International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-8 Issue-6, April 2019

Depth based 3D Indian Sign language Recognition using Adaptive kernels

A.S.C.S. Sastry, A. Ram Kishore, Ch. Bulli Raju, P.V.V. Kishore, D. Anil Kumar, E. Kiran Kumar, M. Teja Kiran Kumar

Abstract*: This work discusses a new approach for Indian sign language recognition (ISLR) using Depth sensor. We propose a novel approach for recognizing sign language gestures from an RGB, depth video sequences by extracting the histogram of oriented gradient (HOG) features and their recognition using adaptive kernel (AK) matching. The kernel-based methods are remarkably effective for recognizing the 2D and 3D actions. This work explores the potential of the adaptive kernels in fusion of RGB and depth kernel scores using ISLR from depth sensor. Accordingly, the HOG features were encoded into adaptive kernels. The recognition is carried out based on the similarity between the query and database features. The performance of the our approach tested on our own 100 class, 5 subjects sign data named as BVCSL3D, captured using Microsoft Kinect v2 sensor and two other publicly available action datasets NTU RGB-D and UTKinect. Our method outperforms when compared to other previous methods on the above datasets.

***Index Terms*: histogram of oriented gradient (HOG), adaptive kernels, Kinect sensor, 3D sign language recognition.**

I. INTRODUCTION

Sign language is the natural communication medium of the hearing-impaired. The signs are characterized by the hand shapes, arm\body movements, head movements. Present, automatic sign language recognition (SLR) system has played an important role in sign gestures convert to text/speech by using human-computer-interface.

Researchers are capturing sign gestures by using differ sensors such as glove-based sensor [1], radio frequency gloves [2], RGB camera [3], leap motion sensor [4], Kinect [5], and Kinect (ToF) [5]. The computer vision-based methods mainly effort on extracting features from 2D images and videos. Digital cameras are used to capturing sign gestures, segmentation and classification of signs very difficult due to various environment noise such as subject invariant, different backgrounds, lighting conditions. The above problems are solved by using 3D sensors. In this paper, we will use an efficient low-cost Microsoft Kinect sensor, that captures both RGB video and depth (RGB-D) information in a range of 2 to 2.5 meters. Depth information supplementing the RGB video sequences has refined sign language recognition.

Over the last few decades, sign language recognition technology has transformed the 1D, 2D to 3D models. In 1D, sign language recognition is based on 1D signals acquired from a glove-based sensor [1], radio frequency gloves [2] and classified using signal processing techniques [1, 2]. These models gave good recognition rates when the emphasis is only on hands. But sign language involves head, hand, torso and face expressions along with finger movements and shapes. The most popular sensor data for sign language recognition task is from video camera, which improves the accuracy of SLR. The data from video camera is widely available at low cost but is influenced by many ambient parameters like background lighting, hand tracking, occultations and capture angle. These parameters are nullified by using 3D models. 3D information from leap motion sensor is good at detecting finger and hand movements, it does not detect other body parts. In this work, we used Microsoft Kinect sensor data to design a 3D Indian sign language machine interpreter to overcome the above difficulties in 2D video-based models.

Figure.1 shows the framework of our proposed SLR system. In this framework, we extract HOG features from RGB and depth sequences and AKs is applied to the extracted features in each 3D sign sequence for classification.

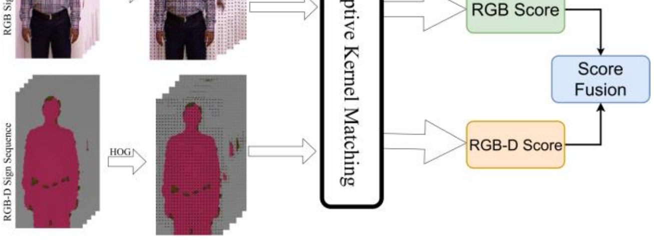

Figure.1. The proposed framework for SLR system

The proposed approach was tested 100 classes of ISL

Revised Manuscript Received on April 07, 2019.

A.S.C.S. Sastry, Department of ECE, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502..

A. Ram Kishore,
Department of ECE, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502..

Ch. Bulli Raju,
Department of ECE, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502..

P.V.V. Kishore,
Department of ECE, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502..

D. Anil Kumar, Department of ECE, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502..

E. Kiran Kumar, Department of ECE, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502..

M. Teja Kiran Kumar, Department of ECE, K.L.E.F. Deemed to be University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh, INDIA – 522502..

dataset (BVCSL3D) for comparison, we used two publicly available action datasets NTU RGBD and UTKinect datasets for validating our proposed algorithm. The proposed method achieved best recognition rates on all the datasets over other state-of-the-art methods.

The rest of the paper is organized as follows: section 2 discusses the related work. Section 3 presents the methodology and procedure. Results and discussion are presented in section 4 and finally, the conclusion is draw in section 5.

II. Related work

In this secession, we briefly review approaches of SLR system. Han et al. [1] reported the applications of gesture recognition system is in human computer interaction that includes sign language recognition, security, gaming and robotics. With the development of SLR system is varied from 1D, 2D to 3D data using multiple signal, image and video processing approaches for feature representation. In 1D SLR use of sensors such as accelerometer or gyroscope, glove-based sensor and radio frequency technology to transmit finger and hand movements to a microcontroller for recognition. The combination of electromyography and 3D accelerometer sensors were used to recognize hand gestures with HMM and Decision Tree based classification, this system has been tested in 72 signs, 40 sentences. Assaleh, K. et al. proposed SLR using sensor-based gloves. In this framework, step wise regression was applied to dimensionally reduction of feature vector for classification. The system was used dynamic time wrapping (DTW) based algorithm with a recognition rate of 95.1%. In, the authors proposed two hand based SLR system using the Cyber-Glove and support vector machine (SVM) to recognize 100 two hand signs from two singers. However, these models will give good recognition rates, but the 1D models is only on the hands. Beside hands, sign language involves hand, head, torso and finger motions. 2D sign video data generates more information compared to 1D data gloves for better recognition.

In the last few decades, many researchers exist a few works that accomplish sign language recognition using 2D image and video sequences. The recognition of sign gestures based on the features are extracted using various image processing approaches like motion, color, shape detection, contour modeling, segmentation. Most of the authors have proposed an optical flow hand tracking and active contour-based segmentation using single camera and the approach was tested on 58 words 10 test subjects with ANN classifier. Yikai Fang et al. have proposed real time adaptive hand tracking and gesture segmentation using motion and color cues. The system tested with 2596 frames recorded for 6 gestures with a recognition of 98% for simple backgrounds and 89% for complex backgrounds. The 2D camera data processing consists of feature extraction and classification is very good. But the difficult due to 2D camera data is background, lighting, blurring and occlusions. The solution for these influences is found with the help of additional information the is somewhat immune to brightness, blurring and occlusions is depth.

In 2010, the Kinect v1.0 was released by Microsoft, it is a popular device for recording depth data in computer vision applications. This sensor used to collect RGB, depth information and 3D skeleton data. Zafrulla et al. proposed a real time system for recognizing hand gestures for hearing impaired people. However, the system performance is decreases due to the tracking errors. Biswas et al. proposed a novel approach for detecting hand gestures using depth images and this approach has been tested on 8 gestures. The depth images were used to remove background of singer's image using histogram approaches and to extract the hand position from rest of the body. This system offers many advantages in image processing techniques using depth images and by employs infrared light for independency of lighting conditions.

In this paper, we are using AKs for finding similarity between the query sign into the database. Recently, the kernel-based methods were performed well and used support vector machine (SVM) to classify the kernel. AKs are being exclusively used for sign language recognition. In this work, we propose a new fusion model where checked by using AK to achieve higher recognition rates compared to other state-of-the-art methods.

Here, the Aks are used to find the maximum similarity between query action and the database actions. This work achieves two major objectives that are currently faced by the skeletal based action recognition algorithms.

1. The process to be independent of the signer and the action frame rate.
2. The database has to be transformed into relational database to show the closeness of the query in the database.

Our proposed architecture is giving its best in classifying or recognizing the human actions from the skeletal data. The detailed methodology is given in next session.

III. Proposed approach

The detailed description of our proposed methodology is given in this section. We start with extracting histogram of oriented gradient using RGB and depth images. The Aks were constructed for each feature type. These AKs are used to find the similarity between the two sign gesture sequences.

A. Feature Extraction

The histogram of oriented gradient (HOG) feature is extensively used for detecting actions or objects. N. Dalal and B. Triggs et al propose to use HOG based method for human action or sign recognition. The proposed method can describe the local object appearance and shape of the gesture within the frame and distribution of local intensity gradients. By calculating HOG features to divides image into several overlapping small region of pixels, called as cell and in the following construction of a 1D histogram.

Let I is the image.

The image is divided into cells of size $N \times N$, and θ is the orientation of each pixel $x = (x_x, x_y)$ is computed by means of following relationship:

$$\theta(x) = \tan^{-1} \frac{I(x_x, x_y + 1) - I(x_x, x_y - 1)}{I(x_x + 1, x_y) - I(x_x - 1, x_y)} \qquad (1)$$

The orientation is collected in a histogram of a predetermined number of bins. Finally, histogram of each cell is collected in a single spatial HOG histogram. The size of feature matrix is dependence on the sign sequence. Figure.2 shows the visualization of the RGB and Depth HOG features.

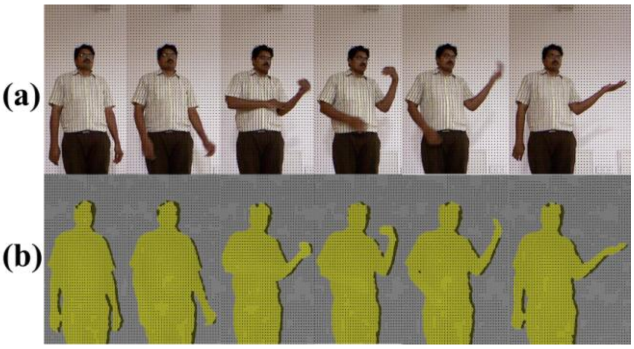

Figure.2. Visualization of the HOG features sign 'Right'

B. Similarity Metric

In the present work, Kinect based sign language recognition based on AKs was implemented. We used HOG features to recognize a sign. The main idea of kernels is to find the similarity between query sign with the database signs to convert input video sequence into text labels. We fuse two kernels for recognizing a sign in the SLR database. Figure 3 shows the block diagram of AK matching proposed in this paper.

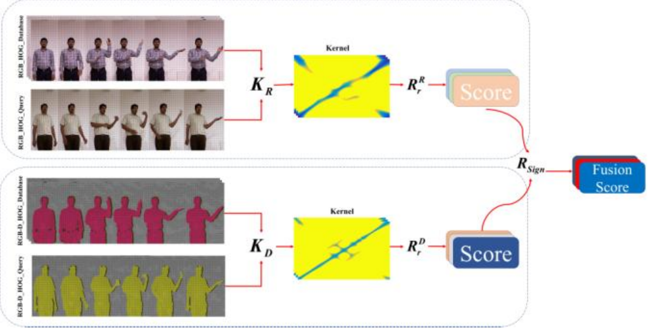

Figure.3. Flow chat of the proposed adaptive kernel matching based sign language recognition

Let $Q(R_{HOG}, D_{HOG})$ and $D^i(R_{HOG}^{'}, D_{HOG}^{'})$ are the query and the dataset signs respectively. Where i is the action index. Here R_{HOG} and D_{HOG} are the metrics of RGB and Depth HOG features of each sign. Where the RGB, depth AKs in a sign data for two signs from query and database is represented as $K_R(R_{HOG}, R_{HOG}^{'})$, $K_D(D_{HOG}, D_{HOG}^{'})$. Here R_{HOG} and $R_{HOG}^{'}$ are the 2D image HOG features of the query sign and the database signs. Similarly, $(D_{HOG}, D_{HOG}^{'})$ is the depth HOG features of the query and the sign from database. The RGB sign kernel is defined as

$$K_R(R_{HOG}, R_{HOG}^{'}) = \exp\left(\frac{\left\| R_{HOG} - R_{HOG}^{'} \right\|_2^2}{2\sigma_1^2} \right) \qquad (2)$$

Where R_{HOG} and $R_{HOG}^{'}$ are the sign RGB HOG features for query and database sign sequence Q and D^i respectively. The gaussian kernel parameter σ_1 is very small $(\sigma_1 > 0)$. The depth kernel is given by

$$K_D(D_{HOG}, D_{HOG}^{'}) = \exp\left(\frac{\left\| D_{HOG} - D_{HOG}^{'} \right\|_2^2}{2\sigma_2^2} \right) \qquad (3)$$

Where D_{HOG} and $D_{HOG}^{'}$ are the depth features in query and database signs. Here $\sigma_2 > 0$ is a gaussian scale parameter.

The constructed two kernels $K_R(R_{HOG}, R_{HOG}^{'})$, $K_D(D_{HOG}, D_{HOG}^{'})$ for a sign with the kernel matrix of size $T_Q \times T_D$. Where T_Q and T_D are the number of frames in query action and dataset action respectively. Cross value analysis shows the matching between query and database. The perfect match gives the maximum score. The proposed algorithm has two advantages: (i) independent of action location in the video frames and (ii) independent of action frames. The results from the one to many frames kernel matching between query and dataset all frames.

The classification of sign label gives the maximum kernel value. The decision boundary for kernel matching scores can be given by

$$R_r^R = \frac{1}{T_Q} \sum_{b \in T_Q} \arg \max_{r \in T_D} \left(K_R^r \left(R_{HOG}, R_{HOG}^{'} \right) \right) \qquad (4)$$

$$R_r^D = \frac{1}{T_Q} \sum_{b \in T_Q} \arg \max_{r \in T_D} \left(K_D^r \left(D_{HOG}, D_{HOG}^{'} \right) \right) \qquad (5)$$

Where R_r^R, R_r^D are recognition matching scores for RGB, Depth features of the sign dataset respectively. R_r^R and R_r^D are in the range of $[0,1]$. The value zero and nearest to zero denotes the nonmatching and the value one indicates perfect matching. Then to extract perfect sign class of the unknown query sign, we apply average score fusion among the scores obtained from two proposed kernels as

$$R_{Sign} = \frac{R_r^R + R_r^D}{2} \qquad (6)$$

IV. Results and Discussion

The proposed algorithm was tested on our own sign language dataset BVCSL3D and two publicly available action datasets like NTU RGB-D and UTKinect. Finally, we give the recognition results with our proposed algorithm and performance compared with other state-of-the-art methods.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

A. Experimental Setup and Datasets

The experiments performed on (BVCSL3D) sign language dataset. It is captured by Microsoft Kinect sensor which gives RGB, depth and 3D coordinate location information. However, the BVCSL3D dataset has 100 signs captured by five signers, the sign labels are : *'Above', 'Absent', 'Accident', 'Alive', 'All', 'Assembly', 'Baby', 'Back', 'Bake', 'Ball', 'Batminton', 'Bed', 'Biscuit', 'Breakfast', 'Cabbage', 'Cake', 'Calcutta', 'Call', 'Calm', 'Cold', 'College', 'Corn', 'Corn', 'Cup', 'Curd', 'Dance', 'Dangerous', 'Dark', 'Eagle', 'Earn', 'Earth', 'East', 'Engineer', 'Face', 'Fail', 'Fall', 'Fan', 'Father', 'Flag', 'Food', 'Games', 'Gardener', 'Garlic', 'Gas stove', 'Good', 'Groundnut', 'Hair', 'Hammer', 'Hand', 'Hang', 'Happy', 'Health', 'Hello', 'Here', 'Himself', 'I', 'Icc cream', 'Idli', 'Improve', 'In', 'Jail', 'Jam', 'Kannada', 'Karate', 'Laddu', 'Lake', 'Lawyer', 'Leader', 'Lose', 'Magic', 'Magnet', 'Medicine', 'Mother', 'My', 'Nail', 'Name', 'North', 'Office', 'Page', 'Pain', 'Power', 'Puri', 'Quick', 'Race', 'Sad', 'Salt', 'Same', 'Shop', 'Some times', 'Sports', 'State', 'Strike', 'Tail', 'Tonic', 'Trophy', 'Under', 'Van', 'Village', 'Vollyball', 'Zoo'*. Figure.4. shows our database samples of RGB and Depth capture at Biomechanics and Vision Computing Research Centre at koneru lakshmaiah education foundation named as BVCSL3D.

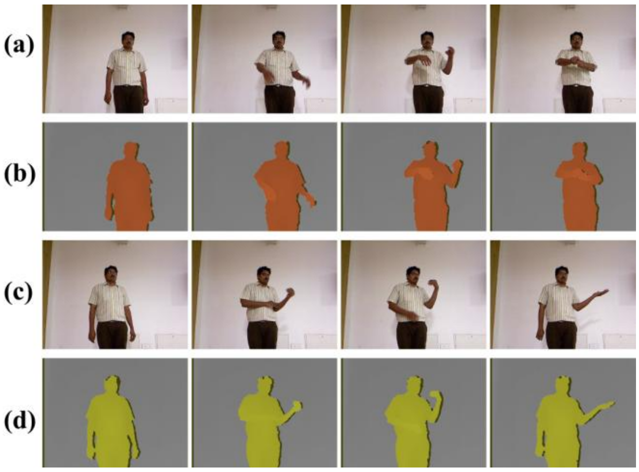

Figure.4. BVC3DSL Kinect sign dataset (a) sign 'Bite' in RGB and its (b) RGB-D (Depth) (c) sign 'Right' in RGB and its (d) Depth.

However, the NTU RGB-D is a very large-scale dataset for action recognition. It has more than 56 thousand labels of 60 different action classes, which are performed by 40 different subjects age between 10 to 35. The dataset consists of four different data samples such as RGB, Depth, 3D Skeleton and infrared videos. This dataset is captured by three Microsoft Kinect V2 sensors and the 3D coordinates of the human body consists of 25 joints. The dataset has two types of standard evaluation protocol, cross view and cross subject.

The UTKinect action 3D dataset, is captured using single stationary Kinect. The dataset contains 10 actions types: *'sit down', 'walk', 'pick up', 'stand up', 'throw', 'carry', 'pull', 'wave hands', 'push', 'clap hands'*. The actions performed 10 subjects (9 males and 1 female), every action performed each subject twice and each action length contain 5 to 120 frames. The dataset contains RGB images, Depth map and skeleton joint locations.

B. Performance of proposed method

We compare the performance of our proposed AK matching algorithm for evaluating the results with respect to precision, recall and percentage of recognition average on 100 signs is around 96.2, 97.1 and 97.8. To test the proposed algorithm against previous methods like weighted graph matching (WGM), adaptive graph matching (AGM), dynamic time wrapping (DTW), histogram, locally preserving positions bag of words (LPP-BOW) and support vector machine.

Table-1 shows the comparison between our proposed method and state-of-the-art techniques is performed on Indian sign language data. The data contain 100 sign labels, each sign performed 5 subjects and 5 different views. Here, we have 2500 signs, form that 500 signs are training, and the remaining 2000 signs are testing.

Table-1 Performance analysis of various methods on BVC3DSL dataset

Method	Precision	Recall	Recognition
WGM	93.2	92.7	93.1
AGM	92.9	92.1	93.4
DTW	78.1	75.7	75.3
Histogram	83.6	85.2	81.4
LPP-BOW	87.9	84.7	85.2
SVM	89.2	91.4	90.2
Proposed method	96.2	97.1	97.8

The results form table-1 shows weighted graph matching and adaptive graph matching came to close to our algorithm. Because of involvement of training phase, where WGM the output weights are updated based on cost function. In AGM, the recognition rate is good, but the computation time is very high.

The signs in a sign language are divided into simple, medium and complex signs. Simple signs use one hand, medium are two hand independent signs and complex signs are two hand torso dependent signs. We draw a confusion matrix in figure.5 showing the nine example signs and the corresponding recognition rates using the proposed method. Most of the methods used for comparison showed high resistance for complex signs when compared to the proposed method.

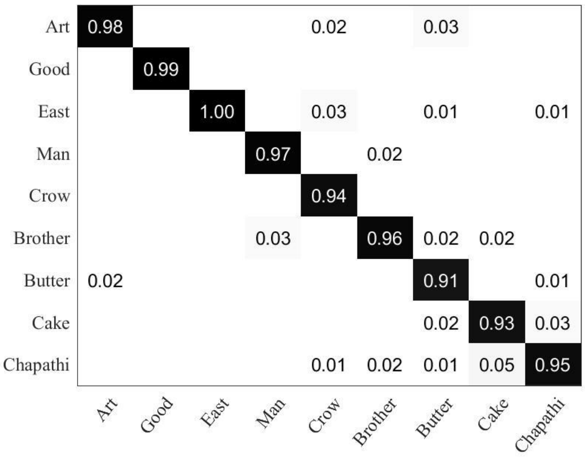

Figure.5. Confusion matrix for simple, medium and complex signs

However, to validate the effectiveness of the proposed method, we performed experiments on two benchmark depth datasets, NTU RGB-D and UTKinect. As shown in table-2, the effect of recognition rates obtained by state-of-the-art techniques against proposed algorithm on two datasets.

Table-2 comparison with proposed method and other state-of-the-art methods with public action datasets.

Method	NTU RGB-D	UTKinect
SGK [6]	91.5	96.2
JAVM-CNN [7]	89.43	--
ST-LSTM [8]	69.2	92.4
Lie Group [9]	52.76	93.6
Proposed Method	92.7	97.9

C. Sensor comparison

The comparison of BVC3DSL against different sensor sign language data and methods reported on that data from literature was presented in table-3. Here, comparing the recognition rate performance between Microsoft Kinect SLR data and other popular sensor data, such as 2D camera, glove-based sensors, EMG, leap motion. The state-of-the-art methods used either Microsoft Kinect or leap motion sensors for gesture recognition. In this work, we used depth data featured along with RGB video by using AK matching algorithm. The recognition rate is close to 98%.

Table-3 Sensor based performance comparison of SLR models

Sensor	Publication	Accuracy
2D Image and Stereovision	Anantha Rao., (2017) [3]	90.00%
	Kishore, P.V.V., (2016) [10]	92.50%
Glove Based Sensors	T.-H. S. Li., (2016) [1]	91.30%
	M. Mohandes., (2014) [2]	97.40%
EMG	Yun Li., (2012) [11]	96.50%
	Xu Zhang., (2011) [11]	92.50%
Leap Motion	I. Nigam., (2014) [4]	91%
	F. R. Khan., (2016) [12]	52.56%
Kinect	A. Memiş., (2013) [5]	91.52%
	Z. Liang., (2018) [6]	83.60%
	Proposed	97.8%

V. CONCLUSION

In this work, we recognize 3D Indian signs captured using Microsoft Kinect. The proposed method extracts histogram of oriented gradient features for depth-based sign language recognition using AKs. The AKs are designed for finding similarity between query sign with the dataset signs. The proposed framework validates on our own sign language dataset (BVC3DSL) and two publicly available action datasets like NTU RGB-D, UTKinect. The results show an improvement in recognition rates using the proposed framework due to adding depth features. The proposed model has shown outstanding performance when compared to other state-of-the-art action recognition methods.

REFERENCES

1. T.-H. S. Li, M.-C. Kao, and P.-H. Kuo, “Recognition System for Home-Service-Related Sign Language Using Entropy-Based K-Means Algorithm and ABC-Based HMM,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 46, no. 1, pp. 150–162, Jan. 2016.
2. M. Mohandes, M. Deriche, and J. Liu, “Image-Based and Sensor-Based Approaches to Arabic Sign Language Recognition,” IEEE Transactions on Human-Machine Systems, vol. 44, no. 4, pp. 551–557, Aug. 2014.
3. Rao, G. A., Kishore, P. V. V., Kumar, D. A., & Sastry, A. S. C. S. (2017). Neural network classifier for continuous sign language recognition with selfie video. Far East Journal of Electronics and Communications, 17(1), 49.
4. I. Nigam, M. Vatsa, and R. Singh, “Leap signature recognition using HOOF and HOT features,” 2014 IEEE International Conference on Image Processing (ICIP), Oct. 2014.
5. A. Memiş and S. Albayrak, “A Kinect based sign language recognition system using spatio-temporal features,” Sixth International Conference on Machine Vision (ICMV 2013), Dec. 2013.
6. Kishore, P. V. V., Kameswari, P. S., Niharika, K., Tanuja, M., Bindu, M., Kumar, D. A., ... & Kiran, M. T. (2018). Spatial Joint features for 3D human skeletal action recognition system using spatial graph kernels. International Journal of Engineering & Technology, 7(1.1), 489-493.
7. Kumar, E. K., Kishore, P. V. V., Kumar, M. T. K., Kumar, D. A., & Sastry, A. S. C. S. (2018). Three-Dimensional Sign Language Recognition With Angular Velocity Maps and Connived Feature ResNet. IEEE Signal Processing Letters, 25(12), 1860-1864.
8. Liu, J., Shahroudy, A., Xu, D., & Wang, G. (2016, October). Spatio-temporal LSTM with trust gates for 3D human action recognition. In European Conference on Computer Vision (pp. 816-833). Springer, Cham.
9. Vemulapalli, R., Arrate, F., & Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 588-595).
10. Kishore, P. V. V., Kumar, D. A., Goutham, E. N. D., & Manikanta, M. (2016, March). Continuous sign language recognition from tracking and shape features using fuzzy inference engine. In 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 2165-2170). IEEE.
11. Yun Li, Xiang Chen, Xu Zhang, Kongqiao Wang, and Z. J. Wang, “A Sign-Component-Based Framework for Chinese Sign Language Recognition Using Accelerometer and EMG Data,” IEEE Transactions on Biomedical Engineering, vol. 59, no. 10, pp. 2695–2704, Oct. 2012.
12. F. R. Khan, H. F. Ong, and N. Bahar, “A Sign Language to Text Converter Using Leap Motion,” International Journal on Advanced Science, Engineering and Information Technology, vol. 6, no. 6, p. 1089, Dec. 2016.