Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 2043–2050

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

Hand sign recognition from depth images with multi-scale density features for deaf mute persons

Taniya Sahana[a], Soumi Paul[b,*], Subhadip Basu[b], Ayatullah Faruk Mollah[a]

[a]*Department of Computer Science & Engineering, Aliah University, Kolkata 700160, India*
[b]*Department of Computer Science & Engineering, Jadavpur University, Kolkata 700032, India*

Abstract

Among many of the fastest growing research fields, sign language recognition is one of the top. Deaf and dumb community uses sign language to express their ideas or views. Sign Language is a methodical coded language where meanings are assigned to every gestures. Many techniques have been developed with the advancement of science and technology to minimize the problem for speech and hearing disabled. The mode of such communication is part of human computer interaction. Hand gesture plays an important role here. The interaction with computer through gesture removes the use of conventional input devices like mouse and keyboards. To create a strong interface between user and computer, recognition of gesture is important. In this paper, a hand gesture recognition method based on multiscale density features is proposed. Depth images of numerals of American Sign Language are considered in this work and recognition rate of 98.20% is obtained, which is comparable with related state-of-the art methods.

Keywords: Sign Language; Gestures Recognition; Human Computer Interaction; Multi-scale Density Features.

1. Introduction

Sign Language is a naturally evolved language like other oral languages. It is used by persons with deafness for daily basis communication. It is considered as a mother tongue of persons with speech impairment. There has not been any standardization of sign languages for hearing impaired people around the globe. Like spoken languages, sign languages are not universally same - they also change when region changes. It is also not possible to find an experienced and qualified interpreters whenever needed. On the contrary, computer can be programmed to translate the sign language to text format and thus it can minimize the distance between normal people and deaf community. Several approaches have been proposed to recognize different hand gestures, that can be broadly grouped into (a)

* Corresponding author. Tel.: +91 33 2413 1766.
E-mail address: spaul.cse.rs@jadavpuruniversity.in

10.1016/j.procs.2020.03.243

sensor based, and (b) vision based. In the first category, the user requires to wear a glove with a sensor or a glove that is colored. During processing, segmentation of the hand portion becomes easy with gloves and it makes the sensor device suitable for digitization of hand as well as finger movements into parametric data. However, this data is often too costly for regular users. On the other side, vision based techniques are more appropriate for real-time applications. In this approach, image processing algorithms are used to detect and track user's hand signs and facial expressions. This is easier compared to other approaches, because wearing additional hardware is not necessary here. However, efficiency of different approaches may differ.

Gesture recognition approaches typically involve three major sub-problems, (i) hand segmentation, (ii) feature extraction, and (iii) classification. Several works have been reported in literature on hand signs recognition. Cabrera et al. [1] proposed a method to recognize hand gestures. A sensor glove was used in this process. A tri-axis accelerometer placed on the back of the hand and the glove conveyed orientation information for each fingers. A neural network was used for classification. In [2, 3], fuzzy rule based classification method was introduced to recognize hand sign gestures. They used accelerometer based hand glove to get a relative angle between fingers and palm of the hand. Li et al. [4] used portable accelerometer and surface electromyographic sensors for automatic Chinese sign language recognition. Continuous sign language sentence is divided into sub words and the three basic components of it i.e. the hand shape, orientation and movement are further modelled and corresponding component classifiers are learned. In [5], a skeleton based method was proposed by Barkoky et al. to recognize numbers 1 to 10 of Persian sign language. Fingertips information were extracted from the end points of skeleton of hand silhouettes. Kaur et al. [6] presented a method for automatic sign recognition using shape features. Otsu's thresholding algorithm was used for hand region segmentation from the images. Gilorkar et al. [7] proposed a method of improvised Scale Invariant Feature Transform (SIFT) for feature extraction. The system was able to recognize a subset of 35 letters of ASL and ISL with an accuracy of 92-96%. In [8], Vieriu et al. reported a method which recognized nine gestures using Hidden Markov Models. Orientation and contour features were extracted from hand silhouettes and used in the recognition process.

Different depth cameras are available in market today, e.g. Microsoft Kinect [9], Creative Senz3D [10], Mesa Swiss-Ranger [11] etc. which have paved the way for new direction of research in hand gesture recognition. For hand gesture recognition system, utilization of depth information is an active topic of research [12]. Isolating hands by depth thresholding is a simpler way that depth camera offers. Subsequent to hand localization and segmentation, several features can be collected from either the Histogram of 3D Facets (H3DF) [13] or Histogram of Oriented Gradients (HOG) [14]. These features are subsequently used for hand gesture recognition. The works [15] and [16] also compared the histogram of center-of-hand to contour distances. In [17], Dominio et al. used a feature combination namely, distance and curvatures of hand contours to recognize hand gestures. Support Vector Machine was used for the classification. Liu et al. [18] represented a model to recognize hand gestures using template matching where Chamfer Distance was used. Furukawa et al. [19] proposed a method for hand detection and fingertips tracking for depth data. The data were obtained through Microsoft Kinect sensor. The method was actually proposed to construct an intelligent room and it was tested in changing environment with complex background. Because of weak classifier, the extracted hand shape was correctly recognized. Wu et al. [20] used 2D appearance feature and TOF sensor. Only ratio and contour based 2D feature were extracted from the hand shapes.

In short, depth based hand sign recognition systems have more advantages than vision based ones. Among different kinds of depth sensors, low cost devices such as Kinect V1 is frequently used to collect input images. In this paper, a publicly available dataset i.e. 10 gesture dataset prepared by [16] is used to evaluate the performance of the multiscale density features for ASL numeral signs recognition. However, the symbols of the above dataset were preprocessed before processing. Rest of the paper is organized as follows. Current work is presented in Section 2, detailed experimental results are reported in Section 3, and finally conclusion is made in Section 4.

2. Current Work

The developed gesture recognition system works in three steps, viz. pre-processing the raw data that is publicly available [16], normalization of orientation and feature extraction. Then, the extracted features are passed to standard pattern classifiers for classification of gesture symbols.

2.1. Data Preprocessing

The RGB-D symbols of the 10 gesture dataset are preprocessed to extract the region of interest (ROI) from the raw data. The ROI regions are obtained with annotated RGB information and histogram thresholding. The cropped ROI regions are subsequently transformed into depth images where the intensity values represent the depth levels. As this dataset was collected using Kinect V1, it was straightforward, because in this case, the depth resolution is 320 x 240 pixels and RGB resolution is double of that, i.e. 640 x 480 pixels. The method is reported in detail in our previous work [21].

2.2. Orientation Normalization

The purpose of orientation normalization is to make the model building process independent of the rotations, movements, stretching and other transformation of the gesture symbols. Following steps are followed for normalization. At first, the centroid of input hand image is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i}{k}, \qquad \bar{y} = \frac{\sum_{i=1}^{k} y_i}{k} \tag{1}$$

Where k is the number of pixels in the region, and x_i and y_i denote the spatial coordinates of the i^{th} pixel in the hand region. Next, the angle of orientation θ, is calculated with respect to the X-axis. Finally, rotation of input gestures by the angle $\varphi = 90 - \theta$ is performed. Orientation normalization for a single numeral class has been shown in Fig. 1.

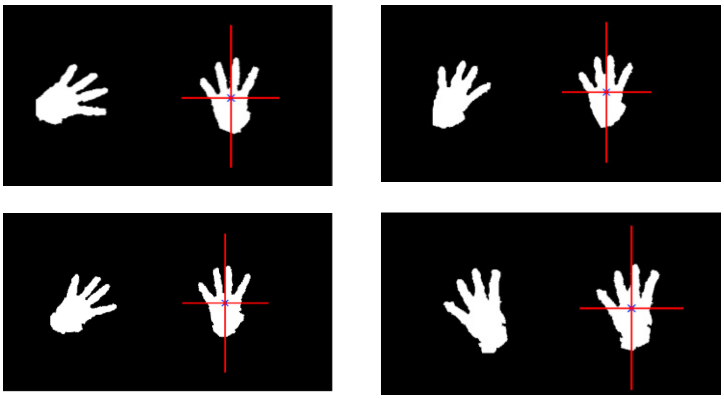

Fig. 1. Orientation normalization of symbols (Some hand symbols of multiple orientations for a single numeral class i.e. 4 have turned to a standard/vertical orientation)

2.3. Feature Extraction

After the orientation normalization features are extracted. A zone based hierarchical framework has been used for feature extraction. In this framework, at each level of the hierarchy input image is subsequently divided into some smaller size images (or zones).

At step 1, a bounding box of the image object is considered for subsequent division. Distance from gesture centroid to left column, right column and top row of the bounding box is calculated. At the 1st level of hierarchy, highest distance among these three is considered as length of a square zone.

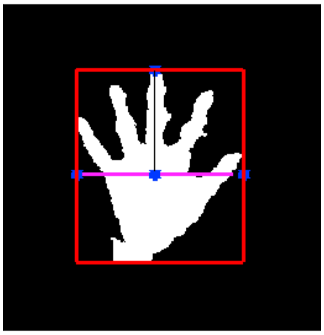

Fig. 2. Finding out the largest distance for length of square zone.

At step 2, based on the centroid input image is divided into two equal parts: upper half and lower half. Upper half consists of all active fingers and hence it is considered in subsequent divisions. Two equal sized square zones are obtained respectively from right and left side of centroid Y-axis: upper right half (z_1) and upper left half (z_2). Consideration of these two equal square zones from input image has been shown in Fig. 3. The zoning process is repeated up to level 4 and at each level density function (*d*) is obtained from each and individual zones, where *d* is defined as shown in Eq. 2.

$$d = \frac{Total\ number\ of information\ pixels\ in\ a\ zone}{Total\ number\ of\ pixels\ in\ a\ zone} \tag{2}$$

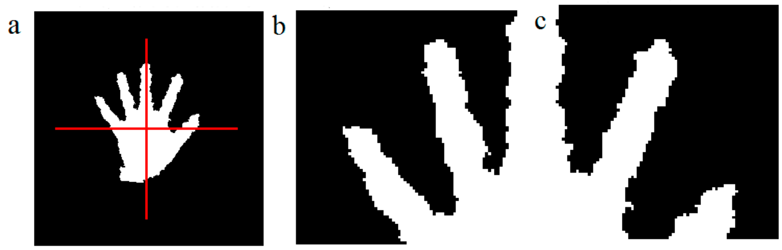

Fig. 3. Two equal square zones have been obtained from an input image. (a) original image (293 rows by 285 columns); (b) upper left half (81 rows by 81 columns); (c) upper right half (81 rows by 81 columns).

Set Z and D will keep a record of updated zones and density features which are obtained at each level of the hierarchy. At level 2, z_1 and z_2 are considered as two separate images. Each image is divided into four equal quadrants. This process is repeated for next two levels of the hierarchy. Division of a sub image into four equal quadrants has been shown in Fig. 4.

Fig. 4. Sub image has been divided into four equal quadrants. (a) upper left half (81 rows by 81 columns); (b) first quadrant (40 rows by 40 columns); (c) second quadrant (40 rows by 40 columns) ;(d) third quadrant (40 rows by 40 columns) ;(e) fourth quadrant (40 rows by 40 columns).

At 1st level of hierarchy $Z = \{z_1, z_2\}$ and $D = \{d_1, d_2\}$.
At 2nd level of hierarchy $Z = \{z_1, z_2 \ldots z_{10}\}$ and $D= \{d_1, d_2\ldots, d_{10}\}$.
At 3rd level of hierarchy $Z = \{z_1, z_2 \ldots z_{42}\}$ and $D = \{d_1, d_2\ldots, d_{42}\}$.
At 4th level of hierarchy $Z = \{z_1, z_2 \ldots z_{170}\}$ and $D = \{d_1, d_2\ldots, d_{170}\}$.

There could be some zones that are empty. So, the value of density of that particular zone image in feature vector is zero. At the end of the 4th level, 170 zones are obtained, so the feature vector is of size 170. This is called multi-scale density feature because each input symbol is divided into some equal sized blocks/zones using different levels of hierarchy. At each level we have obtained some density features.

3. Experimental Results

10 gesture dataset contains 1000 different symbols from 10 different subjects. The extracted 170 density based features are used as input to different classifiers. Random Forest, Support Vector Machine and Multi-Layer Perceptron are used to compute the recognition rates. Cross validation has been used to evaluate the skill of machine learning models. Cross validation is easy to understand and implement. Results in this case generally have lower bias that other cases. Each classifier has used 10 fold cross validation to determine the accuracy. Classifiers are tuned with different parameters. Classifier parameter information and accuracy have been shown in Table 1. A complete description about recognition rate for individual's symbols has been shown in Table 2. A comparative study between the proposed work and other state-of-the-art methods has been shown in Table 3.

Table 1. Classifier parameter information and accuracy.

Classifier used	Parameter information	Accuracy (%)
Random Forest	Number of iteration=100,batch size=100,bag size percentage=100,seed=1,maximum depth size=0	97.30
Support Vector Machine	Seed=1,batch size=100,ploy kernel with catch size=250007,exponent= 1.0,tolerance parameter=0.001	97.80
Multilayer Perceptron	Inputs=170,number of neurons in hidden layer=90,learning rate=0.3,momentum=0.2,training time=500	98.20

From Table 2, it can be noted that recognition rate for gesture zero, one, two, three, four, five and eight are high while recognition rates for six, seven and nine are relatively lesser. The main reason is that hand gesture symbols six, seven and nine are similar in some ways causing the system unable to obtain very distinct feature values.

Table 2. Recognition rate obtained using different classifiers.

Symbol/gesture used	Number of symbol/gesture used	Accuracy using RF (%)	Accuracy using SVM (%)	Accuracy using MLP (%)
0	100	99	99	99
1	100	100	100	100
2	100	100	99	100
3	100	100	100	100
4	100	99	99	99
5	100	100	100	100
6	100	93	97	97
7	100	91	91	91
8	100	100	100	100
9	100	91	93	96

Table 3. Comparative study of our work and other state-of-the art methods

Dataset used	Classification method	Features	Accuracy (%)
American Sign Language (Alphabets and numbers) [22]	Feed forward, back propagation algorithm	Fingertip finder, eccentricity, elongatedness, pixel segmentation and rotation	94.32
American Sign Language Recognition (Alphabets) [23]	Feed forward, back propagation of ANN	Triangle area patches constructed from 3D coordinates	95.00
Thai Sign Language (16 Hand gestures) [24]	Back propagation of neural network	Dimension Measures	83.33
Chinese (20 signs) [25]	Extreme learning machine	Location and Spherical coordinate feature	69.32
Static Indonesian Signs (Alphabets) [26]	SIFT Algorithm	Contours, rectangles, center points	62.60
4 camera-Sign Language Recognition (Alphabets A to Z, Numbers) [27]	ANN back propagation algorithm	Hand shapes	95.10
0-9 Numbers [28]	Support vector machine	Convex points in contour	93.00
Indian Sign Language Recognition (Alphabets) [29]	Dynamic time warping algorithm, nearest neighbor algorithm	shape (scale, rotational and translational invariance)	96.15
10-gesture dataset (1000 different gestures of numbers 0-9)	Random forest, Support vector machine, **Multilayer perceptron**	Density feature (invariant to scale, rotation and translation)	97.30 97.80 **98.20**

A comparison of overall accuracy obtained using different classifiers has been shown in Fig. 5. The performance of different classifiers has been described using confusion matrices in Fig. 6.

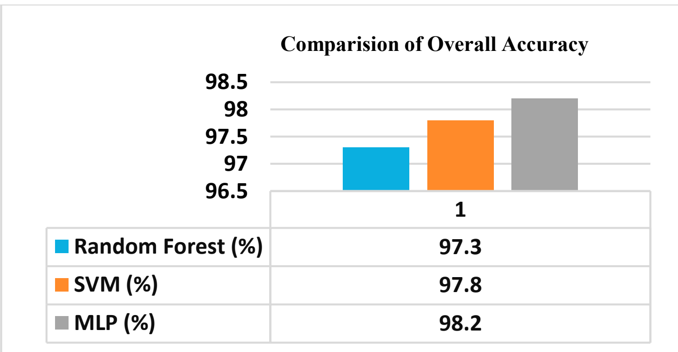

Fig. 5. Comparison of overall accuracy obtained using different classifier.

a

	1	2	3	4	5	6	7	8	9	10
1	99	0	0	0	1	0	0	0	0	0
2	0	100	0	0	0	0	0	0	0	0
3	0	0	100	0	0	0	0	0	0	0
4	0	0	0	100	0	0	0	0	0	0
5	0	0	0	0	99	0	1	0	0	0
6	0	0	0	0	0	100	0	0	0	0
7	0	0	0	0	2	5	93	0	0	0
8	0	0	3	0	0	0	1	91	0	5
9	0	0	0	0	0	0	0	0	100	0
10	1	3	1	0	0	0	0	3	1	91

b

	1	2	3	4	5	6	7	8	9	10
1	99	0	0	0	1	0	0	0	0	0
2	0	100	0	0	0	0	0	0	0	0
3	0	0	99	0	0	0	0	1	0	0
4	0	0	0	100	0	0	0	0	0	0
5	0	0	0	0	99	1	0	0	0	0
6	0	0	0	0	0	100	0	0	0	0
7	0	0	0	0	2	1	97	0	0	0
8	0	0	2	0	0	0	0	91	0	7
9	0	0	0	0	0	0	0	0	100	0
10	0	2	0	0	0	0	0	5	0	93

c

	1	2	3	4	5	6	7	8	9	10
1	99	0	0	0	1	0	0	0	0	0
2	0	100	0	0	0	0	0	0	0	0
3	0	0	100	0	0	0	0	0	0	0
4	0	0	0	100	0	0	0	0	0	0
5	0	0	0	0	99	0	1	0	0	0
6	0	0	0	0	0	100	0	0	0	0
7	0	0	0	0	2	1	97	0	0	0
8	0	0	3	0	0	0	1	91	0	5
9	0	0	0	0	0	0	0	0	100	0
10	0	1	2	0	0	0	1	2	0	96

Fig. 6. Confusion matrix for gesture recognition obtained using (a) RF, (b) SVM, (c) MLP

4. Conclusion

Hand gesture recognition is a challenging problem in designing real life applications for deaf mute community. In this paper, we have presented an efficient method to recognize hand gestures captured with Kinect V1. Experiments on 10 gesture dataset containing hand signs with different orientations is carried out by normalizing orientation of gestures to ensure that the computed feature descriptors are invariant to scale, rotation and translation. Obtained results indicate that our density based feature extraction and recognition method is reasonably accurate. It achieves 98.20% classification accuracy which is comparable with related state-of-the-art methods.

References

[1] Cabrera, Maria Eugenia, Bogado, Juan Manuel, Fermin, Leonardo, Acuna, Raul and Ralev, Dimitar (2012) "Glove-based Gesture Recognition System". *International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines, 10.1142/9789814415958_0095.*

[2] Bui, The Duy and Nguyen, Long Thang (2007) "Recognizing Postures in Vietnamese Sign Language with MEMS Accelerometers", *IEEE Sensors Journal* 7(5): 707–712.

[3] Tabata Y, Kuroda T and Okamoto K (2012) "Development of a glove-type input device with the minimum number of sensors for Japanese finger spelling". *International Conference on Disability, Virtual Reality & Associated Technologies, pp. 305-310.*

[4] Li, Yun, Chen, Xiang, Zhang, Xu, Wang, Kongqiao and Wang, Jane Z (2012) "A sign-component-based framework for Chinese sign language recognition using accelerometer and sEMG data", *IEEE Transactions on Biomedical Engineering* 59(10): 2695-2704.

[5] Barkoky, Alaa and Charkari, Nasrollah Moghadam (2011) "Static hand gesture recognition of Persian sign numbers using thinning method". *International Conference on Multimedia Technology, pp. 6548-6551.*

[6] Kaur, Chandandeep and Gill, Nivit (2015) "An Automated System for Indian Sign Language Recognition", *International Journal of Advanced Research in Computer Science Software Eng*ineering, 5(5):1037-1043.

[7] Gilorkar, Neelam K, Ingle, Manisha M (2014) "Real Time Detection and Recognition of Indian and American Sign Language Using SIFT", *International Journal of Electronics and Communication Engineering & Technology* 5(5):11-18.

[8] Vieriu, Radu Laurentiu, Goras, Bogdan Tudor and Goras, Liviu *(2011) "On HMM static hand gesture recognition," 10th International Symposium on Signals, Circuits and Systems (ISSCS), pp. 1-4.*

[9] Shotton, Jamie, Fitzgibbon, Andrew, Cook, Mat, Sharp, Toby, Finocchio, Mark, Moore, Richard, Kipman, Alex, Blake, Andrew (2011) "Real-time human pose recognition in parts from single depth images", *International Conference on Computer Vision and Pattern Recognition*, *pp. 1297-1304*

[10] She, Yingying, Wang, Qian, Jia, Yunzhe, Gu, Ting, He, Qun and Yang, Baorong (2014) "A real-time hand gesture recognition approach based on motion features of feature points", *IEEE 17th International Conference on Computational Science and Engineering. pp. 1096-1102.*

[11] Kapuscinski, Tomasz, Oszust, Mariusz and Wysocki, Marian (2013) "Recognition of signed dynamic expressions observed by ToF camera", *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), pp. 291-296.*

[12] Suarez, Jesus and Murphy, Robin Roberson (2012) "Hand gesture recognition with depth images: A review", *RO-MAN, IEEE, pp. 411-417.*

[13] Zhang, Chenyang, Yang, Xiaodong and Tian, YingLi (2013) "Histogram of 3D Facets: A characteristic descriptor for hand gesture recognition", *10th IEEE International Conference and Workshop on Automatic. Face Gesture Recognition, pp. 1-8.*

[14] Dalal, Naveneet, Triggs, Bill and Schmid, Cordelia (2010) "Human Detection Using Oriented Histograms of Flow". *Proceedings of the 9th European conference on Computer Vision, pp. 428-441.*

[15] Ren, Zhou, Meng, Jingjing and Junsong, Yuan (2011) "Depth camera based hand gesture recognition and its applications in human-computer-interaction", *Information Communications and Signal Processing 8th International Conference, pp. 1-5.*

[16] Ren, Zhou, Yuan, Junsong, Meng, Jingjing and Zhang, Zhengyou (2013) “Robust part-based hand gesture recognition using kinect sensor”, *IEEE Transactions on Multimedia* 15(5):1110–1120.

[17] Dominio, Fabio, Donadeo, Mauro, Marin, Giulio, Zanuttigh, Pietro and Cortelazzo, Guido Maria (2013) “Hand Gesture Recognition with Depth Data”, *ACM/IEEE International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream, pp. 9-16.*

[18] Liu, Xia and Fujimura, Kikuo (2004) “Hand gesture recognition using depth data”. *IEEE International Conference on Automatic Face and Gesture Recognition, pp. 529-534.*

[19] Takimoto, Hironori, Furukawa, Tatsuya, Kishihara, Mitsuyoshi and Okubo, Kensuke (2012) "Robust Fingertip Tracking for Constructing an Intelligent Room". *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication, pp. 759-763*

[20] Wu, Ying and Huang, Thomas (2001) "Hand modeling analysis and recognition for vision-based human computer interaction" *IEEE Signal Processing Magazine, Special Issue on Immersive Interactive Technology* 18(3)**:**51-60.

[21] Paul, Soumi, Bhattacharyya, Arpan, Mollah, Ayatullah Faruk, Basu, Subhadip and Nasipuri, Mita (2018) “Hand Segmentation from Complex Background for Gesture Recognition”, *International Conference on Emerging Technology in Modelling and Graphics*, pp. 775-782.

[22] Islam, Mohiminul Md, Siddiqua, Sarah and Afnan, Jawata (2017) “Real time hand gesture recognition using different algorithms based on American Sign Language”. *IEEE international conference on imaging, vision & pattern recognition, pp. 1-6.*

[23] Tangsuksant, Watcharin, Adhan, Suchin and Pintavirooj, Chuchart (2014)”American Sign Language Recognition by Using 3D Geometric Invariant Feature and ANN Classification”, *Biomedical Engineering International Conference*, *pp. 1-5*

[24] Chansri, Chana and Srinonchat, Jakkree (2016) “Reliability and accuracy of Thai sign language recognition with Kinect sensor”. *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, pp. 1-4.*

[25] Geng, Lubo, Ma, Xin, Xue, Bingxia, Wu, Hanbo, Gu, Jason and Li, Yibin (2014) “Combining Features for Chinese Sign Language Recognition with Kinect”, *IEEE International Conference on Control & Automation*, *pp. 1393-1398.*

[26] Hartanto, Rudy, Susanto, Adhi and Santosa, P.Insap (2013) “Preliminary design of static Indonesian sign language recognition system”. *International Conference on Information Technology and Electrical Engineering, pp. 187-192*

[27] Kishore, P.V.V., Prasad, Manoj V.D., Prasad, Raghava and Rahul, R. (2015) “4-Camera model for sign language recognition using elliptical Fourier descriptors and ANN”, *International Conference on Signal Processing and Communication Engineering Systems*, *pp. 34-38.*

[28] Lahiani, Houssem, Elleuch, Mohamed and Kherallah, Monji (2015) “Real time hand gesture recognition system for android devices”, *International Conference on Intelligent Systems Design and Applications, pp. 591-596.*

[29] Shukla, Pushkar, Garg, Abhisha, Sharma, Kshitij and Mittal, Ankush (2015) “A DTW and Fourier Descriptor based approach for Indian Sign Language Recognition”, *International Conference on Image Information Processing, IEEE, pp. 113-118.*