Dicle University Journal of Engineering (DUJE)

web: http://dergipark.gov.tr/dumf

Research Article

Recognition of static hand gesture with using ANN and SVM

Julius BAMWEND[1,*], Mehmet Siraç ÖZERDEM[2]

[1] Dicle Üniversitesi, Elektrik Elektronik Mühendisliği Bölümü, Diyarbakır, ORCID iD 0000-0002-6549-940X

[2] Dicle Üniversitesi, Elektrik Elektronik Mühendisliği Bölümü, Diyarbakır, ORCID iD 0000-0002-9368-8902

ARTICLE INFO

Article history:

Received 23 May 2019
Revised 16 June 2019
Accepted 17 June 2019
Available online 19 June 2019

Keywords:

Dynamic / Static Hand Gesture Recognition, Artificial Neural Network, Histogram of Oriented Gradient, Support Vector Machine

Doi:10.24012/dumf.569357

ABSTRACT

Hand gesture recognition is a relevant study topic for a reason that sometimes we may not be in position to communicate verbally. There is need to design Hand gesture recognition systems in order to help people adopt to nonverbal communication mainly sign language. However, there is no clue to understand the meaning of gesture through the computers directly. So this calls for definitions that generalize models in a computer. That is why the machine-learning approaches are implemented in recognition systems. There are generally two types of hand gestures recognition systems which researches have concentrated on. These include static and dynamic Hand gesture recognition systems. However, in building Hand gesture recognition systems, various machine learning approaches have been used. For implementing the proposed system, MS Kinect depth sensor was used as a hardware. The Kinect depth sensor is composed of an infrared camera. This is an advantage to the systems that are designed basing on the depth sensing because factors like color, clothing and background have less effect on the performance. So Kinect based depth sensor systems have a high accuracy and performance making them relevant and applicable in our daily lives. In this paper, we propose a static hand gesture recognition system in real time using two machine learning methods namely Support Vector Machine and Artificial Neural Networks. We use of the newly launched Microsoft Kinect sensor for image extraction. The sensor helps us to extract the hand images. We implement the system on a Matlab platform for reasons that Matlab is widely used by researchers in different fields and that can handle complex computations. In the training of the model, we collect a hundred depth-based Histogram of Oriented Gradient features per alphabet from the hand gesture images which we trained, tested and validated using Artificial Neural Networks (ANN) and Support Vector Machine (SVM). From this dataset, we can generate the generalized gesture model for each alphabet image. For the proposed system, the classification with ANN proves a higher performance then SVM.

Introduction

The use of computers has evolved rapidly in many fields namely leisure industrial, communication, and so on. We utilize those machines almost every time at work, home, school and almost in every field now. In any way, computers are part of us and we can't do away with them. Currently we know that using a computer requires interacting with some devices like mouse and keyboard. People started to use the mouse and keyboard in 1980s as a way to communicate with computers. As we speak today, there is a high progress in technology and sensing devices have been developed hence mouse and keyboards are becoming irrelevant. There have been a lot of studies related with how humans can interact with computers. Technological development has created a number of fields of study such as Gesture Recognition, Image processing and many others. Gesture recognition has become a very important field of study. It provides the basis for body recognition. Many researchers have carried out studies mainly in Face Recognition (FR) and Hand Gesture Recognition (HGR).

* Corresponding author
Julius BAMWENDA
✉ hbamwenda@gmail.com

Please cite this article in press as J. Bamwend, M.S. Ozerdem, "Recognition of Static Hand Gesture with using ANN and SVM", DUJE, vol. 10, no.2, pp. 561-568, June 2019.

In this paper, HGR was interested. Furthermore, there has always been a necessity of communicating with Sign Language (SL) in critical environments such as communicating with people that have speaking and hearing challenges. With the current technological advancements, different kinds of studies have been undertaken in order to counteract the sign language communication related problems. The above information motivated us to also play a part in solving the burning challenges in the HGR field. We developed a HGR system that can help in the interpretation of the different sign languages. In this paper, we propose a human hand detection method based on depth images captured by the MS Kinect (Figure 1). Human detection is achieved using the Infrared (IR) cameras found in the Kinect. We apply segmentation of the right hand from the fore, background and extract the hand details perfectly. We have also come up with algorithms that track in depth image. Our procedures are evaluated on a GIU interface using the Kinect and achieve excellent results.

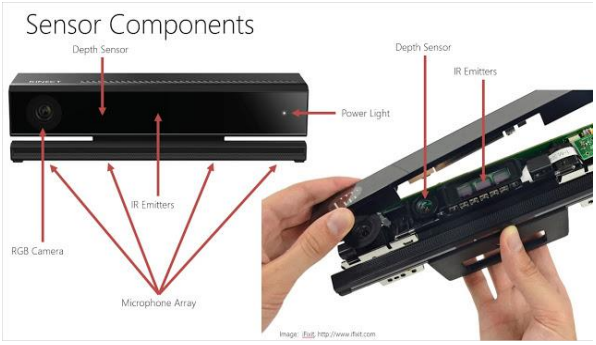

Figure 1. MS Kinect V2 and its components [1]

Related Studies

The aim of our study is to design a system capable of recognizing the hand gestures in real time. In our system, we use the MS Kinect in image extraction and machine learning techniques in the classification of the images. In the literature, we may find many papers based on MS Kinect and some of them are discussed in this section.

[2] presented a skeleton-based dynamic hand gestures recognition system. In order to analyze the figure movements, they extracted the angles of bones from the hand skeleton so they extracted the finger motion features. They extracted the global rotation and global translation of the hand to describe the global movements of the hand (extraction of the Global Motion Features). In the classification process, the finger motion features and global features are fed into a bidirectional RNN together with the skeleton sequence to aid the prediction of the class of input gestures. The publicly available skeleton-based DHG-14/28 dataset were utilized to compare the performance accuracies. Chong et al, (2016) presented yet another HGR system built basing on the 3D point cloud. They applied digital image processing in their research. Considering a 3D point from the depth camera, raw data from the hand was extracted. Thereafter, the data segmentation and preprocessing proceeded and three parameters were considered including the number of stretched fingers, the angles between fingers and the gesture region's area distribution feature. Decision tree method was applied in the classification. The results of experiment demonstrated that the proposed method was quite good in gesture recognition because it yielded an average accuracy of 94.7%. [3] describe a depth image based real–time skeleton fitting algorithm for the hand. They applied an object recognition method by parts approach and the used hand modeler in an American Sign Language. They created a 3D hand model with 21 different parts. Furthermore, Random decision forests (RDF) were trained on depth images hence creating a hand model and per pixel classification performed. The classification results are fed into a local mode finding algorithm to estimate the joint locations for the hand skeleton. Their system processes depth images from Kinect in real–time at 30 fps. They finally used support vector machine (SVM) based recognition module for the ten digits of ASL. The recognition rate attained is 99.9% on live depth images in real–time which is good enough for a system of this nature.[4] designed a number gesture recognition system based on the recognized hand parts in depth images taken by the Kinect. They proposed an approach that consisted of two main stages, hand parts recognition by random forests (RFs) and rule-based hand number gesture recognition. As already stated before, a database of hand depth and their corresponding hand parts-labeled maps was created, and then training followed using

random forest method. Using the trained dataset, labeling of hand parts in depth images was possible. Basing on the information of labeled hand parts, hand number gestures were recognized according to derived rules of features. The system was evaluated with synthetic and real data and an average recognition rate of 97.80 % over the ten hand number gestures from five subjects was attained. [5] carried out a magnificent study of which they employed multi sensors including MS Kinect to build a driver's hand gesture recognition system. In this paper, we see yet another field in which the Kinect can be applied. Basing on this study, we may predict that in future lots of accidents on the roads due to careless driving may be solved. We may also appreciate the power of ANN as a classification method. [6] presented an approach of feature extraction and classification for recognizing continuous dynamic gestures corresponding to Vietnamese Sign Language (VSL). The captured the Input data by the depth sensor of a MS Kinect, considering the advantage that this device is almost not affected by the light of environment. A dataset of 3000 samples corresponding to 30 dynamic gestures in VSL was created by 5 volunteers. They represented the gestures with a sequence of depth images just like other researchers in this field. Feature extraction is performed by dividing the images into 3D grid of same-size blocks in which each one is then converted into a scalar value They applied Support Vector Machine (SVM) in the classification procedure and the Hidden Markov Model (HMM) technique in order to provide a comparison on recognition accuracy. The experiments yielded an average accuracy 95%. [7] deigned a user interface that could help researchers working with Kinect to acquire, edit and even store images from the Kinect. They based their research on the images captured by the Kinect device. They created a dataset of eight gestures of which each gesture had eighty samples. They applied Dynamic Time Warping and Hidden Markov Model in the classification of the gestures. Looking at their accuracy recognition rate (99%), we can conclude that the above classification methods may be appropriate in the classification of the images captured by the Kinect sensor. [8] worked on a Kinect based HGR system motivated by the need to effectively communicate to the people with hearing problems. His system recognizes the hand gesture made by the user, then compares it with the predefined gesture set and in return it gives a matching. He came up with a simple procedure that had basically three main stages namely hand detection, figure identification and gesture recognition. Appling K-means clustering algorithms, he classified the images. He achieved a recognition rate of approximately 99.5%. Hand modeling being one of the complex tasks to do, many researchers have proposed different methods in order to overcome this challenge. In the same way, [9] proposed a hand posture recognition procedure based on the color, depth and skeleton parameters from Kinect images. In the method, hands are first distinguished from background by color data, which can separate non-skin-color background regions from the original image, then refined by depth data, which is used to remove the skin alike noise areas from the remain image, using skeletal data to decide the threshold of the depth of hand, and finally use morphological methods to segment hand and arm. This procedure can provide a solution to the hand modeling challenge at hand. [10] studied on hand gesture recognition where they SVM and RVMs to classify hand images from a Kinect sensor. They only utilized the skeleton data from the images captures by the Kinect sensor. In the feature extraction stage, point velocity, joint angles and joint angle velocities from the skeleton frame were extracted. The extracted features were then classified using SVM and RVMs and their accuracy gesture recognition rates were compared. [11] described method that could recognize alphabet signs using Kinect's depth images. A segmented hand is obtained by using a depth contrast feature based per-pixel classification algorithm. Then a mode-seeking method is developed and implemented to localize hand joint positions under kinematic constraints. Lastly, they applied Random Forest (RF) classify the extracted features. To validate the performance of this method, they used publically available datasets. The results showed an accuracy of 90% considering 24 static gestures from ASL. [12] presented again a system that could detect the hand gestures statically. As an

input device, they used a webcam in the creation of a dataset. The webcam also has a capacity to sense the depth of an image. In the preprocess stage, they created a boundary box that could separate and isolate the hand form the fore and back grounds. Image resizing in order to remove the unnecessary pixies was performed. The resized image could be detected using the canary's edge detection method. HOG features were extracted from the resized images and stored in the database for classification. They made use of the famous SVM for classification of which the procedure was divided into two namely training and testing. They created a user interface in Matlab which facilitated the implementation process. [13] designed a 3D hand gesture recognition system. They employed the kinect and specifically used the details from the depth image. In their methodology, they first of all used the kinect device to acquire the depth image. They then segmented the hand image which they later converted to a point cloud. In order to extract feature from the hand shape, they computed the 3D moment invariants on the point cloud and thereafter applied SVM for classification. They mainly concentrated on three gestures namely paper, rock and scissors hence realizing a high recognition accuracy rate of 97.7%. In the literature it can be seen in the above list, there are two types (static/dynamic) of hand gestures are evaluated. In this paper, we propose a static HGR system in real time using machine learning methods.

Material and Method

Proposed System

Our system consists of five main steps as illustrated in the Figure 2.

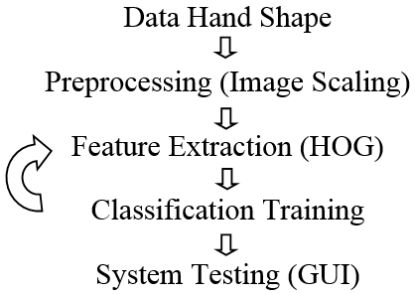

Figure 2. The overall procedure for gesture recognition

With Kinect 2.0 for Windows, we can capture raw data streams and skeletal information in our method. The device has features such as RGB camera and a depth sensor, which provides full-body 3D motion capture capability. In our study, we make use of the skeletal data stream as well as depth and color data streams.

Acquisition Dataset

Matlab Image Acquisition Toolbox provides functions and blocks that enables us to connect industrial and scientific cameras like Kinect to Matlab and Simulink. It includes a Matlab application that let researchers interactively detect and configure hardware properties. The toolbox enables acquisition modes such as processing in-the-loop, hardware triggering, background acquisition, and synchronizing acquisition across multiple devices and many other properties. By attaching the kinect device, we handled the color and depth frame in real time. Image Acquisition Toolbox gathers 30 frames per sec from the Kinect. Each frame contains the depthMetaData. From the depthMetaData, we extract the skeleton data which is one of the most useful parameters in getting the hand position. The skeleton data helps us in getting the joint positions of right hand. We set the Region of Interest (ROI) based on the center of mass. We then apply the background subtraction to reduce the effects that may result from reflection and other factors. At the end of the above processes, a handDepthImage is generated and stored further processing. The quality of the Kinect depth sensing is still inherently noisy and greatly affected by natural conditions. Depth measurements often fluctuate and depth maps contain numerous holes where no estimations of range are obtained. In the depth images taken by the Kinect, all the points that are not measured by the sensor are offset to 0 in the output array. For reasons of estimating the true depth value, we make the assumption that the space is continuous, and the missing point is more likely to have a similar depth value to its neighbors. Figure 3 represents our training dataset. It is sorted by left to right, top to bottom. It contains alphabets A to Y except J and Z because they are dynamic hand gestures. Our goal is to recognize the static postures so for that reason J and Z were excluded.

Our training data is based on the American Sign Language. However, in order to improve on the performance, some gestures that are closely similar to each other were replaced by gestures that we found easy to classify. The replaced gestures include A, M, N, S, T and Y respectively. All experiments were executed on a PC with CPU (Intel Core i5-4200M) 2.50GHz, 8GB RAM, and 500GB hard disk.

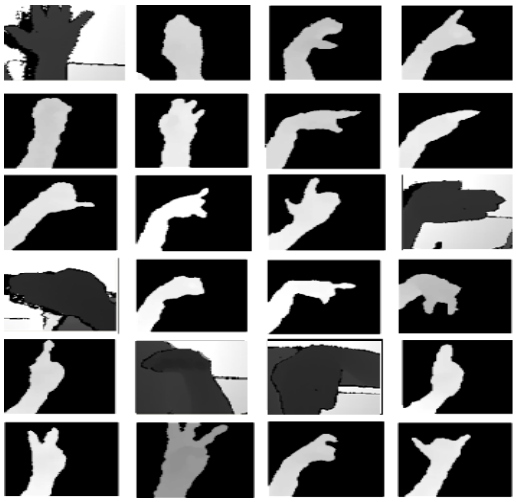

Figure 3. Training data, from A to Y except J, order by left to right, up to down

The machine was operated by Windows 8. In the software part, all the codes were implemented in MATLAB R2016b. For reasons that MS Kinect V2 works with the later versions of Matlab (from R2014 and above) in order to use the Kinect, we installed Kinect for windows SDK V2 and Kinect runtime. Those are very essential software for using the MS Kinect. On the other side, we installed the following hardware support packages namely Computer Vision System Toolbox, Image Acquisition Toolbox, Image Processing Toolbox, Neural Network Toolbox, Signal Processing Toolbox, Statistics and Machine Learning Toolbox, Bioinformatics Toolbox and Matlab Wrapper for SVMlight. With all this setup we carried out all our experiments.

For reasons of evaluating the performance of our approach, we created our own database using Kinect for windows device. Our dataset consists of 24 different hand gestures of which 18 of them belong to the American Sign Language. The other 6 gestures were created by us in order to improve the performance of the classification algorithms. J and Z letters weren't included into dataset since they are dynamic letters. So, only static letters were evaluated in this study. In the dataset creation, we used gestures from three different people in order to cater for some divergences that may arise due to differences in hand sizes. A total of 2400 depth images were created implying that each alphabet has 100 images (patterns). In addition, we train and test our system on both real and synthetic data that we downloaded from the internet. This helped us in providing a comparison between the dataset results.

Hand Feature Extraction

In this study, Histogram of Oriented Gradient (HOG) was employed. The HOG returns the count of occurrences of oriented gradients. So, the image oriented gradients can be represented by HOG. All hand gestures have different orientations. We then assume that each image would have unique histogram because an image has identical oriented gradients. An example of histogram procedure is given in Figure 4.

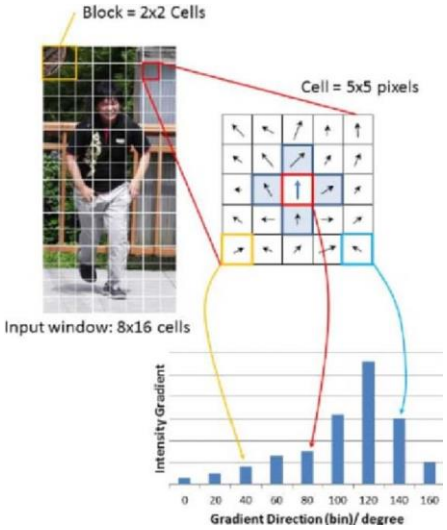

Figure 4. Histogram based feature extraction of image with HOG. M. [14]

The hand image features are extracted and converted by HOG. HOG is an edge orientation histogram based on the orientation of the gradient in localized region called cells. A histogram is a graphical representation of the distribution of numerical data. Therefore, it is easy to express the rough shape of the object and it's robust to variations in geometry and illumination changes.

Gesture Classification with SVM

In this procedure, we make use of the Support Vector Machine (SVM) which is a supervised learning model with associated algorithms that analyze data and recognize patterns. SVM was developed by Vapnik and used to supervised learning, Burges, C. J. C. (1998). Basically, this machine is a classifier of two sets which can be separable. It uses the support vector and kernels for learning. The kernel machine gives a framework which is flexible to the different domain by selecting the appropriate kernel functions. Unlike other machines, SVM makes a hyper plane or set of hyper planes in a high-dimensional space, which can be used for classification, regression, or other tasks. A good separation would be achieved by the largest distance to the nearest training data point of any class hyper plain, since the classification contained large margin can get the lower error and higher generalization.

The SVM (primal) optimization problem is as follows:

$$\max_{\mathbf{w}} \frac{2}{||\mathbf{w}||} \text{ subject to } \mathbf{w}^\top \mathbf{x}_i + b \begin{matrix} \geq 1 & \text{if } y_i = +1 \\ \leq -1 & \text{if } y_i = -1 \end{matrix} \text{ for } i = 1 \dots N$$

Or equivalently

$$\min_{\mathbf{w}} ||\mathbf{w}||^2 \text{ subject to } y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 \text{ for } i = 1 \dots N \quad (1)$$

This problem is quite complex to solve directly, we can compute this problem by formulating unconstrained optimization using Lagrange multipliers.

$$L(w, b, \alpha) = \tfrac{1}{2}||w||^2 - \sum_{t=1}^{N} \alpha_t [y_t (w^T x_t + b) - 1]$$
$$\text{where } \alpha_t \geq 0 \quad (2)$$

In the above formula, we apply Karush-Kuhn-Tucker

$$\text{maximize } \sum_{t=1}^{N} \alpha_t - \tfrac{1}{2} \sum_{s=1}^{N} \sum_{t=1}^{N} \alpha_s \alpha_t y_s y_t x_s^T x_t$$
$$\text{subject to } \sum_{t=0}^{N} \alpha_t y_t = 0$$
$$\text{where } \alpha_t \geq 0, \forall t \in [1, N] \quad (3)$$

This dual problem has simpler setting of involved constraints. Although the original problem would imply a finite dimensional space, it only occurs when the discriminated sets are not linearly separable in that space. Because of this, it was suggested that the original finite-dimensional space should be converted into a higher-dimensional space, for making the classification easily. To keep this process reasonably good, the SVM is designed to ensure that data may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function. Since the number of our dataset gestures is 24, we then need a multi-classifier and proper utility. In this case, we use LIBSVM which supports multi-class classification for classifying the images.

Classification with Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) refers to the simulations performed on the computer to complete a number of machine learning tasks such as pattern recognition, clustering, classification. The ANNs are biologically inspired. ANNs are similar to the human brain in ways that they acquire knowledge through learning and store that knowledge within inter-neuron connection strengths known as synaptic weights. Those two characteristics make neural networks unique in nature. We applied Multilayer perception in our study. With Multilayer perception, the network is composed of more than one hidden layer of neurons as compared to single layer perception architecture. Different types of structure were tried and we can deduce that the highest classification rate was observed in 20736x23x9x24 ANN structure.

Results

Due to the fact that we needed multi-classification, we employed LIBSVM as stated under methods and procedure. We used a total of 2400 input hand gesture images for SVM training. This implies that the training data is composed of 100 images per alphabet. The test was executed by real-time. The Kinect gets 30

frames per second so that we count the correct recognition frame.

In the test, we created a graphical user interface (GUI) which is composed of four windows. The two windows on the top are taken through Kinect. The left one is an original image returned by the color stream of the Kinect and right one is a depth frame returned by the depth stream. From this depth data, the human body and hands are distinguished as shown in the Figure 5. The segmented hand image is shown in left-bottom window. The last window shows the alphabet which is represented by a hand gesture. When the Kinect detects the human, the framework generates the 25 skeleton points from depth data. Using right hand skeleton point (12), only right hand image is segmented, and compared it with trained model. If the decision is right, the right bottom window changes the appropriate alphabet. We present a sample cases from the 24 alphabets. The experiment was repeated 10 times and the average of performance was obtained as 93.4.

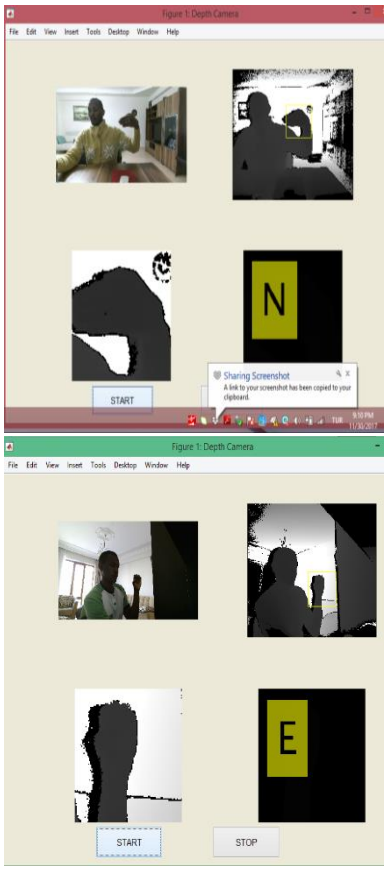

Figure 5. The GUI showing letters E and N

Considering ANN classification, we used scaled conjugate gradient backward propagation as the training function. The performance was evaluated using mean squared error. The dataset was randomly divided up into training, testing and validation data. During the training stage, the performance graphically was obtained shown in Figure 6.

The experiment was repeated 10 times and average of performance was obtained as 98.2.

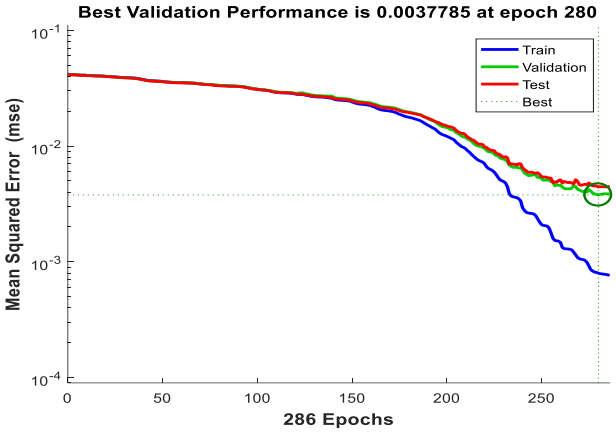

Figure 6. Represents the best validation performance at epoch 280

Conclusions

The aim of our study is to design a system capable of recognizing the hand gestures in real time. In our system, we use the MS Kinect in image extraction and machine learning techniques in the classification of the images. The system is motivated by the importance of real-time communication under specific situations such as communication under water and communicating to people with hearing problems. We reviewed quite a number of systems in our study providing information about sign language recognition systems and hand gesture recognition systems.

The proposed system in this paper, the right hand is distinguished from the background by the depth information and that constitutes the preprocessing procedure. The hand detection is in a range is between 0.5m to 0.8m from the Kinect. HOG features were used to extract the hand positions from the images. A set of hand positions was then passed to ANN and SVM for classifying the images. The dataset was organized as consisting

static letters. The average of recognition accuracy was 93.4 and 98.2 for SVM and ANN, respectively. For the proposed system, the better performance was observed with ANN method.

References

[1] T. Cook, R. Cargill, *Say Hello to My Little Friend : JavaScript!* 2014.

[2] X. Chen, H. Guo, G. Wang, and L. Zhang, "Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition," in *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 2881–2885.

[3] L. Keskin, C., Kıraç, F., Kara, Y. E., Akarun, "Real time hand pose estimation using depth sensors," *Consum. depth cameras Comput. Vis.*, pp. 119–137, 2013.

[4] T. S. Dinh, D. L., Lee, S., Kim, "Hand number gesture recognition using recognized hand parts in depth images," *Multimed. Tools Appl.*, vol. 75, no. 2, pp. 1333–1348, 2016.

[5] M. P. . G. S. . K. K. . P. K., "Multi-sensor system for driver's hand-gesture recognition," in *11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, 2015, pp. 1–8.

[6] J. Vo, D. H., Huynh, H. H., Doan, P. M., Meunier, "Dynamic Gesture Classification for Vietnamese Sign Language Recognition," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, pp. 412–420, 2017.

[7] M. Ibanez, R., Soria, A., Teyseyre, A., Campo, "Easy gesture recognition for Kinect," *Adv. Eng. Softw.*, vol. 76, pp. 171–180, 2014.

[8] L. Y., "Hand gesture recognition using Kinect," in *IEEE International Conference on Computer Science and Automation Engineering*, 2012, pp. 196–199.

[9] X. Xie, J., Shen, "Hand posture recognition using kinect," in *International Conference on Virtual Reality and Visualization (ICVRV)*, 2015, pp. 89–92.

[10] H. S. Nguyen, D. D., Le, "Kinect gesture recognition: Svm vs. rvm.," in *Seventh International Conference on Knowledge and Systems Engineering (KSE)*, 2015, pp. 395–400.

[11] Z. Dong, C., Leu, M. C., Yin, "American sign language alphabet recognition using microsoft kinect," in *IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 44–52.

[12] R. K. Nagashree, R. N., Michahial, S., Aishwarya, G. N., Azeez, B. H., Jayalakshmi, M. R., Rani, "Hand gesture recognition using support vector machine," *Int. J. Eng. Sci.*, vol. 4, no. 6, pp. 42–46, 2005.

[13] L. Song, L., Hu, R., Xiao, Y., Gong, "Real-Time 3D Hand Gesture Recognition from Depth Image," in *nd International Conference On Systems Engineering and Modeling (ICSEM-13)*, 2013.

[14] M. Hatto, "Acceleration of Pedestrian Detection System using Hardware-Software Co-design.," Lund University MSc Thesis, 2015.