Hindawi
Advances in Human-Computer Interaction
Volume 2018, Article ID 1069823, 13 pages
https://doi.org/10.1155/2018/1069823

Research Article

Recognition of Symbolic Gestures Using Depth Information

Hasan Mahmud,[1] **Md. Kamrul Hasan**,[1] **Abdullah-Al-Tariq**,[1] **Md. Hasanul Kabir**,[2] **and M. A. Mottalib**[3]

[1]*Systems and Software Lab (SSL), Department of Computer Science and Engineering, Islamic University of Technology (IUT), Dhaka, Bangladesh*
[2]*Department of Computer Science and Engineering, Islamic University of Technology (IUT), Dhaka, Bangladesh*
[3]*Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh*

Correspondence should be addressed to Hasan Mahmud; hasan@iut-dhaka.edu

Received 10 July 2018; Revised 24 September 2018; Accepted 28 October 2018; Published 19 November 2018

Academic Editor: Marco Porta

Copyright © 2018 Hasan Mahmud et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Symbolic gestures are the hand postures with some conventionalized meanings. They are static gestures that one can perform in a very complex environment containing variations in rotation and scale without using voice. The gestures may be produced in different illumination conditions or occluding background scenarios. Any hand gesture recognition system should find enough discriminative features, such as hand-finger contextual information. However, in existing approaches, depth information of hand fingers that represents finger shapes is utilized in limited capacity to extract discriminative features of fingers. Nevertheless, if we consider finger bending information (i.e., a finger that overlaps palm), extracted from depth map, and use them as local features, static gestures varying ever so slightly can become distinguishable. Our work here corroborated this idea and we have generated depth silhouettes with variation in contrast to achieve more discriminative keypoints. This approach, in turn, improved the recognition accuracy up to 96.84%. We have applied Scale-Invariant Feature Transform (SIFT) algorithm which takes the generated depth silhouettes as input and produces robust feature descriptors as output. These features (after converting into unified dimensional feature vectors) are fed into a multiclass Support Vector Machine (SVM) classifier to measure the accuracy. We have tested our results with a standard dataset containing 10 symbolic gesture representing 10 numeric symbols (0-9). After that we have verified and compared our results among depth images, binary images, and images consisting of the hand-finger edge information generated from the same dataset. Our results show higher accuracy while applying SIFT features on depth images. Recognizing numeric symbols accurately performed through hand gestures has a huge impact on different Human-Computer Interaction (HCI) applications including augmented reality, virtual reality, and other fields.

1. Introduction

Gesture-based interaction has been introduced in many HCI applications which allow users to interact intuitively through computer interfaces in a natural way. Rather than using traditional unimodal inputs, blending alternative style of interactions, such as hand gestures along with mouse and keyboard, introduces more degree of freedom (DoF) to the computer users. Nowadays, hand gesture-based interaction is a prominent area of research which has a huge impact in the design and development of many HCI applications like controlling robots through hand gestures, manipulating virtual objects in an augmented reality environment, playing virtual reality games through different hand movements, communicating through sign languages, etc. We need these types of interaction to achieve interaction design goals like effectiveness, efficiency, affordance, and feedback.

Hand gesture can be defined as the movement of hands and fingers in a particular orientation to convey some meaningful information [1] like pointing to some object through index fingers, expressing victory sign or OK sign, waving hands, grasping an object, etc. Symbolic hand gestures represent some specific symbols like 'OK' sign or gesture that represents numeric symbol '1' (raising the index finger and bending all other fingers). In most of the cases, these gestural movements conveys single meaning in each culture

having very specific and prescribed interpretations. More importantly, symbolic gestures are alternative to verbal discourse structure, different from everyday body movement which is consciously perceived. These gestures are observed in the spatial domain and are called static hand gestures characterized by the position of fingers (finger joint angle, orientation, and finger bending information). Unlike static hand gestures, dynamic gestures are considered in the temporal domain, presenting gesture as a sequence of hand shapes which includes starting through ending hand pose (e.g., hand waving, boxing).

There are different approaches to capture and recognize these gestures. Computer vision-based approach imposes restrictions on the gesturing environment, such as special lighting conditions, simple and uncluttered background, and occlusions (the gesturing hand is occluded by other parts of the body) [1]. Due to these restrictions segmentation of hand may cause the reduction in hand gesture recognition accuracy. Hand poses, generated in the process of gesticulation, can also be detected by means of wearable sensor like data-gloves. The data-gloves are embedded with the accelerometer, gyroscope, bend sensor, proximity sensor, and other forms of inertial sensors [2]. These sensors collect hand-finger motion information as multiparametric values. However, the sensor-based gesture recognition approaches have limitations in terms of naturalness, cost, user comfort, portability, and data preprocessing.

The recent advancements in stereo vision camera that utilizes depth perception from smaller to larger distances have opened a huge scope for the researchers to work with depth information [3]. Traditional web cameras do not provide the depth values (the distance of the gesturing hand from the camera). Depth information can help eliminating occlusion problems easily and can quicken the segmentation process with less error. In an occluded background, using depth information it is possible to extract the gesturing hand movement information including other important features (e.g., finger bending information) which can be effectively utilized in feature representations. Moreover, static gesture can be performed by the users with varying hand size, changes in hand position (orientation, rotation), and different illumination conditions. Scale-Invariant Feature Transform (SIFT) [4] is an algorithm that works better for these types of variation. The algorithm generates key points from images and provides 128-dimensional feature vectors.

In this research work, we try to recognize symbolic hand gestures representing 10 numeric symbols from 0 to 9. These are very close gestures, differing only in slight variations (e.g., the difference between numeric symbol 2 and numeric symbol 3 is due to the presence/absence of one finger only) of finger positions. With the help of depth data stream, after a quick and robust segmentation process, we have calculated depth threshold based on which the contrast varying depth images are generated according to the depth map of the individual gesture. This process was applied to 100 image instances per gesture. In each image for the same gesture, we got the different number of SIFT keypoints. By combining the keypoints, we have generated bag-of-feature (BoF) vector with the help of the k-means clustering technique to generate uniform dimensional feature vectors and classified using a multiclass SVM.

The main contributions of this paper are as follows:

(i) Generating contrast varying grey-scale depth images according to the depth map to utilize local shape information of hand fingers which has contributed to the improvement of recognition accuracy.

(ii) Applying SIFT over depth images to achieve image invariant properties (translation, scaling, rotation, illumination, and local geometric distortion).

The remainder of the paper is described as follows. Section 2 elaborates the related research. Section 3 describes the proposed approach. Section 4 presents the experimental results. The last Section 5 describes the conclusion and the future scope of the work.

2. Related Works

Human hand is a highly articulated model, prominent in making deft poses. To recognize those hand poses many research works have utilized RGB cameras and applied either template-based approaches or model-based approaches on RGB images. Conventional RGB image-based gesture recognition techniques need to consider many research challenges, such as light sensitivity, cluttered background, and occlusions. However, the recent emergence of depth sensors has given an opportunity for the researchers to utilize the depth information in order to overcome those challenges. The depth data stream provided by the depth sensors (e.g., Microsoft Kinect, Intel Real Sense, Asus Xtion Pro) corresponding to the hand gesture images has given new dimensions to conduct research in hand segmentation process, finger identification techniques, finger joint detection, and finger tracking. Depth value indicates the distance of the gesturing hand from the RGB-Depth (RGB-D) camera in millimeters appropriate to make the segmentation process faster. Among the depth sensors, we have used Microsoft Kinect depth sensor that captures depth image in 640×480 resolution in a frame rate of 30 fps and 11-bit depth under the environment consisting of any ambient light. Depth information helps to extract additional features which can significantly improve the recognition results. Many researchers have developed depth sensor-based applications like interactive displays through Kinect [5], a system for therapeutic interventions [6], robot navigation through gestures [7–9], Kinect-based American Sign Language (ASL) recognition [10], etc. Other different applications of Kinect depth sensor includes categorizations of indoor environments by mobile robots equipped with Kinect [11], measuring canopy structure for vegetation [12], just to name a few.

From the depth sensors, the most common features used in hand posture recognitions [13] are skeleton joint positions, hand geometry, hand-finger shape, area, distance features, depth pixel values, etc. Generally, these features can be categorized as local features or global features. The major challenges of these feature descriptors are variations of gesturing hands while articulating an emblem or symbolic gesture. A gesture may slightly differ in terms of hand shape

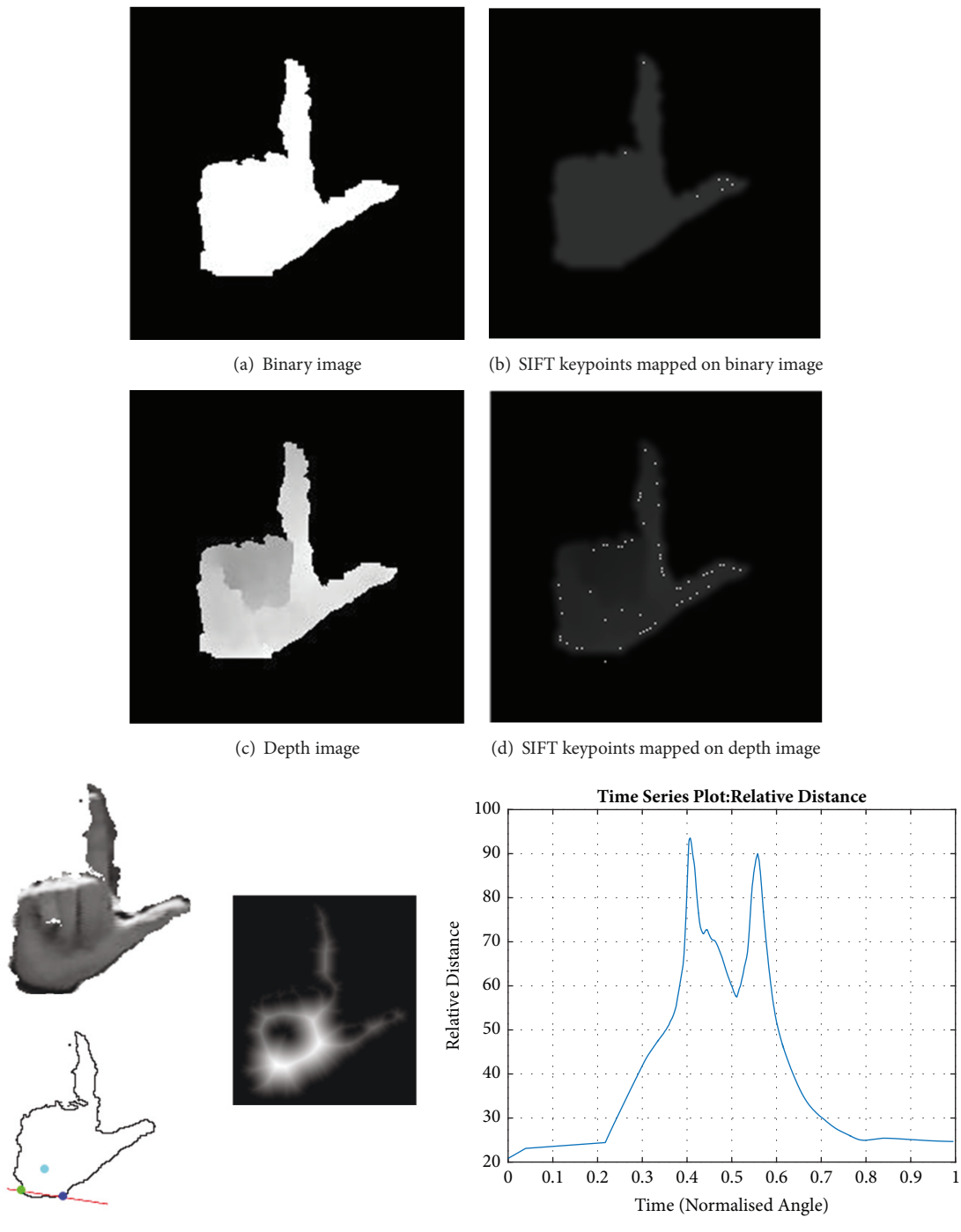
and size, variations in translation, or rotation of the fingers for the same gesture. A robust hand gesture recognition system should be invariant to the scale, speed, and the orientation of the gesture performed.

The approaches that are followed by static gesture recognition system from binary images in [14] and time-series curves in [15] do not facilitate the possibility of extracting local finger context information. The authors in [14] have captured RGB images from webcam and converted them to binary images and applied SIFT algorithm to determine the recognition accuracy. In binary images, the finger context information, shape, orientation, bending fingers, and occlusion, cannot be preserved, a limitation that can be overcome by utilizing depth map information of the gesturing hand. SIFT keypoints are important feature points which are well distributed and contain information about not only thumb and baby fingers but also finger bending information of index, middle, and ring fingers. Figure 1 shows the differences of SIFT keypoints in gesture 8 mapped into the binary image (7 keypoints) (Figure 1(b)) and into depth image (56 keypoints) (Figure 1(d)). This information is not present in the case of binary image or time-series curve. SIFT works on local oriented features rather than topological shapes of opening fingers which are considered as the global features. In [15], global features are used to generate time-series curves (Figure 1(f)) after the segmentation process as shown in Figure 1(e) from the hand shape represented in binary image. The edit-distances are calculated to apply distance-based matching algorithm, such as Finger-Earth Mover's Distance (FEMD). Edit-distance-based matching algorithms are not completely rotation, orientation invariant because they are measured by comparing time-series trajectories based on the proximity distance and not based on the local shape information. Moreover, the temporal information is better for dynamic gesture recognition rather than static gesture recognition [16].

Local features measure the characteristics of a particularly important region of the object, superior in discriminating fine details. In [17], shape descriptor-based algorithm and weak learning-based strong classifier were applied to recognize three symbolic gestures (palm, fist, six). Their goal was to get orientation invariant property of those gestures. They have used SIFT features as local features in weak classifier for hand detection and trained each classifier independently. The accuracy, in this case, depends on the large set of training images which they have not considered. They have used a varying number of training images for individual gestures. They have not considered the fact that SIFT features extracted from the different gesturing image can form a natural group of clusters having feature vectors of the unified dimensions appropriate to feed into a classifier that can recognize more than two classes. We have achieved this by clustering feature descriptors and generating BoF features. In [18, 19], the researchers have considered Haar-like features, applying learning-based techniques to recognize hand gestures. They required a huge number of images for training and testing with high computational power and they have not considered the scale-invariant property for object detection.

Global features measure the characteristics of the whole image and face difficulties in capturing fine details. An example would be the contour representation of a hand gesture image (e.g., the hand contour image of Figure 1(e)) which gives hand-finger shape information from the whole image. The limitations of contour-based recognition methods are that they are not robust on local distortion, occlusion, and clutter [20]. To extract the complete hand posture information while a finger and a palm are overlapped, such as bending fingers, as shown in Figure 1(d), the consideration of hand contour as the global feature representations is not enough. The similar problems are also mentioned in the recognition approaches like skeleton-based recognition methods [21], shape contexts based methods [22], and inner-distance methods [23]. A solution to these problems was proposed using a novel distance-based measurement technique called Finger-Earth Mover's Distance (FEMD) [15]. They represented the shape of hand fingers as a global feature (the finger cluster) by analyzing time-series curve. In the curve, the Euclidean distance between each contour point and the center point is considered in one dimension and the angle of these contour points made with the initial point relative to the center point is considered as another dimension. Figure 1(f) shows the time-series curve of the topological hand shape considered as finger parts and matches those fingers only, not the whole hand shape. Features only from opening finger parts may not give good recognition results. Rather features including bending finger parts as local features will play a significant role to improve the recognition accuracy. We have considered those features in our proposed approach. Moreover, for gesture recognition, they [15] have applied template matching with minimum dissimilarity distance which may not give improved recognition accuracy on both changes in orientation and rotation of a particular pose. We propose to overcome this problem using local features found as SIFT keypoints. Edit-distance-based time-series matching approaches are more applicable for dynamic gesture recognition due to their spatiotemporal features, rather than static symbolic gesture recognition. Template-based approaches are good to recognize the shape as a whole but lack in terms of invariance. SIFT algorithm is known to be robust for its distinctiveness and invariance to rotation, scale, and translation in object recognition. Depth image acquired using Kinect depth sensor suffers from low grey level contrast that can cause an unstable set of keypoints. Recently in [24], the researchers used Kinect-based depth map information to discard the SIFT keypoints that are located at the boundaries of an object. They applied Canny's edge detection algorithm [25] on depth images and generated an object model to store depth values and distance to the nearest depth edge for the remaining SIFT keypoints. They have used Euclidean distance-based nearest neighbor algorithm to rank the keypoints matches and performed RANSAC-based homograph estimation for object pose estimation. Their aim was to identify predefined objects in the surrounding environment for the visually impaired. To extract a stable set of SIFT keypoints different techniques were proposed by the researchers. Preprocessing on the medical image (retina image) was done to reduce the number of SIFT keypoints in [26].

In [27], the researchers have extracted the SIFT keypoints from both the color and the depth image and tried to find



(a) Binary image (b) SIFT keypoints mapped on binary image

(c) Depth image (d) SIFT keypoints mapped on depth image

(e) Hand segmentation (f) Its time-series curve

Figure 1: Differences in the number of SIFT keypoints in both (a) binary image and (c) depth image and the use of finger bending information.

out the correspondence of SIFT keypoints between those two images. They have combined SIFT descriptor with Harris corner detector to compute SIFT features at predefined spatial scales. They enhanced the depth image contrast by applying histogram equalization without utilizing the depth values explicitly of the gesturing hand to generate contrast varying depth images. However, we have considered the depth map information to determine the contrast level and generate depth silhouettes accordingly.

SIFT algorithm along with its different variants like PCA-SIFT [28], SURF [29], and GLOH [30] has been applied in various applications such as image stitching, object recognition, and image retrieval. SIFT and SURF algorithm were also applied in simultaneous localization and mapping (SLAM)

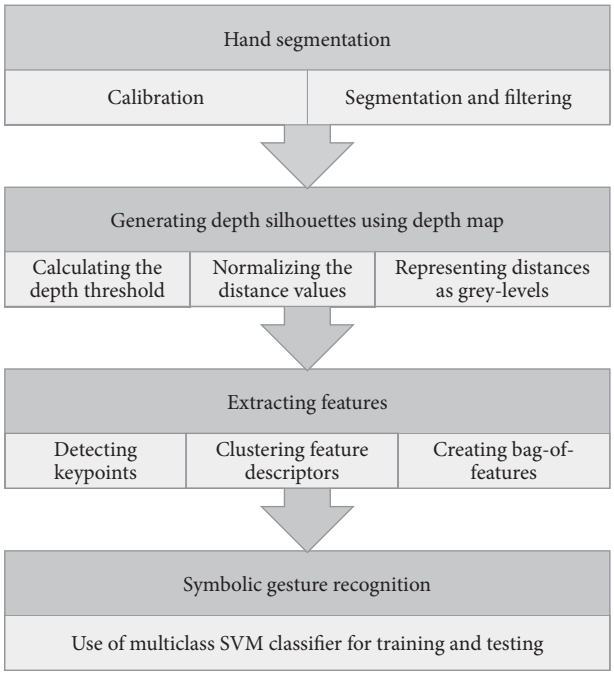


Figure 2: The architecture to recognize symbolic gestures.

with RGB-D Kinect sensor on robots [31]. SURF is the fast approximation of SIFT that uses box filter instead of Gaussian filter. However, SURF is not good at different illumination conditions [29]. To improve the time complexity of SIFT several alternatives were proposed, such as Binary Robust Independent Elementary Features (BRIEF) [32] and Oriented FAST and Rotated BRIEF (ORB) [33] that uses binary descriptor instead of floating point descriptor to achieve faster performance suitable for real-time applications.

In [34], the authors showed the comparisons among different image matching algorithms, such as SIFT, SURF, and ORB. They have manually performed transformation and deformation on the images in respect to rotation, scaling, fish eye distortion, noise, and shearing. The comparison was done based on different evaluation parameters, such as the number of keypoints in images, execution time, and matching rate. For most of the scenarios they have found SIFT performed best. The researchers in [35] tried to use depth map to perform smoothing process in the scale-space. They smoothed the scene surface considering smoothing quantity as a function of the distance given by the depth map so that 'the further a given pixel is, the less it is smoothed'. They tried to inject the smoothing filter in the SIFT algorithm and determined the repeatability score to evaluate the keypoint detection performance. Their goal was to find the keypoint repeatability under viewpoint position changes. However, the dataset we have used in our research was generated using single depth camera without changing the viewpoint positions.

Bag-of-Feature (BoF) representation was used in [36] to obtain a global information of visual data out of arrays of local point descriptors generated by SIFT algorithm. SIFT algorithm can extract higher dimensional feature points from the images even with lower resolutions but compromises the efficiency in terms of computation. To address this problem, BoF approach has been applied in reducing feature dimensions, redundancy elimination, and extracting global information from local SIFT features [36]. Moreover, the BoF approach has been considered as an efficient method to represent visual contents in hand gesture recognition [37]. The local feature points extracted from SIFT are fed into clustering algorithm to learn visual codebook and then each feature vector is mapped to a visual codeword represented by a sparse histogram. We have applied this technique to depth images for the classification using a multiclass SVM.

3. Proposed System

The proposed system consists of (1) hand segmentation and depth silhouette generation, (2) SIFT keypoints extraction, (3) clustering keypoints and generating BoF descriptors, and (4) symbolic gesture recognition using SVM.

The architectural diagram of the proposed approach is shown in Figure 2. The standard dataset [15] has considered 640×480 image resolution to capture the RGB image and the depth map of gesturing hand using Microsoft Kinect. Depth values are stored in millimeters. After calibration, we have applied the segmentation process as described in [15], except generating grey-scale variations on depth images.

3.1. Hand Segmentation. Segmentation is the process of removing the noninteresting area from the pertinent object. Many of the techniques in hand region segmentation worked on color space-based detection like skin-color detection, YCbCr/HSV color space filtering, and so on. These color-based techniques have limitations due to the noise, lighting variations, and background complexities. However, utilizing depth map information combined with color information improves the segmentation process which in turns gives better recognition accuracy.

Before segmenting the hand shape or region of interest (ROI), some preprocessing is performed. This involves calibrating the RGB and Depth Images. The RGB image is also converted into grey-scale. To extract the region of interest, first, we locate the smallest depth value from the depth image. This corresponds to the closest point of the hand from the camera plane. We call this value minimum distance. Next, an empirical threshold value is added to the minimum distance to give the segmentation threshold. This segmentation threshold is then used to segment the hand region from the rest of the image. This approach has proven to be robust in cluttered and noisy environments [38]. It is important to note that the hand should be the closest object to the camera for proper segmentation. The segmentation threshold is the sum of a minimum distance and a depth threshold. The minimum distance is easily obtained from the depth image as the minimum value in the depth matrix. The depth threshold is estimated based on different possible orientations of the hand shape.

After multiple measurements and testing, an upper bound is chosen as the depth threshold, such that the sum of the depth threshold and the minimum distance will allow us to isolate or segment the hand shape including the black

belt from the rest of the image. In our scenario, the depth threshold was estimated at 200 mm. The depth threshold is useful for filtering cluttered background containing an overlapped image (e.g., gesturing hand is overlapped with the face having the same color). We followed the same segmentation process as described in our previous work in [39]. However, in this research, the segmentation process is applied to a larger and challenging dataset [15]. Earlier, we used smaller dataset containing only 5 (five) static hand gestures representing numeric symbols 1 to 5 in a restricted environment, collected from a limited number of users.

3.1.1. Generating Depth Silhouettes Using Depth Map. The images from the Kinect depth stream are in 640 × 480 resolution which does not show enough contrast variations. Keypoints with low contrast will not give enough gradient variations to identify finger bending information. If we can generate contrast variation according to the depth values, then we can get more discriminative keypoints. These keypoints would be the salient features to improve the recognition accuracy. So, we have done some preprocessing where the depth values of gesturing hands were used to produce grey-scale levels. The closer a point is, the brighter its shade is. To do that, we cropped out depth values of the hands and got an m-by-n matrix with depth values of hands and its background.

Let $dist(x, y)$ be the distance of a point in the millimeter at (x, y). $f(x, y)$ is the corresponding grey level of the generated image used in extracting the key features by SIFT. Now, we select η as the number of grey levels between $greyLevel_{min}$ and $greyLevel_{max}$. We also selected η number of distance segments between $dist_{min}$ (minimum distance) and $(dist_{min} + dist_{th})$; where $dist_{th}$ is the distance, we assumed the hand would be from $dist_{min}$ and the depth threshold. We let the background be black in the generated image to get the better result using SIFT. We have applied (1) to generate the grey-scale image using only the depth values.

$$f(x, y) = \begin{cases} 0, & \text{if } dist(x, y) > dsit_{min} + dist_{th} \\ greyLevel_{min} + \left(\left\lfloor \left(\dfrac{dist(x, y) - dist_{min}}{dist_{th} - dist_{min}} \times \eta \right) + 0.5 \right\rfloor \times \left\lfloor \dfrac{greyLevel_{max} - greyLevel_{min}}{\eta} \right\rfloor \right), & \text{Otherwise} \end{cases} \quad (1)$$

We can see from the equation that any point in the depth image within the threshold distance is going to be a nonblack pixel depending on the grey levels determined from depth information. To assign grey levels to those pixels we segmented the depth values in η levels. Any distance value under the threshold is rounded off and normalized. The normalized distance values are converted into appropriate grey levels. After that, we find a grey-scale image which is the depth silhouette of a hand with the dark background and the grey levels corresponding to the depths of different parts of the subject hand. To emphasize the contrasts, the η number of segments was used. If we had used all the 256 levels of the grey image, the contrasts would not be prominent enough to get fair results. We considered $\eta = 10$ grey-scale levels from 155 to 255, dividing the levels equally to get a good contrast ratio. The number of levels was heuristically determined based on the assumption that more levels of grey will mean that the hand segments' contrast will be low. Thus, one of our main focuses (to represent distances in distinctive grey levels) would be undermined. Representing the distances using fewer grey levels would have the similar effect as the binary images. The shape would be distinct but the local features would be lost. Moreover, the grey-scale images with proper contrast are useful enough to distinguish the curves and angles of finger joints in different gestures. Both of the characteristics helped the SIFT to generate feature descriptors for the gestures, indifferent of the orientation of the hands.

For each gesturing image, we have extracted depth values within 200 from the depth image of the resolution 640 × 480. Actually, the 200 region contains the gesture information which we have used to generate the depth silhouettes. The process of segmentation and grey-scale varying depth silhouette generation are shown in Figure 3.

3.2. Feature Extraction. Features to be extracted by the feature extraction algorithm should present a high degree of invariance to scaling, translation, and rotation. Feature representation depends on the algorithm to be used for classification. We have used SIFT algorithm to represent the features as 128-dimensional feature points that are extracted from the depth images.

3.2.1. SIFT Features. The SIFT algorithm detects keypoints from a multiscale image representation consisting of blurred images at different scale. The keypoint location and the scale values of each keypoint are accurately determined using the Difference of Gaussian (DoG). Then the key points are filtered by eliminating edge points and low contrast points. After that, the orientation of the keypoint is determined based on the local image gradient within an image patch. Finally, the keypoint descriptor is computed which defines the center, size, and orientation of normalized patch [4]. We have used the SIFT implementation code as in [40].

Features generated by SIFT algorithm are invariant to scales and robust against changing position of object, slight rotation of object, and object in noisy and varying illumination condition in different images. These feature points can be found in the high-contrast regions and we have generated those contrast varying images based on depth values of the

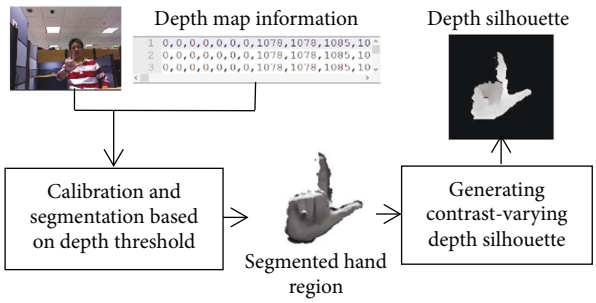


Figure 3: Hand gesture segmentation.

gesturing hand. SIFT algorithm effectively determines the keypoints on those depth images and represent them as feature descriptors.

The main objective of our approach is to improve the recognition accuracy for static gestures using depth information compared to binary and time-series representation of the images. We have utilized depth information and generated depth silhouettes which can be fed to any keypoint detector and descriptor-based algorithm, such as SIFT, SURF, and ORB. However, we have chosen SIFT to generate training and testing images. The training images with corresponding keypoints mapped over the gesturing image are shown in Figure 4. The first and third columns in Figure 4 represent the depth silhouette generated using depth map information of the gestures 1-10 (G1-G10) of the numeric symbols 0-9. The second and fourth columns in Figure 4 represent the corresponding hand gestures G1-G10 with 27, 41, 51, 61, 77, 101, 55, 56, 32, and 80 SIFT keypoints, respectively.

While extracting the keypoints we have found that the number of keypoints varies according to the type of gestures. As different symbolic gestures consist of a different number of fingers to be articulated, hence we got these variations. We captured 100 images per gesture as the candidates to generate keypoint descriptors and we got 41273 keypoints by considering 1000 images in total training images. The distribution of the number of keypoints per gesture is shown in Figure 5.

The keypoint descriptors that we have found are 128-dimensional feature vectors. Due to the changes in orientation, scale, and illumination of the same gesturing image by multiple persons the number of keypoints varies. Moreover, the dimensions of the gesturing images become larger which increases computations. Hence, we have used the strategy of a bag-of-visual-words and clustering technique to reduce dimensions.

3.2.2. Clustering Feature Descriptors. The dimension of the feature vector in each gesturing image varies based on the number of keypoints found for each gesture. The problem is that we need unified dimensional feature vectors as the training set to classify using multiclass SVM [41]. For the depth image that has 27 keypoints, the dimension of that image becomes $27 \times 128 = 3456$ and if another image from the same gesture contains 80 keypoints then the dimension becomes $80 \times 128 = 10240$. So, we have used the bag-of-word for which we need clustering to reduce the dimensions.

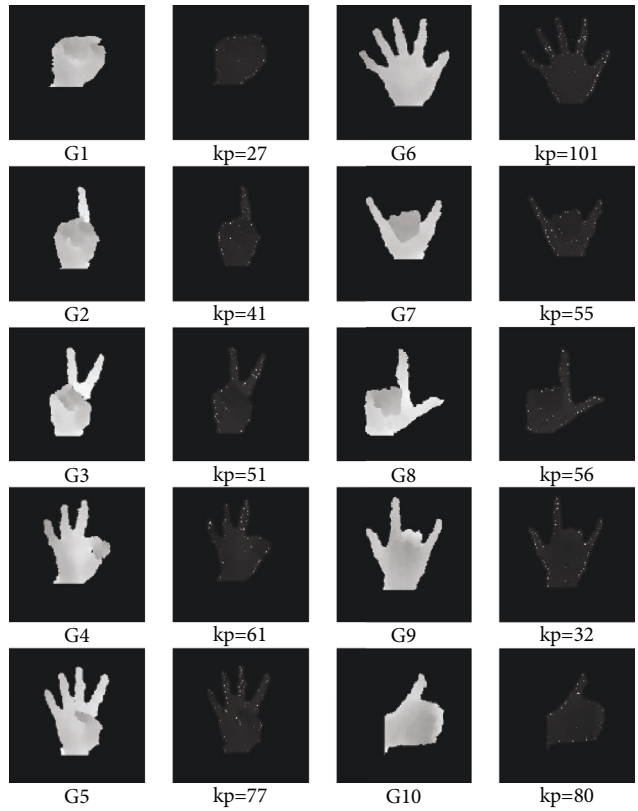


Figure 4: Example images containing generated depth silhouettes (first and third columns) and the corresponding SIFT keypoints mapped in to depth images (second and fourth columns) showing numeric symbols (0-9) representing the gestures (G1-G10).

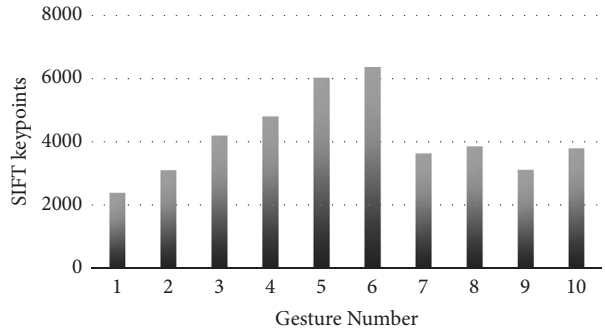


Figure 5: Number of SIFT keypoints at $\sigma = 1.8$.

The basic k-means clustering served our purpose because k-means converge faster than hierarchical-based clustering approaches. It also gives efficient performance for larger datasets. The keypoint distributions for different gestures are found to be almost Gaussian and distinctive as shown in Figure 5. In the concept of bag-of-word, the clusters are defined as codebooks and the size of the cluster determines the convergence property of the clustering technique. If we took smaller codebook size then bag-of-word vectors may not contain all the important keypoints. The larger codebook size may raise the overfitting problem. As the keypoints in

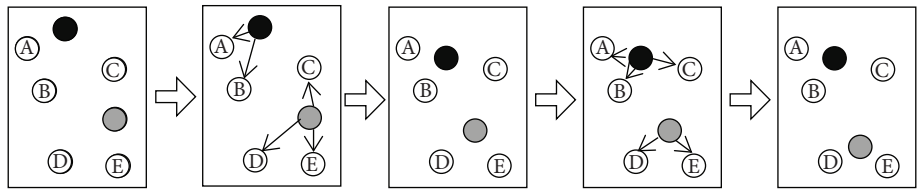


Figure 6: Demonstration of k-means clustering.

depth images are well distributed containing information about opening finger parts as well as bending finger parts, intuitively, we should get better accuracy.

To build our k-means clustering model, we have chosen 1600 as the cluster size which is the size of the visual vocabulary. An individual feature vector is assigned based on the nearest mean value while partitioning the feature vectors. After that, the code vectors were updated to reform the clusters until the grouping stops.

The goal of the k-means clustering approach is to minimize total intracluster distance using the following.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (2)$$

where k is the cluster size, n is the number of instances, and c is the cluster centroid of cluster j. An illustration of k-means clustering is shown in Figure 6 for five keypoints, A, B, C, D, and E, to form two clusters.

We develop the cluster model from each of the training images consisting feature vectors and encoded each of the keypoints with the clustered index. Keypoint and the cluster centroid are mapped according to the minimum distance criteria based on Euclidean distance measurement.

We got k disjoint subgroups of keypoints after assigning the keypoints to the corresponding cluster centers. So, the dimension of each training image consisting of n keypoints ($n \times 128$) reduced to $1 \times k$. k determines the cluster numbers.

3.2.3. Creating Bag-of-Features. We have created the bag-of-feature representation of each training image from the SIFT feature extracted. In order to learn visual vocabulary, we have built the k-means clustering model. Keypoints from each training image are mapped to the centroid of the corresponding cluster to represent visual vocabulary, known as feature vector quantization (VQ) process [42]. After that, we have represented each training image by the frequencies of visual words and found a unified dimensional histogram vector. The histogram representations of images of each gesture are ready for the classification. The process of creating Bag-of-features is shown in Figure 7

We updated the feature extraction process which is applied to two types of images; one is the depth image and the other one is the edge image, generated from the same dataset [15]. This is because we tried to establish more reliability in our approach through experimental evaluation compared to our previous work [39].

3.3. Recognition of Gestures Using SVM Classifier. The bag-of-feature vectors are now the input feature vectors for the classification algorithm. In order to recognize the performed symbolic gestures, we have applied a multiclass SVM training algorithm which is a supervised machine learning algorithm. It performs nonlinear mapping and transforms the training dataset into higher dimensional datasets. The algorithm tries to find out an optimal hyperplane which is linear.

SVM determines that the support vectors are closest to the separating hyperplane. The margins are also defined by those support vectors. Maximum separation is ensured by the maximum margin hyperplane.

We have applied the one-against-all approach to implement the SVM classifier [41] that built the model with respect to the training set supplied with group vector (class label indicator from gesture classes 1 to 10).

4. Experimental Evaluation

In order to evaluate the symbolic gesture recognition results, we have considered NTU hand gesture recognition dataset [15] which is a benchmark dataset in static hand gesture recognition. The dataset was collected using Kinect depth camera from 10 subjects. Each subject has performed 10 symbolic gestures 10 times. So, the dataset contains total of 1000 instances. Each gesturing instance contains a color image and the corresponding depth map. The dataset was prepared in a very challenging real-life environment containing the situations like the cluttered background and pose variations in terms of rotation, scale, orientation, articulation, changing illumination, etc.

We have conducted the 5-fold cross-validation process to evaluate our results. In each fold 4 of the image groups were used as training set and one of them was used as validation testing set. Each fold contains 20 images and we permuted the process, calculating the accuracy of SVM classifier. All the experiments were executed on an Intel Core I7 2.60 GHz CPU having 16 GB RAM.

Our system is robust to cluttered background due to the process of segmentation where the depth threshold and minimum hand-finger distance from the depth camera are used to determine the segmentation threshold. Good contrast varying depth silhouettes guarantee SIFT keypoints to be extracted in different scale-rotation-orientation changing conditions as shown in Figure 8.

SIFT extracted local features which produce good recognition results compared to global features considered in FEMD-based approach [15]. We tested our results in two types of images produced from the same dataset: binary images and image with edge information. The former was generated along with depth silhouette by converting the

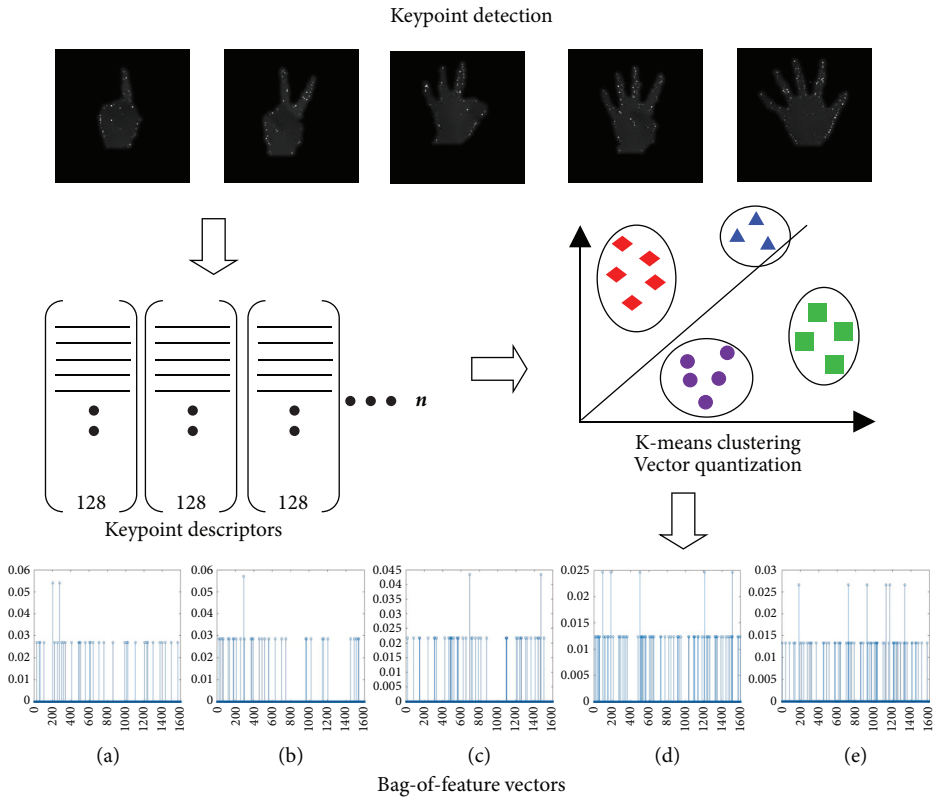


FIGURE 7: Generating bag-of-feature for training. (a)-(e) Bag-of-feature generated of gesture 2-6 from individual depth silhouette for 1600 clusters.

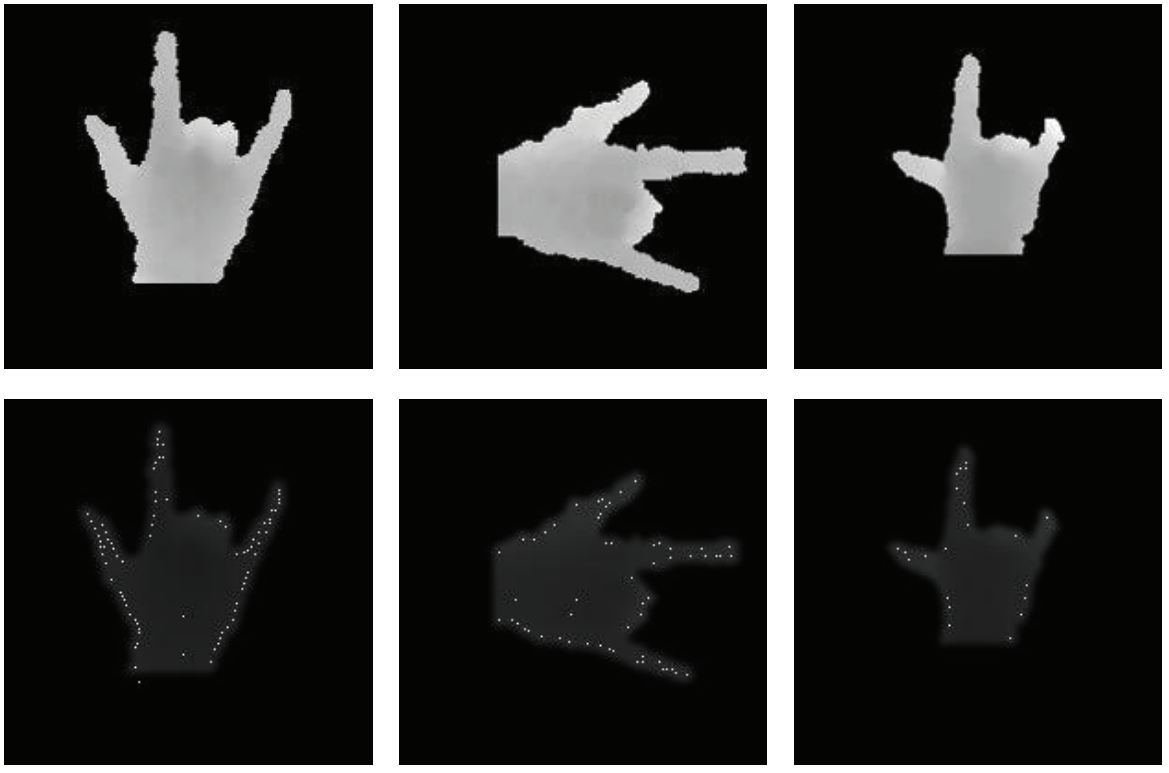
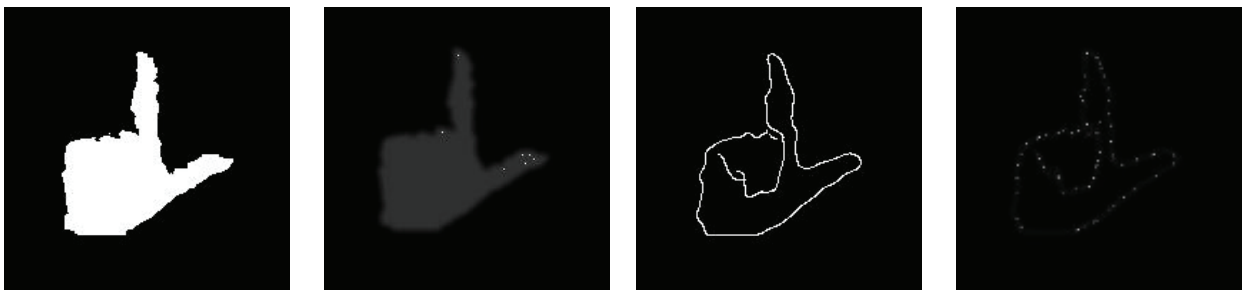


FIGURE 8: SIFT features are robust to orientation changes (b) and scale changes (c) along with normal pose (a).



(a) (b)

Figure 9: SIFT keypoints on binary image (a) and edge image (b).

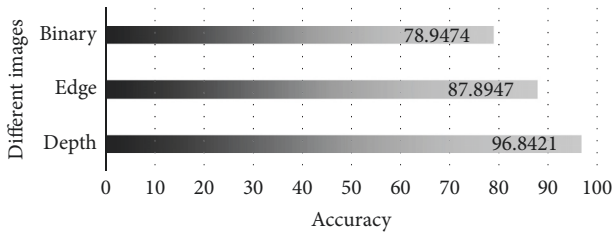


Figure 10: Accuracy comparison among different images.

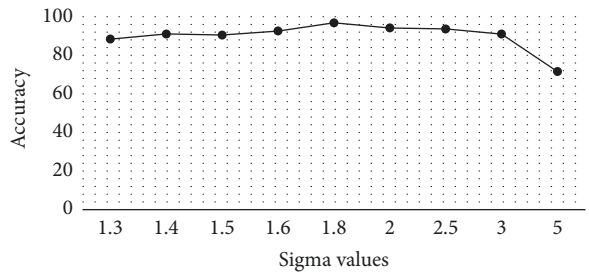


Figure 11: Accuracy at different sigma values.

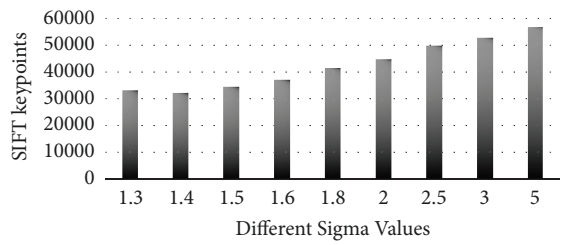


Figure 12: Number of SIFT keypoints at different Sigma values.

depth silhouettes into binary images and the latter was generated applying Canny's edge detection algorithm [25] on depth silhouettes. Example binary and edge images are shown in Figure 9. The image contains internal finger bending edge overlapped with palm and the external hand shape edge, but this information is not present in binary image or time-series images. So, the accuracy of our approach should vary on these different datasets.

Previously in [39], we demonstrated that the SIFT works better on depth images rather than binary images for static hand gesture recognition consisting of symbolic gestures (numeric symbol 1-5). The dataset used in the previous work was generated by us in a constrained environment. To create the dataset, we considered a limited number of hand gestures from a limited number of users. The comparison of experimental results was not performed among depth images, binary images, and edge images. However, in this research work, we have compared our experimental results among all the images and also compared the result with FEMD-based approach [15] and got higher accuracy for depth images (recognition accuracy is shown in Figure 10). Moreover, we elaborated the processes of depth silhouette generation with equations which illustrates the fact that the intensity of a pixel in grey-scale depends on the distance of that pixel from the depth camera. This, in turn, determines the contrast of the image based on depth values suitable for key point detector and descriptor-based algorithms.

To evaluate the accuracy of our approach, we generated different SIFT keypoints by varying the sigma (scaling parameter) value and found the highest accuracy at $\sigma = 1.8$. The mean accuracy at different σ values is shown in Figure 11.

With the increased value of sigma, we found more keypoints (Figure 12) which results in spurious DoG extrema considered as less stable and not linked to any particular structure in the image. These cause the differences in accuracy.

We evaluated the accuracy with the different number of clusters. We considered 100, 200, 400, 800, 1200, 1600, and 2000 clusters to validate our proposed method and compared the results for depth, binary, and edge images. The comparison result is shown in Figure 13. We observed that the accuracy increments commensurate with the higher number of clusters. The highest accuracy we attained has been with a cluster size of 1600. This phenomenon can be traced back to depth images which significantly contribute to the salient keypoints identification for it is the depth images from which we can distinguish the positions of each fingers. However, the same cannot be said for binary images or images containing only edge information. FEMD has considered the shape distance metric which matches only opening finger parts or finger shapes, not the whole hand. While making a pose the bending finger parts are also important to distinguish slightly varying gestures, which can be found in the local features. To avoid local distortion we have chosen the correct scale factor.

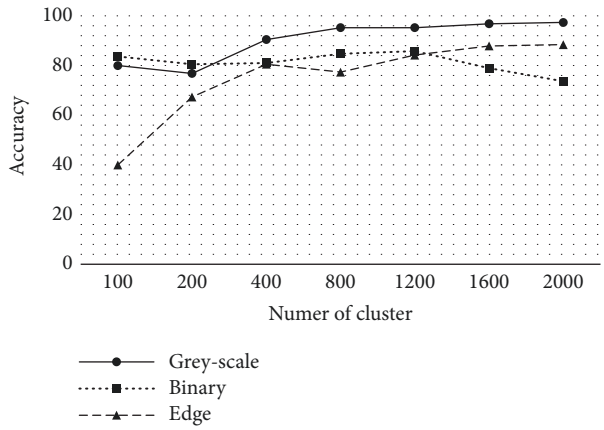


Figure 13: Overall accuracy comparison among different images.

1	95	0	0	0	0	0	0	0	0	5
2	0	98	0	0	0	0	0	1	0	1
3	0	0	94	0	0	1	0	0	0	5
4	0	0	0	98	1	1	0	0	0	0
5	0	0	0	0	97	2	0	0	0	1
6	0	0	0	2	0	91	0	0	0	7
7	0	0	0	0	0	0	100	0	0	0
8	0	0	0	0	0	0	0	100	0	0
9	0	0	0	0	5	0	0	0	95	0
10	0	0	0	0	0	0	0	0	0	100
	1	2	3	4	5	6	7	8	9	10

(a)

1	95	1	0	0	0	0	0	3	1	0
2	3	86	4	2	0	0	1	4	0	0
3	0	2	94	2	0	0	2	0	0	0
4	0	0	4	87	6	0	3	0	0	0
5	0	0	0	7	89	3	1	0	0	0
6	1	2	0	0	0	95	0	0	2	0
7	0	0	1	0	0	1	96	2	0	0
8	6	2	0	0	0	0	0	92	0	0
9	1	0	0	0	0	1	0	0	98	0
10	0	0	0	0	0	0	0	0	0	100
	1	2	3	4	5	6	7	8	9	10

(b)

Figure 14: Confusion matrix of (a) proposed approach and (b) FEMD-based approach.

We have presented the input hand as a contrast varying grey-scale image depending on the depth map information but FEMD has presented the hand image as a global feature using time-series curves. Shape contour presentation introduces lower accuracy in terms of scale, rotation, or orientation changes which we have overcome through depth images and got accuracy up to 96.8421% whereas the FEMD has produced 93.2%. The confusion matrix of our approach and FEMD is given in Figure 14.

We have also calculated True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) and based on these the F-Score values using the following.

$$F - Score = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (3)$$

The class-wise F-Score comparison between our approach and FEMD is given in Figure 15.

From Figure 14(a), we find that the accuracy of gestures 2, 4, 5, 7, and 8 has been improved significantly as expected because SIFT features are found more robust in the benchmark dataset. Moreover, we prove this by comparing the results with binary and edge images. In binary or edge images, a small variation in the shape may cause significant changes on the tangent vectors at the points on the shape. Since we are considering local hand-finger features for the hand

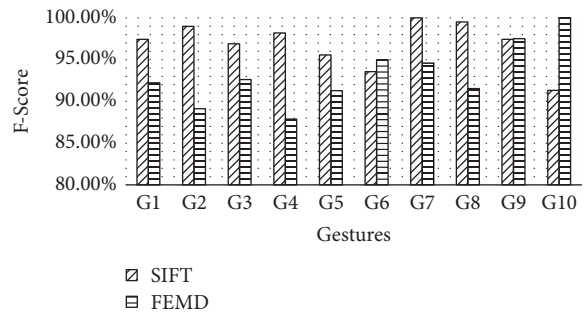


Figure 15: F-Score comparison between proposed approach and FEMD [15].

poses we are getting better results. Shape changes over time-series data are not required to be considered. Recognition accuracy of gestures 4 and 5 has increased to 98% and 97%, respectively, compared to FEMD-based approach. In gestures 6 and 10, we are getting most confusing results. Gesture 6 is all finger open gesture and contains a maximum number of keypoints (6374) as shown in Figure 5 and includes no bending finger information. The same is for gesture 10 and it is the only gesture in the dataset which contains no bending finger information like other gestures. The pose was given by the user opposite to other gestures; the bending fingers were facing towards the user, not the camera.

5. Conclusion

This paper describes a symbolic hand gesture recognition system and presents an effective way of utilizing depth map information. The use of depth value in determining grey-scale levels to generate contrast varying depth images is one of the significant contributions to this research. Moreover, hand-finger's context information of the gesturing hand represented by local invariant feature descriptors has contributed the recognition accuracy up to 96.84% which is better compared to binary images, images containing edge information, and images represented in time-series curves.

Preparing depth silhouettes of the gesturing hand is one the factors that affect the accuracy of gesture recognition system. With the help of depth map information, we were able to produce those gesturing images using fast and effective segmentation process. Choosing the right cluster size is also important. Our empirical results indicate that 1600 is the most desirable number of clusters to attain the best accuracy. This large number of clusters is contributed by the fact that images with only edge information or binary images contain far less keypoints than that of depth images. The number of training samples that we have taken was sufficient to develop the cluster model as well as the SVM classification model.

In future, we will analyze gesture recognition accuracy in terms of variations in cluster size using the principal component analysis (PCA), adaptive grey-scale levels, combining local, and global features (containing contour information) using hierarchical classification techniques. We will also try to compare the performance result of different detector and descriptor-based algorithms, such as SURF, BRIEF, and ORB.

Data Availability

The dataset used to support the findings of this study has been deposited in the following link: https://drive.google.com/file/d/0B_9saHAqGFITODNmNzU0ZjctMjk0Yi00YjI-5LWJmZDMtYTdiYTE2YzM5OTQ4/view.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] H. Hasan and S. Abdul-Kareem, "Human-computer interaction using vision-based hand gesture recognition systems: a survey," *Neural Computing and Applications*, vol. 25, no. 2, pp. 251–261, 2013.

[2] Y. Park, J. Lee, and J. Bae, "Development of a wearable sensing glove for measuring the motion of fingers using linear potentiometers and flexible wires," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 1, pp. 198–206, 2015.

[3] J. Suarez and R. R. Murphy, "Hand gesture recognition with depth images: A review," in *Proceedings of the 2012 RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pp. 411–417, Paris, France, September 2012.

[4] D. G. Lowe, "Distinctive image feature from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] S. Zhang, W. He, Q. Yu, and X. Zheng, "Low-cost interactive whiteboard using the Kinect," in *Proceedings of the 2012 International Conference on Image Analysis and Signal Processing (IASP)*, pp. 1–5, Hangzhou, November 2012.

[6] Y. Chang, S. Chen, and J. Huang, "A Kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities," *Research in Developmental Disabilities*, vol. 32, no. 6, pp. 2566–2570, 2011.

[7] A. Ramey, V. Gonzalez-Pacheco, and M. A. Salichs, "Integration of a low-cost RGB-D sensor in a social robot for gesture recognition," in *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI '11)*, pp. 229-230, Lausanne, Switzerland, March 2011.

[8] M. Van Den Bergh, D. Carton, R. De Nijs et al., "Real-time 3D hand gesture interaction with a robot for understanding directions from humans," in *Proceedings of the 20th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2011*, pp. 357–362, USA, August 2011.

[9] D. Xu, Y. Chen, C. Lin, X. Kong, and X. Wu, "Real-time dynamic gesture recognition system based on depth perception for robot navigation," in *Proceedings of the 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 689–694, Guangzhou, China, December 2012.

[10] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the kinect," in *Proceedings of the 2011 ACM International Conference on Multimodal Interaction, ICMI'11*, pp. 279–286, Spain, November 2011.

[11] O. M. Mozos, H. Mizutani, R. Kurazume, and T. Hasegawa, "Categorization of indoor places using the Kinect sensor," *Sensors*, vol. 12, no. 5, pp. 6695–6711, 2012.

[12] G. Azzari, M. L. Goulden, and R. B. Rusu, "Rapid characterization of vegetation structure with a microsoft kinect sensor," *Sensors*, vol. 13, no. 2, pp. 2384–2398, 2013.

[13] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," *Computer Vision and Image Understanding*, vol. 141, pp. 152–165, 2015.

[14] W. Lin, Y. Wu, W. Hung, and C. Tang, "A Study of Real-Time Hand Gesture Recognition Using SIFT on Binary Images," in *Advances in Intelligent Systems & Applications, SIST 21*, pp. 235–246, Springer-Verlag, Berlin Heidelberg, 2013.

[15] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

[16] T. D'Orazio, R. Marani, V. Renò, and G. Cicirelli, "Recent trends in gesture recognition: How depth data has improved classical approaches," *Image and Vision Computing*, vol. 52, pp. 56–72, 2016.

[17] C. Wang and K. Wang, *Hand Gesture Recognition Using Adaboost With SIFT for Human Robot Interaction*, vol. 370, SpringerVerlag, Berlin, Germany, 2008.

[18] Q. Chen, N. Georganas, and E. Petriu, "Real-time vision-based hand gesture recognition using Haar-like features," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference Proceedings (IMTC '07)*, pp. 1–6, Warsaw, Poland, May 2007.

[19] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 2, no. 57, pp. 137–154, 2004.

[20] D. Lisin, M. Mattar, M. Blaschko, E. Learned-Miller, and M. Benfield, "Combining Local and Global Image Features for Object Class Recognition," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pp. 47-47, San Diego, CA, USA.

[21] K. Siddiqi, S. Bouix, A. Tannenbaum, and S. W. Zucker, "Hamilton-Jacobi skeletons," *International Journal of Computer Vision*, vol. 48, no. 3, pp. 215–231, 2002.

[22] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[23] H. Ling and D. W. Jacobs, "Shape classification using the inner-distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 286–299, 2007.

[24] K. Matusiak, P. Skulimowski, and P. Strumillo, "Depth-based descriptor for matching keypoints in 3D scenes," vol. Volume 64, Issue 3, pp. 299–306, 2018.

[25] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[26] X. J. Meng, Y. L. Yin, G. P. Yang, and X. M. Xi, "Retinal identification based on an improved circular gabor filter and scale invariant feature transform," *Sensors*, vol. 13, no. 7, pp. 9248–9266, 2013.

[27] S. Zhao, X. Xu, W. Zheng, and J. Ling, "Registration of Depth Image and Color Image Based on Harris-SIFT," in *Proceedings of the IEEE 2012 Second International Conference on Electric Information and Control Engineering (ICEICE*, 2012.

[28] Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 2, pp. II-506–II-513, July 2004.

[29] H. Bay, A. Ess, T. Tuytelaars, and L. Gool, "SURF: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[30] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[31] L. Zhang, P. Shen, G. Zhu, W. Wei, and H. Song, "A fast robot identification and mapping algorithm based on kinect sensor," *Sensors*, vol. 15, no. 8, pp. 19937–19967, 2015.

[32] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 6314, no. 4, pp. 778–792, 2010.

[33] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 2564–2571, Barcelona, Spain, November 2011.

[34] E. Karami, S. Prasad, and M. Shehata, "Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images," in *Proceedings of the 2015 Newfoundland Electrical and Computer Engineering Conference, St*, johns, Canada, November 2015.

[35] M. Karpushin, G. Valenzise, and F. Dufaux, "Keypoint detection in rgbd images based on an anisotropic scale space," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1762–1771, 2016.

[36] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2169–2178, June 2006.

[37] D. A. Leite, J. C. Duarte, J. C. Oliveira, V. d. Thomaz, and G. A. Giraldi, "A System to Interact with CAVE Applications Using Hand Gesture Recognition from Depth Data," in *Proceedings of the 2014 XVI Symposium on Virtual and Augmented Reality (SVR)*, pp. 246–253, Piata Salvador, Bahia, Brazil, May 2014.

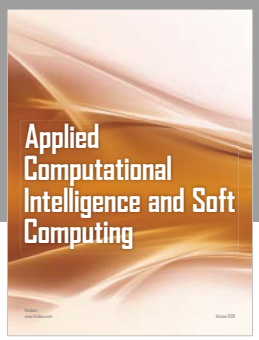
[38] Y. Li, "Hand gesture recognition using Kinect," in *Proceedings of the 2012 IEEE 3rd International Conference on Software Engineering and Service Science, ICSESS 2012*, pp. 196–199, China, June 2012.

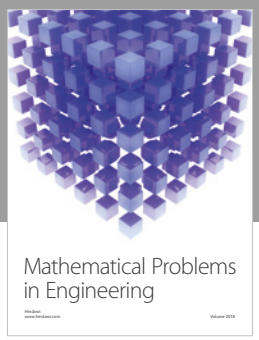
[39] H. Mahmud, M. K. Hasan, A. A. Tariq, and M. A. Mottalib, "Hand Gesture Recognition Using SIFT Features on Depth Image," in *Proceedings of the the Ninth International Conference on Advances in Computer-Human Interactions (ACHI)*, pp. 359–365, Venice, Italy, 2016.

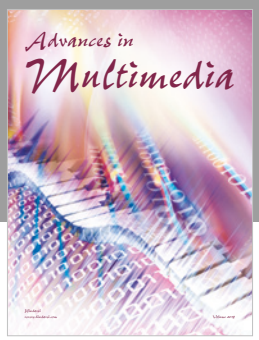
[40] C. Naveen, *SIFT algorithm*, vol. 11, MATLAB Central File Exchange, 2018, https://www.mathworks.com/matlabcentral/fileexchange/43723-sift-scale-invariant-feature-transform-algorithm.

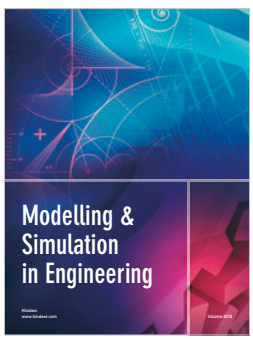
[41] "Multiclass Support Vector Macine Classifier," http://www.mathworks.com/matlabcentral/fileexchange/33170-multi-class-support-vector-machine/.

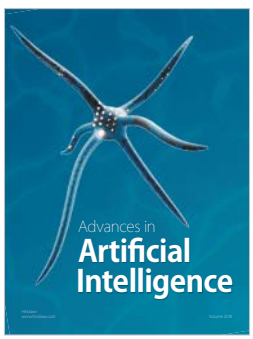
[42] A. Bosch, X. Muñoz, and R. Martí, "Which is the best way to organize/classify images by content?" *Image and Vision Computing*, vol. 25, no. 6, pp. 778–791, 2007.

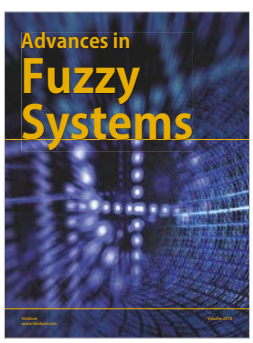


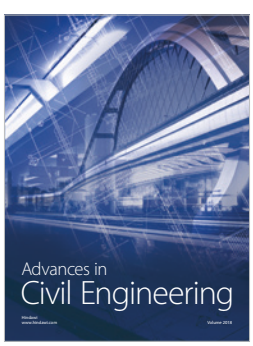


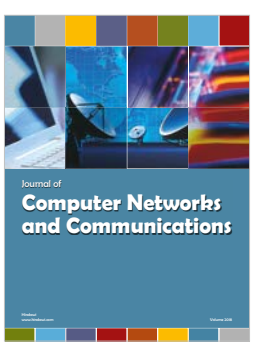


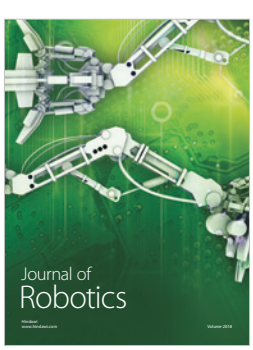


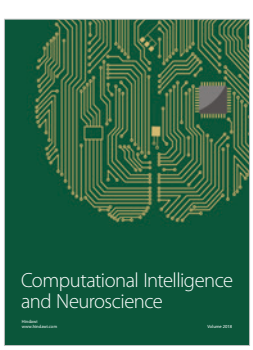












Hindawi

Submit your manuscripts at
www.hindawi.com