusion of features. The system is experimented on the four different datasets: Hand Gesture Dataset and Action 3D dataset Cambridge Hand Gesture Dataset and Northwestern Dataset. The recognition accuracy was found 99.21% 98.98 % 98.23% and 96.89%, for the four datasets respectively. The fusion of features was done by combining appearance feature and motion features. The appearance features were extracted by SURF and LBP and the motion features extracted by LBP-TOP and SIFT 3D. In the recognition process, the SVM is popularly used to recognize the hand gesture. The author is selected the number of selected key frames is five and the databases contains only the hand region image with simplest background. The literature [5, 6, 20, 21] used Multi-class SVM for classification of the hand gesture. The Multi-class SVM is generally used in several applications and few of these include optical character recognition, intrusion detection, speech recognition, facial expression recognition.

Rahim et al., (2019) [22] proposed a dynamic hand gesture recognition system and tested on own dataset with 15 labels. In the pre-processing stage, the authors used YCbCr colour space to detect colour of human skin by applying to the region of interest (ROI) images. After detecting the human skin colour erosion, the process of obtaining the holes was performed on the binary image. The feature extraction was done by using deep learning applied over original ROI image and pre-processed image and then the features were combined. The recognition accuracy was found as 96.96 % with the use of softmax function. The literature [23, 24] implemented end-to-end hand gesture recognition system by using convolutional neural network (CNN) with any auxiliary method for pre-processing purpose. Hoang et al., established dynamic hand gesture

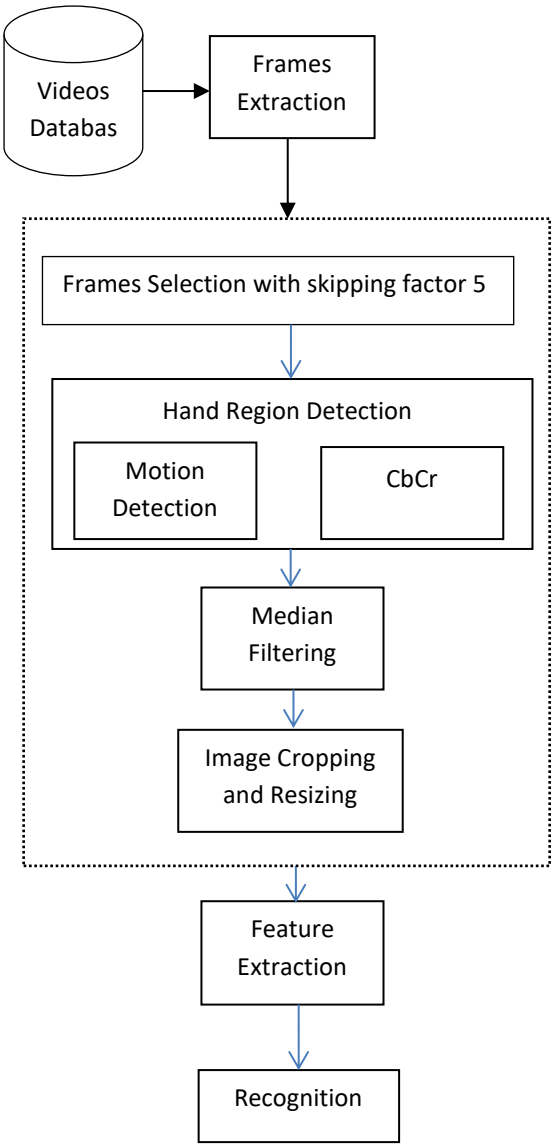

Figure. 2 The architecture of the propose system

recognition system using deep learning and key frames selection. The system is selected number of key frames N=3, 4, 5, 6 and 5 key frames obtained the highest accuracy. The recognition accuracy is 77.71 % on SHG dataset and 97.8% on Cambridge hand gesture benchmark dataset. The limitation of this paper is heavy computation [25]. Mangla et al., (2020) implemented sign language recognition system using novel key frames selection method. The numbers of selected key frames are 3 and used Discrete Wavelet Transform (DWT) to extract features. The system is implemented on self-constructed Pakistani sign dataset. The system cannot solve the hand and face overlap condition [26]. Azar et al., (2020) established trajectory based Persian sign language recognition system on self-constructed dataset. The recognition accuracy is reached 97.48% based on trajectory and shape features. The number of selected frames is 30. The signer wears white colour glove when capturing the data that is naturalness [27]. Kraljevic` et al., (2020) implemented Croatian sign language recognition system by using end-to-end process of deep learning. The numbers of selected frames are 8, 16 and 32 from all of videos and input size 32 frames are obtained highest recognition accuracy. The system is heavy computation and 70.1% on combined SHSL-RGD and SHSL-Depth data [28]. In [29], the author proposed hand gesture recognition based on the shape and orientation features using the HOG feature extraction method. The various size of the hand has degraded the performance. So, our method used scale and orientation based features using the GIST feature descriptor. Most of the existing research works in Myanmar use static images for sign language recognition and we have used videos in our work.

3. Methodology for static and dynamic gesture recognition

The overview architecture of our proposed system is described in the following Figure. The dynamic Myanmar sign word recognition system is implemented based on hybrid hand region detection and hand-craft shape, orientation features.

3.1 Frame selection and hand region detection

The dynamic sign word video for study contains the large number of image frames. In the frame's sequences, many redundant frames are also present and such frames are not required to be recognized as valid sign word gestures and therefore only few important frames are sufficient. The amount of data and time is reduced by eliminating the redundant frames; else it would have needed extra time as well as data storage. In our experiment, the frames are extracted through skipping five frames or using skipping factor 5 from the sequences of frames with the following equation.

$$SF = [F(i), (i = 1 \leq NOF; i + c)] \quad (1)$$

$$SF = \{F_{i1}, F_{i2}, \dots\dots, F_{ik}\} \leq T_{NSF} \quad (2)$$

In Eq. (1), SF is the sequence of selected frames. The constant c is 5 and NOF is the total number of frames in video. In Eq. (2), k is the size of selected frames and TNSF is the total number of selected frames by using skipping factor 5. Hand region detection (HRD) is the first and most crucial process that needs to solve for the hand gesture recognition

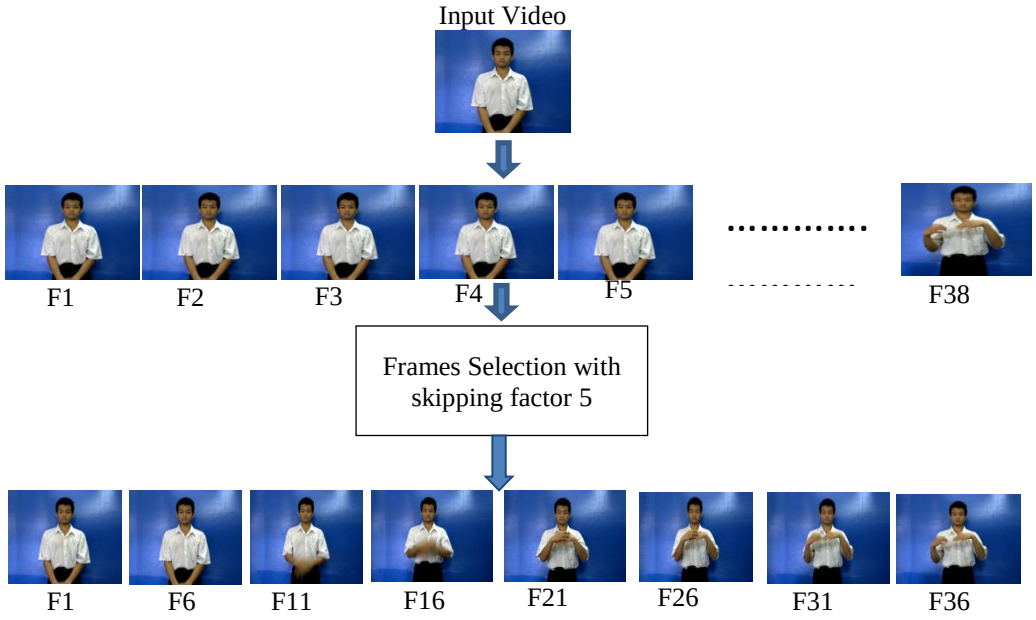

Figure. 3 The frames selection by using skipping factor 5

system. Hand region detection is the extracting only hand region from the background by eliminating the background, face and other not-hand region. There are many hand region detection methods such as colour threshold-based, region-based, edge-based, pixel-based, model-based, ANN-based, watershed-based, and clustering-based and many other hand detection methods. The colour threshold-based method is the most usable approach in the start-of-the-art study. However, the colour threshold-based method faces the challenge when removing the face, arm and skin-like object in the background. To address this problem, we proposed the new hand region detection process. In the proposed detection process used hybrid segmentation method CbCr and Motion detection (CbCr+Motion) to detect the hand from the background region. The proposed hand region detection process can be solved the issue of miss hand region detection when the background contain like human skin colour such as wood, door and skin colour curtain by motion mask and eliminate non-skin moving object by skin mask.

3.1.1. YCbCr skin colour segmentation

There different ten colour space used to detect the hand using human skin colour [30]. YCbCr is often used for hand region extraction or human skin colour detection [31]. The Y is the luminance channel and CbCr is the chrominance channel and there are many colour spaces to detect the human skin colour such as RGB, YCbCr, HSV, Lab, YIQ, CMYK and many others. According to the state-of-art research, YCbCr colour space model is the most suitable for human skin detection as compared to other colour space models. In YCbCr colour space, we eliminate the luminance Y because the best result can be achieving only after changing environment lighting condition. Firstly, the original RGB image is converted into YCbCr colour image by using the Eq. (3).

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \frac{1}{256} \times \begin{bmatrix} 65.738 & 129.057 & 25.064 \\ -37.945 & -74.494 & 112.439 \\ 112.439 & -94.154 & -18.285 \end{bmatrix} \times \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (3)$$

The skin region image is defined on the foreground region image using the standard range threshold values. The appropriate threshold value of chrominance channel (Cb,Cr) is used to detect the human skin colour using the Eq. (4). We defined the skin colour when the chrominance Cb channel value is between (77, 127) and the Cr channel value is between (133,173) [32]. The literature [33], [34] and [35] and also used YCbCr colour space with different threshold values.

$$I_{SkinRegion} = \begin{cases} 1, if(77 < Cb < 127 and 133 < Cr < 173) \\ 0, \quad otherwise \end{cases} \quad (4)$$

3.1.2. Motion detection

In general, there are three basic movement (motion) detection methods such as background subtraction, optical flow and frame differencing. Motion detection or Background Subtraction is to detect the motion of hand while doing the sign gesture. The first step of background subtraction is to establish the background. In the proposed system, the background is estimated by using mean filtering with the Eq. (5). N is the number of frames in the video. BF is the background frame.

$$BF(x,y,t) = \frac{1}{N}\sum_{i=1}^{N} I\,(x,y,t-i) \tag{5}$$

It segments the hand motion region by using the difference between the background frame and input frames with the Eq. (6). The background subtraction method is simple and quick method. The key advantage of background subtraction method is more efficient when the background is stable, quickly and easily.

$$I_{MotionRegion} = \begin{cases} 1, if\,(CF(x,y,t) - BF(x,y,t) > T) \\ 0, \quad otherwise \end{cases} \tag{6}$$

In hand region segmentation, the AND operation performed on skin region and motion region to get the only hand region with the following Eq. (7). In our experiment, the value of threshold T is 25.

$$I_{i(HandRegion)} = \prod_{i=1}^{n} (I_{(iSkinRegion)}, I_{(iMotionRegion)}) \tag{7}$$

3.1.3. Median filtering

Median filtering technique is applied on the hand region image to remove the salt and pepper noise. In median filtering, the size of filter is 5 (5×5) is the most suitable than other filter sizes such as (3×3, 7×7 and 9×9). The results of filtering image obtain by the following Eq. (8). After the median filtering, the filtered image remained small connected salt and pepper noise region. All the connected regions are then sorted by descending order of area and the largest connected area is represented as hand region [1].

$$I_{i(filterimage)} = \left[I_{i(HandRegion)} * filter\right]^{i=1.....n} \tag{8}$$

Where, the * is the filter operation and the size of filter is 5.

3.1.4. Image cropping and resizing

A bounding box is created on the filtering hand region image and then the image is cropped by using the bounding box. After that, the cropped image is resized to 32×32 image size. The block diagram of the hand region detection process is shown in the following Fig. (4). The processing output of the combination of segmentation is shown in the following Fig. (5).

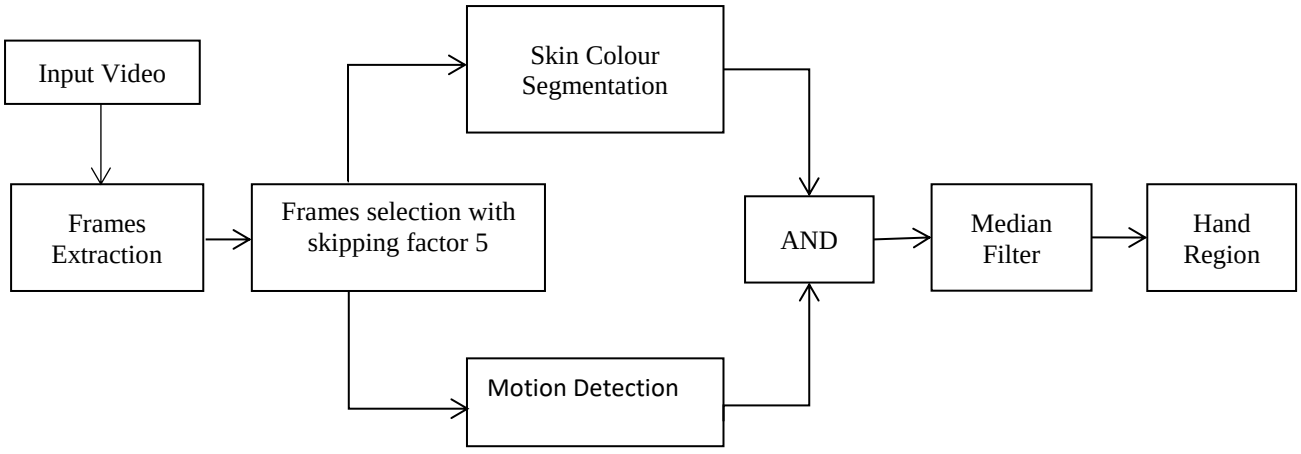

Figure. 4 Block diagram of hand region detection

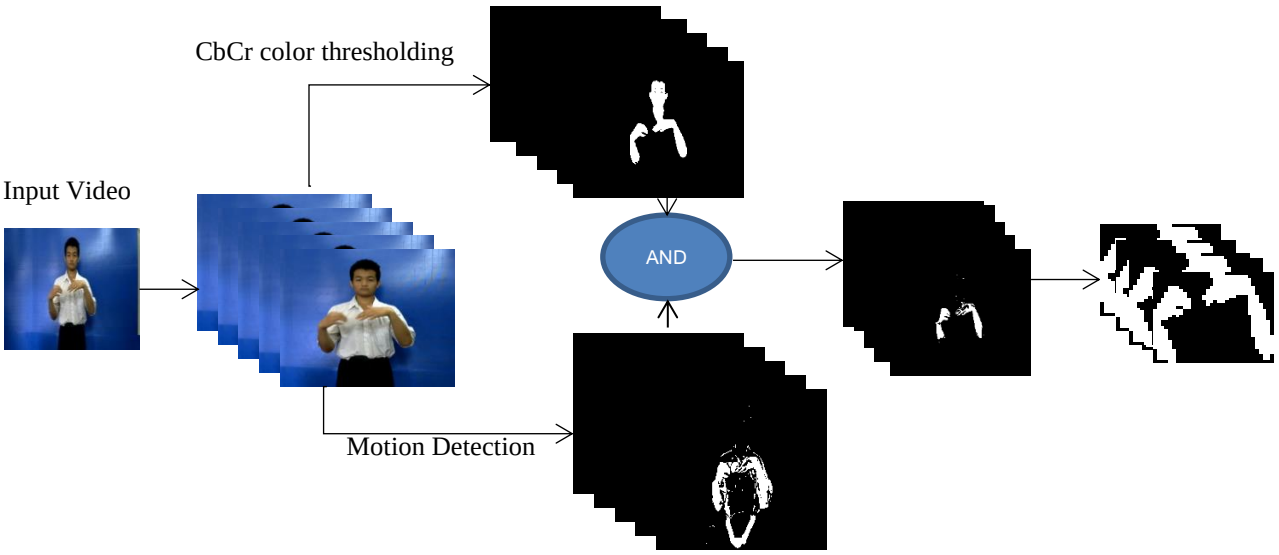

Figure. 5 Hand region detection process from input video

3.2 Feature extraction

The shape, orientation, texture and (scale or size) features are very useful in the task of sign language recognition using hand gesture and thus we also have used these features for hand gesture recognition. There are many feature extraction methods available in literature for this purpose such as GIST, HOG, HOG-LBP, SURF (speeded up robust feature), CNN (Convolutional Neural Network) and SIFT (scale invariant feature transform). Among these methods, Generalized Search Tree (GIST), Histogram of Gradient (HOG), HOG-LBP (Local Binary Pattern) feature descriptor method is chosen which emphasizes on orientation, scale, texture, and structure or shape of an object. The HOG also provides the information related to the edge direction as well and rotation invariant [36]. The GIST descriptor is extracted features emphasis on the shape, orientation and local properties of the image that represent abstract representation for overall image [37]. The LBP descriptor extracts texture information from hand gestures [10].

3.2.1. GIST descriptor

The global GIST feature descriptor obtained good result for scenes classification. The benefit of the GIST descriptors is compact and fast to compute. The GIST feature descriptor extracts the main different orientation and frequencies contained in the image. There are three parameters in GIST feature descriptors such then number of orientations, number of scales and the number of grids. The main parameters of the GIST descriptor are tuned. The experiments are performed with 8 orientations, (4 and 5) different scales and the (4×4 to 9×9 or 16 regions to 81 regions) grid size for each image. In our experiments, the number of features dimensions increased when the numbers of grid size and scale values are increased. So, the time complexity is more increased but the recognition accuracy is not increased. Among different scale, different orientation and different grid size: the number of orientations =8, the number of scales=4 and the number of grid size = 4×4 was obtained the best accuracy. Finally, the GIST descriptor concatenates 32 features map for 16 regions. The lengths of feature vector are 512 for each image.

3.2.2. HOG descriptor

The detailed HOG process is described as:

Step 1: The input image for pre-processing is divided into 8×8 pixel patches called cell to extract the features.

Step 2: The gradient Gx and Gy for every pixel in the image are calculated using following equations. The gradients indicate the small changes in the x and y direction respectively.

$$Gx = (x, y+1) - (x, y-1) \tag{9}$$

$$Gx = (x+1, y) - (x-1, y) \tag{10}$$

Step 3: The magnitude and direction (orientation) for each pixel value is determined.

$$Magnitude(g) = \sqrt{(Gx^2 + Gx^2)} \tag{11}$$

$$Direction(\theta) = \tan^{-1}\left(\frac{Gy}{Gy}\right) \tag{12}$$

Step 4: We calculate the histogram of gradient for each cell by using 9 bin histogram that obtained 1×9 features matrix for each cell.

Step 5: We then normalize the features for 16×16-pixel pitches called block. The overlapping blocks are used to handle the illumination changes. The normalization is calculated by using the following equation.

$$V = \lfloor f_1, f_2, f_3 \dots\dots\dots, f_{36} \rfloor \quad (13)$$

$$k = \left(\sqrt{\left({f_1}^2, {f_2}^2, {f_3}^2 \dots\dots\dots, {f_{36}}^2 \right)} \right) \quad (14)$$

The root of the sum of the square calculates with the following equation.

$$NormalizedVector = \left[\frac{f_1}{k}, \frac{f_2}{k}, \frac{f_3}{k}, \dots\dots, \frac{f_{36}}{k}, \right] \quad (15)$$

In normalization step, the all of the values in the vector V is divided with the value k.

3.3 Hand gesture recognition

The recognition is the final process and also important step of the system. The role of classifier is important to increase the performance of the system. In our research, the three different classifiers SVM (Quadratic), SVM (Cubic) and Linear discriminant are used to recognize hand gesture based static and dynamic sign recognition. The experiments of the system are performed in 10 times runs and the performance of the system is evaluated in term of mean accuracy.

4. Evaluation result

All experiments were performed on Intel Core 2.3 GHz processor with 8 GB RAM and 1280×720 resolution pixels dynamic Myanmar sign words gesture videos.

4.1 Dataset used

The performance of the system is evaluated on two different datasets. The first dataset is public available static posture dataset and the next one is self-constructed dynamic Myanmar sign words dataset.

MUDB dataset: MUDB dataset is published from Massey University called MU_HandImages_ASL [38]. This dataset contains American Sign Language postures with five different illumination conditions (left, right, top, bottom and diffuse), five volunteers and different hand size. The experiments are performed with (A-Z) ASL alphabets and 65 images for each sign.

The dynamic Myanmar sign word dataset contains 17 sign words are collected by using a web camera with resolution 1280 X 720 px. In this system, we used both male and female deaf signers in School for the Deaf Mandalay, Myanmar. The videos are recorded in mp4 format with frame rate of 25 fps. The duration of gesture length is between 1.5 s and 2.5 s.

The video data are recorded in indoor environment under normal lighting conditions. The hand gesture videos are captured without the aid of any data gloves or wearable tracking devices or special markers like other systems. We collected 2 times for each gesture from 9 subjects. There are 18 videos for each gesture. So, the database contains of 306 video files (13857 frames) for 17 isolated dynamic sign gestures. The dataset contains 17 dynamic Myanmar sing words as Amaranth, Bamboo shoot, Cauliflower, Chayote, Chinese Chives,

Table 1. List of dynamic Myanmar sign words in the dataset

Sign Code	Sign	Sign Code	Sign
1	Amaranth	10	Durian
2	Bamboo Shoot	11	Long Bean
3	Cauliflower	12	Mango
4	Chayote	13	Mangosteen
5	Chinese Chives	14	Papaya
6	Chinese Cabbage	15	Pear
7	Coriander	16	Pomelo
8	Cucumber	17	Pumpkin
9	Custard Apple		

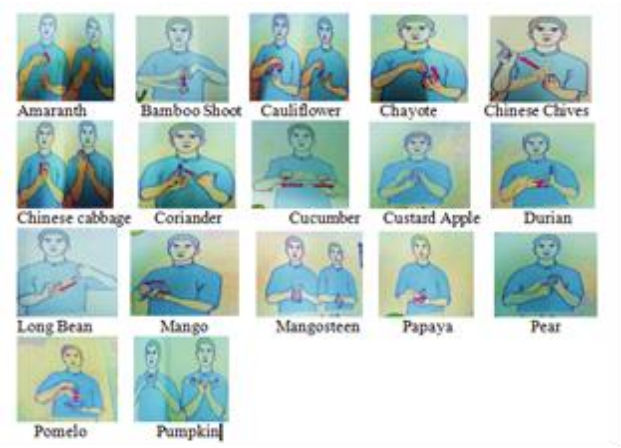

Figure. 6 Sample gesturing sign words images in the dataset

Chinese Cabbage, Coriander, Cucumber, Custard Apple, Durian, Long Bean, Mango, Mangosteen, Papaya, Pear, Pomelo and Pumpkin. The list of sign words is described in Table 1. The samples of 17 dynamic Myanmar sign words are shown in Fig. 6.

4.2 Data analysis

In our experiment, we used cross validation with the value of k=10. In 10-Fold-Cross validation procedure, the cross validation is performed with ten iterations. The dataset is divided into ten subsets. In the first iteration, the first subset is selected as testing samples and the rest of nine subsets are used for training. In the second iteration, the next second subset is selected for testing and the rest nine subsets are chosen as training. Therefore, the numbers of iterations are repetitive ten times. Finally, the recognition accuracy is obtained by computing the mean accuracy.

The MUDB and sign word dataset is static dataset and contain segmented hand region images. So, the features directly extracted without using any hand region detection process. In self-constructed dataset, the total number of selected frames 3 frames, 4 frames and 5 frames (k=3, 4, 5) for all videos in our dataset. The features are extracted on 3 frames, 4 frames and 5 frames by using GIST, HOG and HOG-LBP for all videos. The extracted features are concatenation and then classification by using three different classifiers (quadratic SVM, Cubic SVM and Linear Discriminant). The dimensions of HOG features are 972, 1296 and 1620 and GIST features are 1536, 2048 and 2560 for 3 frames, 4 frames and 5 frames respectively.

The performance comparison table 2 describes the results of different feature extractions (HOG, HOG-LBP, GIST) and different classification methods (Cubic SVM, Quadratic SVM and Linear Discriminant) on MUDB dataset. The table 3 describes the result of mean accuracy on self-constructed dynamic Myanmar sign words with different features extraction methods, different classification methods and hybrid segmentation. According to the results, it is visible that the best recognition accuracy results were obtained by using using 5 (k=5) frames selection for all videos. The confusion matrix of Hybrid segmentation with 5 frames selection for all videos in our dataset is described in Fig. 7. The confusion matrix with best classification accuracy on MUDB datasets is described in Fig. 8.

4.3 Discussion with other methods

The proposed method compares with other features extraction methods: HOG [29] and HOG-LBP [10]. The HOG extracted orientation and shape based features vector from hand region images for twos datasets. The HOG features vector size is changed based on image size. So, the features vector size increased when the image size increased. HOG features extraction method has degraded the performance when scale noise increased. In HOG-LBP, the texture information extracts using LBP and combined with contour information extracted by using HOG. The combined features obtained slightly higher accuracy than the proposed method on the MUDB static dataset. However, the result of proposed method is higher than other twos feature extraction methods on dynamic Myanmar sign words dataset. The classification process performed with three different classifiers. Among these three classifiers, the linear discriminant classifier takes less time and obtained more accuracy on the dynamic Myanmar sign words dataset.

Table 2. the comparison of different features extraction and different classifiers on MUDB dataset

Datasets	Feature Extraction	SVM (Quadratic)	SVM (Cubic)	Linear Discriminant
MUDB	HOG[29]	98.6%	98.7%	98.0%
	HOG-LBP[10]	99.19%	99.22%	98.83%
	GIST	99.01%	99.03%	98.71%

Table 3. the comparison of different features extraction and different classifiers with hybrid segmentation on self-constructed dynamic Myanmar sign words dataset

Frame Selection	Feature Extraction	Classifier	Training Time	Accuracy
3 frames selection with skipping factor 5	HOG [29]	SVM(Cubic)	40s	68.75%
		SVM(Quadratic)	40s	68.84%
		Linear Discriminant	7s	71.18%
	HOG-LBP [10]	SVM(Cubic)	32s	67.46%
		SVM(Quadratic)	38s	68.86%
		Linear Discriminant	10s	71.79%
	GIST	SVM(Cubic)	31s	69.61%
		SVM (Quadratic)	39s	69.74%
		Linear Discriminant	10s	77.46%
4 frames selection with skipping factor 5	HOG [29]	SVM(Cubic)	36s	72.20%
		SVM(Quadratic)	36s	73.33%
		Linear Discriminant	8s	78.48%
	HOG-LBP [10]	SVM(Cubic)	45s	72.71%
		SVM(Quadratic)	44s	73.23%
		Linear Discriminant	10s	78.95%
	GIST	SVM(Cubic)	48s	74.95%
		SVM(Quadratic)	48s	75.00%
		Linear Discriminant	13s	81.90%
5 frames selection with skipping factor 5		SVM(Cubic)	41s	76.86%
	HOG [29]	SVM(Quadratic)	42s	77.52%
		Linear Discriminant	10s	83.41%
	HOG-LBP [10]	SVM(Cubic)	48s	77.18%
		SVM(Quadratic)	48s	77.09%
		Linear Discriminant	14s	84.02%
	GIST	SVM(Cubic)	60s	79.09%
		SVM(Quadratic)	59s	79.13%
		Linear Discriminant	19s	86.70%

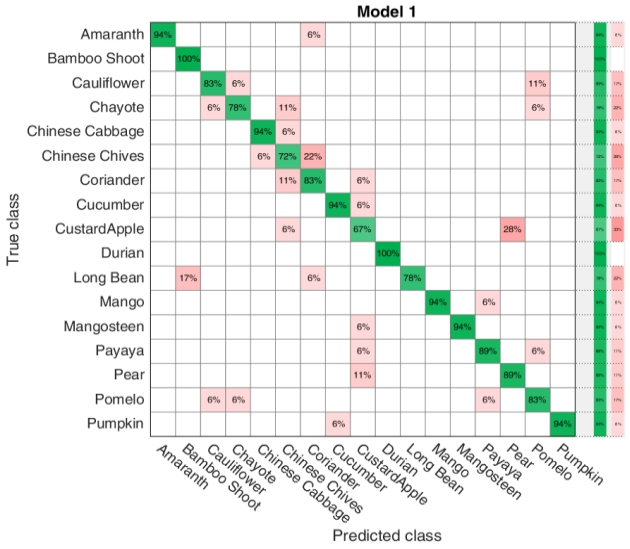

Figure. 7 Confusion matrix on self-constructed dataset

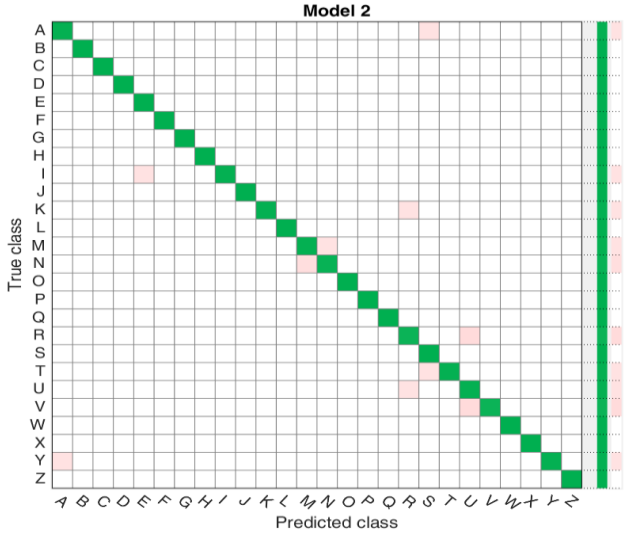

Figure. 8 Confusion matrix on MUDB

5. Conclusion

The system demonstrates vision based static and dynamic hand gesture recognition system. In dynamic, the system used skipping factor 5 for frames selection to reduce the redundant frames and computation time. The combination of CbCr colour channel from YCbCr colour space and motion detection process from background subtraction used to pre-process the original input frames of dynamic Myanmar sign words. Then, generalized search tree, histogram of gradient oriented (HOG) and HOG-LBP (Local Binary Pattern) is used to extract features form each pixel and compared them. In the classification stage, Quadratic SVM, Cubic SVM and linear discriminant was applied on the self-constructed dynamic Myanmar sign words dataset and MUDB dataset. The best recognition accuracy was achieved 86.7 % and 99.22%. The sing gestures performed by two hands and dynamic hand gesture recognition have many challenging problems. There is no standard dataset and has only little research for Myanmar dynamic sing gestures. In the future, we planned to construct big dynamic Myanmar sign words dataset with more subjects and then will implements with deep learning approach on the dataset. In the field of hand gesture detection and recognition, the vision based hand gesture recognition is more challenging because the highly sensitivity of individual differences, different background condition, illumination changing and skin like object.

Conflicts of Interest

The authors declare no conflict of interest with respect to the research, authorship, and publication of this article.

Author Contributions

Phyu Myo Thwe contribute to the design and implementation of the research, analysis, dynamic Myanmar sign word dataset and writing the the paper. Dr.May The` Yu supervised the research.

Acknowledgments

I would like to special thanks, Dr. May The` Yu, Professor, Faculty of Information Science, University of Computer Studies, Mandalay, for her continuous guided, supporting and suggestions. I would like to special thanks the signers from School for the Deaf, Mandalay, Myanmar.

References

[1] M. Mohandes, M. Deriche, and J. Liu, “Image-Based and Sensor-Based Approaches to Arabic Sign Language Recognition”, *IEEE Trans. Hum.-Mach. Syst.*, Vol. 44, No. 4, pp. 551–557, 2014.

[2] Steinberg, T. M. London, and D. D. Castro, “Hand Gesture Recognition in Images and Video”, *Department of Electrical Engineering*, p. 1- 20, 2010.

[3] Dadashzadeh, A. T. Targhi, M. Tahmasbi, and M. Mirmehdi, “HGR-Net: A Fusion Network for Hand Gesture Segmentation and Recognition”, *IET Computer Vision*, Vol.13, No.8, pp-700-707, 2019.

[4] N. H. Aung, Y. K. Thu, and S. S. Maung, “Feature Based Myanmar Fingerspelling Image Classification Using SIFT, SURF and BRIEF”, *Seventeenth International Conference on*

Computer Applications (ICCA 2019), p. 11, 2019.

[5] J. P. Sahoo, S. Ari, and D. K. Ghosh, "Hand gesture recognition using DWT and F-ratio based feature descriptor", *IET Image Process.*, Vol. 12, No. 10, pp. 1780–1787, 2018.

[6] M. A. Rahim, M. R. Islam, and J. Shin, "Non-Touch Sign Word Recognition Based on Dynamic Hand Gesture Using Hybrid Segmentation and CNN Feature Fusion", Appl. Sci., Vol. 9, No. 18, p. 3790, 2019.

[7] P. Bao, A. I. Maqueda, C. R. del-Blanco, and N. García, "Tiny hand gesture recognition without localization via a deep convolutional network", *IEEE Trans. Consum. Electron*, Vol. 63, No. 3, pp. 251–257, 2017.

[8] D.-L. Lee and W.-S. You, "Recognition of complex static hand gestures by using the wristband-based contour features", *IET Image Process*., Vol. 12, No. 1, pp. 80–87, 2018.

[9] D. A. Contreras Alejo and F. J. Gallegos Funes, "Recognition of a Single Dynamic Gesture with the Segmentation Technique HS-ab and Principle Components Analysis (PCA)", *Entropy*, Vol. 21, No. 11, p. 1114, 2019.

[10] H. Lahiani and M. Neji, "Hand gesture recognition method based on HOG-LBP features for mobile devices", *Procedia Comput. Sci*., Vol. 126, pp. 254–263, 2018.

[11] M. A. Moni and A. B. M. S. Ali, "HMM based hand gesture recognition: A review on techniques and approaches", In: *Proc. of 2009 2nd IEEE International Conference on Computer Science and Information Technology*, Beijing, China, pp. 433–437, 2009.

[12] N. Tubaiz, T. Shanableh, and K. Assaleh, "Glove-Based Continuous Arabic Sign Language Recognition in User-Dependent Mode", *IEEE Trans. Hum.-Mach. Syst.*, Vol. 45, No. 4, pp. 526–533, Aug. 2015.

[13] Q. Zheng, X. Tian, S. Liu, M. Yang, H. Wang, and J. Yang, "Static Hand Gesture Recognition Based on Gaussian Mixture Model and Partial Differential Equation", *International Journal of Computer Science (IAENG)*, Vol.45, No.4, p. 15, 2018.

[14] K. B. Shaik, P. Ganesan, V. Kalist, B. S. Sathish, and J. M. M. Jenitha, "Comparative Study of Skin Colour Detection and Segmentation in HSV and YCbCr Colour Space", *Procedia Comput. Sci*., Vol. 57, pp. 41–48, 2015.

[15] L. Y. Deng, J. C. Hung, H.-C. Keh, K.-Y. Lin, Y.-J. Liu, and N.-C. Huang, "Real-time Hand Gesture Recognition by Shape Context Based Matching and Cost Matrix", *J. Netw.*, Vol. 6, No. 5, pp. 697–704, May 2011.

[16] Q. Zhang, J. Lu, M. Zhang, H. Duan, and L. Lv, "Hand Gesture Segmentation Method Based on YCbCr Colour Space and K-Means Clustering", *Int. J. Signal Process. Image Process. Pattern Recognit.*, Vol. 8, No. 5, pp. 105–116, 2015.

[17] S. Thakur, S. Paul, A. Mondal, S. Das, and A. Abraham, "Face detection using skin tone segmentation", In: *Proc. of 2011 World Congress on Information and Communication Technologies*, Mumbai, India, pp. 53–60, 2011 .

[18] D. J. Sawicki, and W. Miziolek, "Human colour skin detection in CMYK colour space", *IET Image Process*., Vol. 9, No. 9, pp. 751–757, 2015.

[19] S. Reshna, and M. Jayaraju, "Spotting and recognition of hand gesture for Indian sign language recognition system with skin segmentation and SVM", In: *Proc. of 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai* , pp. 386–390, 2017.

[20] H. Tang, H. Liu, W. Xiao, and N. Sebe, "Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion", *Neurocomputing*, Vol. 331, pp. 424–433, 2019.

[21] B. Pathak, A. S. Jalal, S. C. Agrawal, and C. Bhatnagar, "A framework for dynamic hand Gesture Recognition using key frames extraction", in 2015 *Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, Patna, India, pp. 1–4, 2015.

[22] M. A. Rahim, J. Shin, and M. R. Islam, "Dynamic Hand Gesture Based Sign Word Recognition Using Convolutional Neural Network with Feature Fusion", In: *Proc. of 2019 IEEE 2nd International Conference on Knowledge Innovation and Invention (ICKII)*, Seoul, Korea (South), pp. 221–224, 2019.

[23] Köpüklü, A. Gunduz, N. Kose, and G. Rigoll, "Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks", In: *Proc. of 14th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1-8, 2019.

[24] J. Gangrade, and J. Bharti, "Vision-based Hand Gesture Recognition for Indian Sign Language Using Convolution Neural Network", *IETE Journal of Research*, p. 1-10, 2020.

[25] N. N. Hoang, G.-S. Lee, S.-H. Kim, and H.-J. Yang, "Effective Hand Gesture Recognition by Key Frame Selection and 3D Neural Network",

Korean Inst. Smart Media, Vol. 9, No. 1, pp. 23–29, 2020.

[26] F. U. Mangla, A. Bashir, I. Lali, A. C. Bukhari, and B. Shahzad, "A novel key-frame selection-based sign language recognition framework for the video data", *Imaging Sci. J.*, Vol. 68, No. 3, pp. 156–169, 2020.

[27] S. G. Azar and H. Seyedarabi, "Trajectory-based recognition of dynamic Persian sign language using hidden Markov model", *Comput. Speech Lang.*, Vol. 61, p. 101053, 2020.

[28] L. Kraljević, M. Russo, M. Pauković, and M. Šarić, "A Dynamic Gesture Recognition Interface for Smart Home Control based on Croatian Sign Language", *Appl. Sci.*, Vol. 10, No. 7, p. 2300, 2020.

[29] V. Sombandith, A. Walairacht, and S. Walairacht, "Hand gesture recognition for Lao alphabet sign language using HOG and correlation", In: *Proc. of 2017 14th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Phuket, pp. 649–65, 2017.

[30] J. M. Chaves-González, M. A. Vega-Rodríguez, J. A. Gómez-Pulido, and J. M. Sánchez-Pérez, "Detecting skin in face recognition systems: A colour spaces study", *Digit. Signal Process*., Vol. 20, No. 3, pp. 806–823, 2010.

[31] J. Ekbote, and M. Joshi, "Indian sign language recognition using ANN and SVM classifiers", In: *Proc. of* 2017 *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, pp. 1–5, 2017.

[32] D. Chai and K. N. Ngan, "Face segmentation using skin-colour map in videophone applications", *IEEE Trans. Circuits Syst. Video Technol*., Vol. 9, No. 4, pp. 551–564, 1999.

[33] T. J. McBride, N. Vandayar, and K. J. Nixon, "A Comparison of Skin Detection Algorithms for Hand Gesture Recognition", in 2019 Southern African Universities Power *Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, Bloemfontein, South Africa, pp. 211–216, 2019.

[34] D. K. Ghosh and S. Ari, "Static Hand Gesture Recognition Using Mixture of Features and SVM Classifier", in 2015 *Fifth International Conference on Communication Systems and Network Technologies*, Gwalior, India, pp. 1094–1099, 2015.

[35] S. N. Endah, H. A. Wibawa, and R. Kusumaningrum, "Skin Detection Based on Local Representation of YCbCr Colour Moment", In: *Proc. of 1st International Conferences on Informatics and Computational Sciences (ICICoS)* p. 65-70 , 2017.

[36] T. Kobayashi, A. Hidaka, and T. Kurita, "Selection of Histograms of Oriented Gradients Features for Pedestrian Detection", *Neural Information Processing*, Vol. 4985, M. Ishikawa, K. Doya, H. Miyamoto, and T. Yamakawa, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 598–607, 2008.

[37] Oliva, and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope", *International Journal of Computer Vision*, Vol 42, No. 3, p. 145-175, 2001.

[38] L. C. Barczak, N. H. Reyes, M. Abastillas, A. Piccio, and T. Susnjak, "A New 2D Static Hand Gesture Colour Image Dataset for ASL Gestures", *Res. Lett. Inf. Math. Sci*., Vol. 15, p. 12-20, 2011.