ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 944-954 (2020)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Dense SIFT–Flow based Architecture for Recognizing Hand Gestures

Suni S S[*,1] K Gopakumar[2]

[1]*LBS Centre for Science & Technology, University of Kerala, Kerala, 695033, India*

[2]*TKM College of Engineering, Kollam, Kerala, 691005, India*

ARTICLE INFO

Article history:
Received: 04 July, 2020
Accepted: 29 September, 2020
Online: 20 October, 2020

Keywords:
Human-computer interaction
Hand segmentation
Scale Invariant Feature Transform
Support Vector Machine
K-Nearest Neighbour
Hand gesture recognition

ABSTRACT

Several challenges like changes in brightness, dynamic background, occlusion and inconsistency of camera position make the recognition of hand gestures difficult in any vision-based method. Diversity in finger shape, size, distribution and motion dynamics is also a big constraint. This leads to the motivation in developing a dense Scale Invariant Feature Transform (SIFT) flow based architecture for recognizing dynamic hand gestures. Initially, a combination of three frames differencing and skin filtering technique is used for hand detection to reduce the computational complexity followed by a SIFT flow technique to extract the features from the detected hand region. SIFT flow vectors obtained from every pixel can lead to overfilling, data redundancy and dimension disaster. A dual layer belief propagation algorithm is utilized to optimize the feature vectors to resolve the dimensionality problem. Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifiers are used to evaluate the performance of the developed framework. Experiments were conducted on hand gesture database for HCI, Sebastien Marcel Dynamic Hand Posture Database and RWTH German finger spelling database. The simulation results demonstrate that the developed architecture has excellent performance on the uneven background and varying camera position and it is robust against image noise. A comparative analysis with the state of the art methods illustrates the effectiveness of the architecture.

1. Introduction

Man's innate, innovative and intentional desire to grab and cope up with the smart devices and interactive systems in the modern era is fathomless. Recent studies show that 80% of the population uses smart decision making systems in one or another way [1]. For example, a person who is doing multiple tasks can change display of computer using a hand movement or gesture, while reading newspaper. Artificial intelligence leaps to the zenith that it becomes an art in presenting artistic perfection and accuracy increasing the percentage of digitalized people resulting restless efforts to propose new algorithms for interacting with machines. Numerous approaches are there to tap and convert gesture movement captured using vision sensors to decision support systems [2,3]. The work of Liu and Zhang was a human computer conversation based on hand gestures from the video streams using Hu invariant moments and support vector machine [2]. A system to recognize the basketball referee signals from videos using image segmentation and support vector machine classifier was build [3].

In human computer interaction systems, the present and widely used trend is dynamic texture based techniques [4,5] and multimodal feature based frameworks [6,7] because of its powerful representation. One of the most broadly used features is local binary pattern (LBP) introduced by Ojala [8] for texture classification, that proved to be useful in computer vision applications. The authors of [4] was developed a dynamic texture based technique for human-computer interaction using volumetric spatiograms of local binary patterns. The approach is congenial to detect static and dynamic gestures with the expense of high dimensional feature vectors. A spatio temporal descriptor called directional local ternary pattern from three orthogonal planes was introduced by the authors of [5] to classify the dynamic scenes and utilized convolutional neural network for detection. However, performance of these systems are found to be affected by external environment such as complex background and image noise and most of the systems perform well only for specific dataset. Human

[*] Corresponding Author: Suni S S, suni.ss@gmail.com

gestures are unique in velocity and motion scale. Another work was to build a multimodal feature based framework that fuses shape and movement information for recognizing facial expressions. The design uses histogram of gradients and optical flow features to obtain the video descriptor [6]. The authors of [7] developed a framework that uses combination of moments and directional wavelet local binary pattern for recognizing human actions. These methods give higher recognition rate at the cost of high dimensionality of feature vectors. The present attempt is to reduce the changes between distinct people performing the same gesture and to exaggerate the coherence of the gestures. Some of the constraints observed in the above mentioned literature are summarized as follows:

- Most feature vectors are calculated on limited number of pixels in an image. So it will not take the micro scale variations of movements (eg: finger movements). This will affect the overall accuracy of the system.
- Suggested methods are not strong enough to avoid image noise.
- High dimensionality of feature vectors leads to increase in computational time.
- Most of the systems perform well only for specific dataset. Robustness is needed.

To solve the above-mentioned drawbacks, a newly refined descriptor, dense SIFT flow that preserves the local appearance features of the image and the temporal variations (movement) of the video can be utilized. Dense SIFT flow algorithm calculates the flow vectors of the SIFT image. The proposed architecture is invariant to rotation, scale, illumination, view point and noise.

The main contributions of this work are,

- An architecture that accounts the spatial and temporal consistency, and can be manage to adapt itself to diverse lighting environments. Moreover, it can improve the system accuracy when connected with gesture identification methods.
- In the pre-processing stage to reduce the computational complexity and dimensionality of feature vector a hand segmentation method is applied.
- A sift flow algorithm that captures features from video frames in terms of spatial coherence and motion. It improves the discriminative power and effectiveness of the feature vector.
- System is tested for three different databases with varying background and viewpoints. The results obtained indicates the robustness of the system.

The rest of the article is framed as follows: Related works in the field of gesture recognition are detailed in Section 2. Theoretical framework of the proposed approach is given in section 3. The classifiers and its detection process is explained in section 4. Simulations, performance evaluations, comparative study of proposed algorithm with other existing methods are illustrated in section 5 followed by conclusion in section 6.

2. Related works

Attention has been focused in describing hand gesture using different feature based approaches, i.e., local and global feature based approaches. Recognition of hand gestures involves segmentation, preprocessing, feature extraction, gesture description and representation & classification.

Mostly, the state of movement of objects in the whole frame is represent as optical flow at a given time. Optical flow enables the foreground object motion study under static background. The methods based on the optical flow [9-11] are mostly used one to find the displacement of an entity in a image sequence. Estimation of optical flow captures local dynamics in the temporal domain and is computatinally efficient. The authors of [9] improved the classical optical flow algorithm by adding median filtering during optimization, that help to preserve better motion details. Dense optical flow algorithm is applied to label data for convolutional neural networks for motion detection in one work [10]. Another work [11] which utilized optical flow to detect moving objects and to estimate depth using triangulation. Optical flow algorithm with Kalman filter is successfully used for tracking multiple objects [12]. The authors of [13] proposed a solution with very low time complexity and accuracy for the estimation of dense optical flow. To detect obstacles at the time of landing of a moving air vehicles, a self-supervised learning (SSL) technique based on optical flow cues was developed [14]. All these approaches, however, do not take into account the characteristic temporal dynamics of all micro variations in movements and are limited to preserve local appearance representations.

A new system is introduced by authors of [15] based on a Single-Shot Multibox Detector (SSD) deep learning algorithm for hand gesture recognition. The system identifies the hand gestures in complex scenes with limited test cases. The authors of [16] build a framework that uses knowledge from multiple modalities by utilizing separate 3D convolutional neural networks (3D-CNNs) for recognizing dynamic hand gestures. The authors of [17] integrated density clustering and image entropy to obtain the key frames to enhance the performance of fast and robust gesture recognition and feature combination technique is also introduced. They also created two datasets named Action3D and Hand Gesture datasets. A dynamic hand gesture recognition system using skeletal representation of hand introduced in [18]. The system used a hand topology technique called later hand skeletal data to capture kinematic descriptors from the gesture frames and were encoded in using a multi-level temporal pyramid and a Fisher kernel. Mahayuddin and Saif [19] has described a method to improve learning process of mathematics using modified extrusion method for recognizing hand gestures which converts 2D shape to 3D shape. The method achieved good accuracy, but robustness of the system has to tested. Handling the large dataset is also a big concern.

The idea of using local appearance based feature like SIFT was introduced by Lowe [20]. SIFT algorithm converts an image into a set of local feature vectors. Each of these feature vectors is is distinctive and invariant to any image translation, rotation or scaling. After introduction of this feature, it has been applied in the different areas of video processing problems like gesture recognition, image registration and image retrieval. Hossein-Nejad and Nasri utilized SIFT features and RANdom SAmple Consensus (RANSAC) transform in image registration for extracting and matching features [21]. SIFT features are applied in gender and gaze recognition problem and produced good results in [22]. One

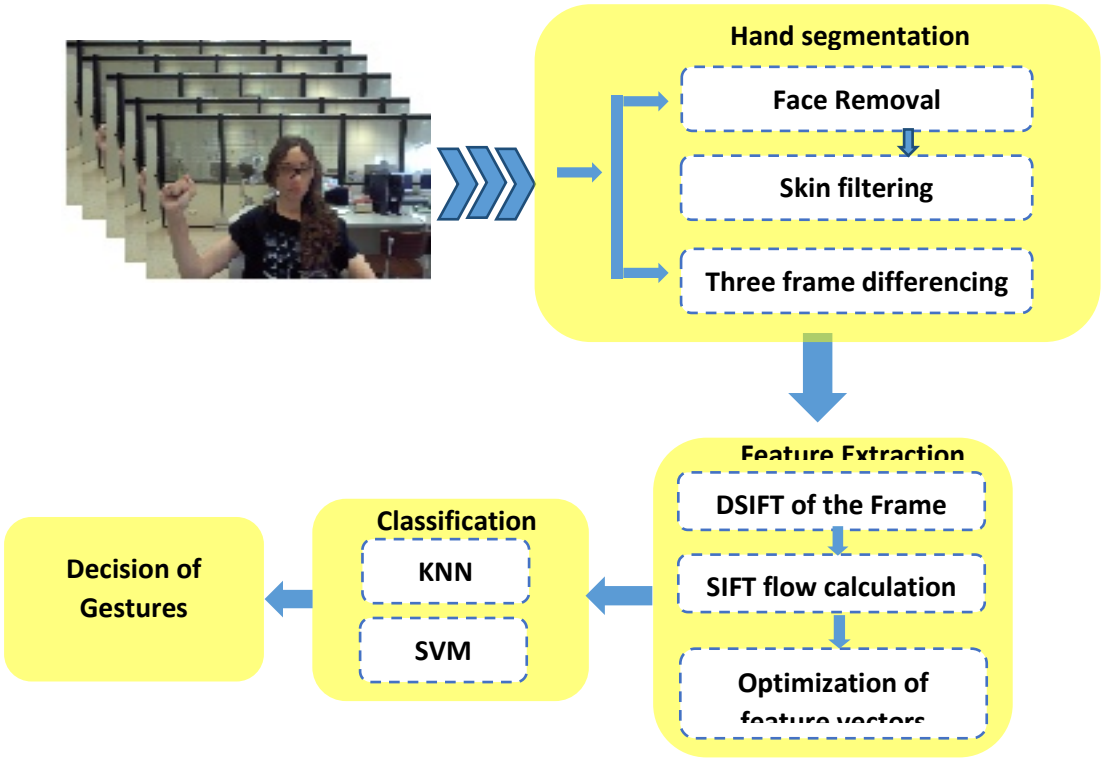


Figure 1: Proposed architecture of hand gesture recognition

disadvantage of using this feature matching method is that they become computationally expensive. Furuya and Ohbuchi [23] introduced the dense sampling of SIFT and a fast encoding algorithm for the application of 3D model retrieval. The dense SIFT algorithm takes the local gradient information of all pixels in an image and has the limitation for capturing motion dynamics of the video. This leads to the motivation in developing a SIFT flow based framework that captures spatial and motion features for detection.

SIFT flow vectors generated are high dimension. Optimization is necessary to reduce the dimensionality of feature vectors. Numerous methods are used for optimizing the SIFT descriptors. Daixian adopt key point location optimization using Chamfer distance [24]. The technique is proved to be reduce computation time and achieves good accuracy. Here the optimization was done before descriptor generation. VSA(Vector Symbolic Architecture) is an encoding technique which converts the high dimensional vectors into the sparse discrete representation [25]. The mapping is one of the key constraints when working on real time data. The double layer belief propagation was proved to be a good optimization technique for SIFT flow vectors [26].

A framework is developed for dynamic hand gesture recognition, wherein multimodal behavior of the image sequences in terms of spatial coherence and motion is extracted. In this application, a descriptor called dense sift flow is utilized which takes flow vectors of the local gradients. Dense SIFT flow refers to taking dense sampling in space of all image sequences. An algorithm is applied that performs gesture analysis to recognize and take decisions to interact with machines. It yields dense sift flow fields that are robust to image noise. Thus, the final video descriptor achieves highly discriminative hand gesture representation. The method is invariant to scale, rotation, illumination, view point and noise. It is observed that the developed architecture gives a better recognition rate.

3. Our Architecture

The prime motivation of the proposed work is to build an efficient framework which can capture shape and motion features from image frames for recognizing hand gestures. There are four phases in the present approach: pre-processing, feature extraction, optimization and detection. In the first stage, the hand is segmented by using three frame differencing and skin filtering [24]. The feature vectors are evaluated using dense SIFT flow algorithm and create a descriptor representation that can express the particular gesture. Classifiers are trained with feature vectors and then assign the new descriptors into various categories of gestures. The proposed architecture of a hand gesture recognition system is shown in Figure 1. The subsections will describe the SIFT flow algorithm and its optimization.

3.1. Hand Segmentation

Hand segmentation is the first step in any hand gesture recognition process. The work flow of hand segmentation process is shown in Figure 2. There are three main processes that combined together to achieve the segmented image. They are hand region tracking, skin filtering, and three frame differencing of both color and grayscale image frames [27].

Initially, the face of the person performing gesture is identified and eliminated from the second image of the video using the Viola–Jones algorithm [28, 29]. Skin filtering operation [30] is performed after removing the face to retrieve only the skin

coloured objects from the image frames. In the meantime, three-frame differencing is carried out with the first three frames of both coloured and grayscale sequences. Morphological operations (AND/OR) are done as shown in Figure 2 to attain the expected results. The desired hand is obtained from the combined results of the three-frame differencing and skin filtering. But, it has been notifying that some small skin coloured objects are also present with the desired hand. Thus, the largest binary linked object is selected and is treated to be the required hand in the scene.

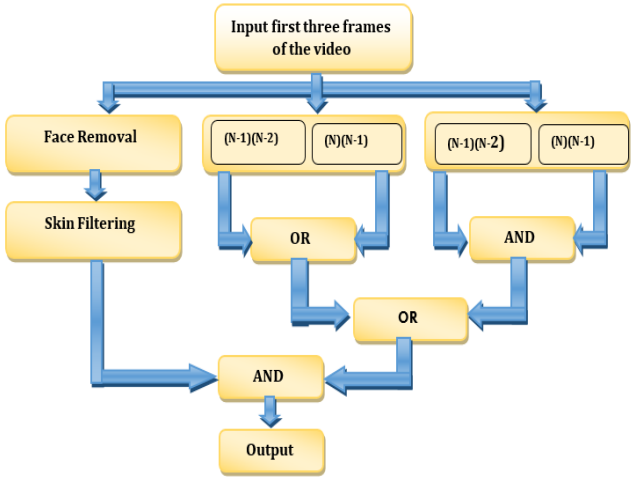


Figure 2: Work flow of Hand segmentation process

Algorithm for Hand Segmentation

Input :*Colour image frame from the video*
Output: *Segmented Hand Image*

1. For n =1 to 3 do
2. I(n) = input colour image frame
3.%%%Colour to gray scale conversion of image
4. I1(n) = rgb I(n) to the grayscale image conversion
5. if n =2 then
6.%%%Face detection of image
7.J = face detection using Viola – Jones algorithm
8.%% Removing face
9.J1 = Remove face J from frame I(n)
10.J2 = Skin filtering of J1 image
11. end if
12. if n =3 then
13.K1 = Skin filtering of (I(n-2) – I(n-1))
14. K2 = Skin filtering of (I(n-1)-I(n))
15.L = OR (K1, K2)
16.M1 = Grey to binary (I(n-2) – I(n-1))
17.M2 = Grey to binary (I(n-1)-I(n))
18.P = AND(M1,M2)
19.Q = OR(P,L)
20.R = AND (Q,J2)
21. %% Biggest binary linked object is selected
22.Seg_image = BBLOB(R)
23.end if
24.end for

Figure 3: Result of hand segmentation algorithm

Figure 3 shows the result of hand segmentation algorithm. The yellow box is the segmented hand region.

3.2. SIFT flow algorithm

3.2.1. Dense SIFT

Object tracking problems are usually solved using visual extraction algorithms. Dense SIFT (Dense scale invariant feature transform) is an innovative step in this field conducive even there is change in the background and appearance of the tracked object during tracking [31]. The process of extraction of dense SIFT is done by taking SIFT histogram for all pixels at a single scale having overlapping patches. A dense sampling assures a good characterization of foreground gestures as well as the consistency of features in the adjacent image frames. The following steps have been used to obtain dense SIFT descriptor. For each pixel of a segmented grey scale image the derivative value ρ and the derivative direction θ can be represented as

$$\rho(x,y) = \sqrt{dx^2 + dy^2}$$
$$\theta(x,y) = \tan^{-1}\frac{x}{y} \quad (1)$$

SIFT descriptor indicates the neighbourhood gradient information. As such, every pixel in an image, select 16 x16 neighbourhood, into 4 x4 cell array. The histogram of derivative direction comprising each block of size 4x4 is obtained and quantizing the orientation into 8 bins. Thus obtaining 4 x4 x 8 = 128 dimensional SIFT description for a pixel. Figure 4 shows the principle of SIFT local feature generation.

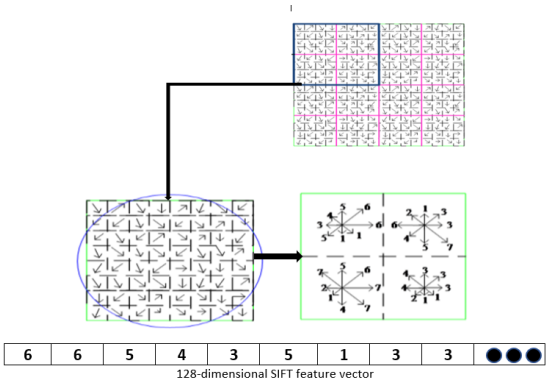


Figure 4: Principle of SIFT feature generation

3.2.2. SIFT flow algorithm

The Dense optical flow [32] opened the avenue of tracking the flow fields very densely from two consecutive SIFT images. SIFT flow is the latest innovative step by authors of [26] and is formulated to adjust an image frame to its nearest neighbor for huge image changes in scale. SIFT flow vector is obtained as follows. Assume $p = (x, y)$ be the grid coordinates of the image frames and $\omega(p) = (v(p), u(p))$ be the flow vector at Where $v(p)$ and $u(p)$ are integers with L possible states. s_1 and s_2 be the two SIFT images and ϵ contains all spatial neighborhoods (4 neighbourhood). The objective energy function of SIFT flow is designed as,

$$E(\omega) = \sum_p \min(\|s_1(p) - s_2(p + \omega(p))\|_1, t) \quad + \tag{3}$$

$$\sum_p \eta(|v(p) + u(p)|) \quad + \tag{4}$$

$$\sum_{(p,q)\in\varepsilon} \min(\alpha|v(p) - v(q)|, d) \quad + \quad \min(\alpha|u(p) - u(q)|, d) \tag{5}$$

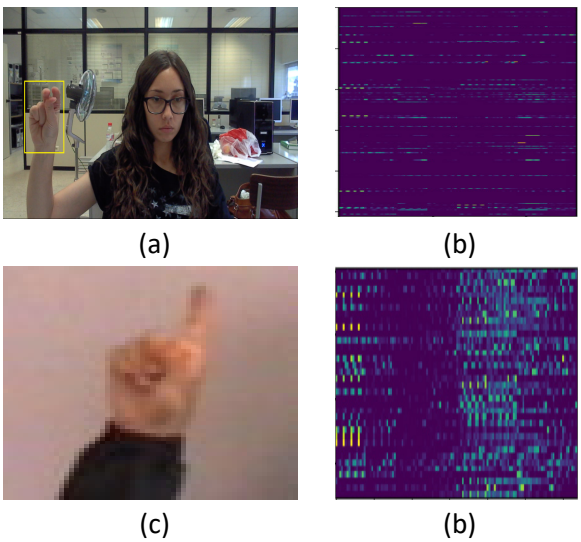


Figure 5: Sample image and its dense SIFT image (a) shows the sample image from hand gesture database for HCI. (b) shows its dense SIFT image. (c) shows sample image from Sebastien Marcel Dynamic Hand Posture Database (d) its dense SIFT image

The data term in (3) is a SIFT descriptor match constraint to be matched via the flow vector $\omega(p)$. The small displacement constraint in (4) allows the flow vector to be as small as possible when no other information is available. The smoothness term inhibits (5) the neighboring pixels to have similar displacement. In this objective energy function, truncated L1 norm is used as a matching threshold in both the smoothness and data term with $d$and t respectively (d=0.1, t=0.01). Figure 5 shows the sample image and its dense SIFT image.

3.2.3. Optimization of feature vector

A dual layer loopy belief propagation method is utilized for optimizing the objective energy function (5) [31]. Figure 6 shows the components in horizontal and vertical direction using dual layer loopy belief propagation. The direct application of SIFT flow vectors to the high dimensional image will result in poor performance. Hence, a coarse to fine SIFT flow matching method is used to formulate the correspondence between two consecutive frames, to accelerate the speed and matching performance.

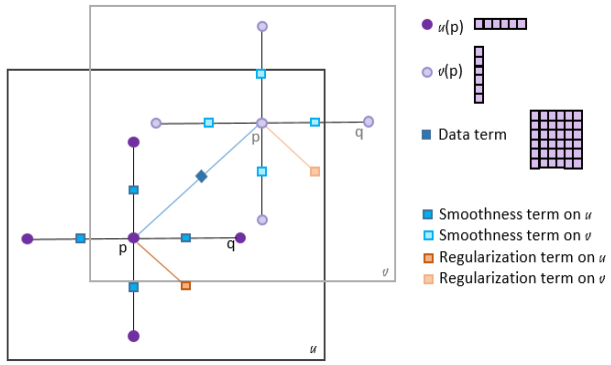


Figure 6: Illustration of Dual-layer belief Propagation [32]. Vertical and horizontal components of objective function

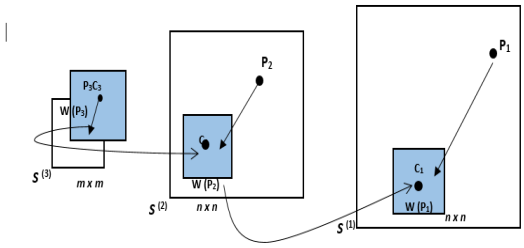


Figure 7: Coarse-to-fine SIFT flow matching process on a pyramid. Blue Square shows the searching window for P_k at pyramid level k.

The foremost aim is to calculate the flow at the coarse level of image patch, constantly resulting and the flow is refined from coarse to fine level [26]. The process is demonstrated in Figure 7. The SIFT images s_1 and s_2 are represented as 's'. A sift pyramid $\{s(k)\}$ is created where $s(1) = s$ and $s(k+1)$ is smoothed and down sampled from $s(k)$. The pyramid level at K, Pk is the coordinate of the pixel to match. Ck is the centroid or offset of the searching window and the right match is $\omega(Pk)$. At the loop pyramid levels $s(2), P3$ is the centroid of the searching window with size mxm. Then the belief propagation converges, at this

level. The system transfers the optimized flow vector $\omega(P3)$ to the next finer level $C2$, where searching window of $P2$ is centered. The process continues from $s(3)$ to $s(1)$ till the flow vector $\omega(P1)$ is obtained. The computational complexity is reduced to a factor of $O(h2logh)$. Thus, the algorithm become very energy efficient by doubling η and retain α and d. Figure 7 shows the process of matching method from coarse to fine SIFT flow.

4. Recognition of hand gestures

4.1. KNN classifier

The SIFT flow vectors generated in the feature extraction stage is used to train the classifiers to assess the performance of the designed system. KNN classifier is selected here, because of its capability of handling large training data. Training was done in the system for two values of k such as 3 and 5. The draw votes in recognition process can be eliminated by selecting odd values of k. 'Leave one group out cross validation' strategy is utilized for testing.

4.2. Support Vector Machine Classifier

The aim of the detector stage is to create a SVM classifier with all data samples of extracted feature vectors. This supervised learning classifier splits the data samples of various categories by calculating a maximum-margin boundary between them. The system is trained using feature vectors of videos from selected database with quadratic, radial, linear and polynomial kernel based SVM classifiers. The classifier was selected because it has been tested successfully in various image classification and object detection tasks. SVM with four kernels is trained with different training videos in hand gesture database. It is seen that radial function based SVM performs better than others. Hence SVM with RBF kernel is used for testing. 'Leave one group out cross-validation' technique is used for testing.

5. Results and Discussion

All test procedures are conducted using Matlab 2014a on a i5 2.7GHz computer with 8.00GB RAM. To analyze the proposed algorithm, we performed the set of experiments with three databases such as hand gesture database for HCI [4,33], Sebastien Marcel Dynamic Hand Posture Database [34] and RMTH German finger spelling database [35]. The performance parameters of the system are analyzed on the basis of certain metrics such as sensitivity, specificity and accuracy [36,37]. The parameters are defined as follows.

$$\text{Sensitivity} = TP/(TP+FN) \quad (6)$$

$$\text{Specificity} = TN/(TN+FP) \quad (7)$$

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (8)$$

$$\text{F1 Score} = 2TP/(2TP+FP+FN) \quad (9)$$

$$\text{Negative predictive value} = TN/(TN+FN) \quad (10)$$

where TN, TP, FN, and FP are true negatives, true positives, false negatives and false positives respectively. A comparative analysis has been done with different state of art methods.

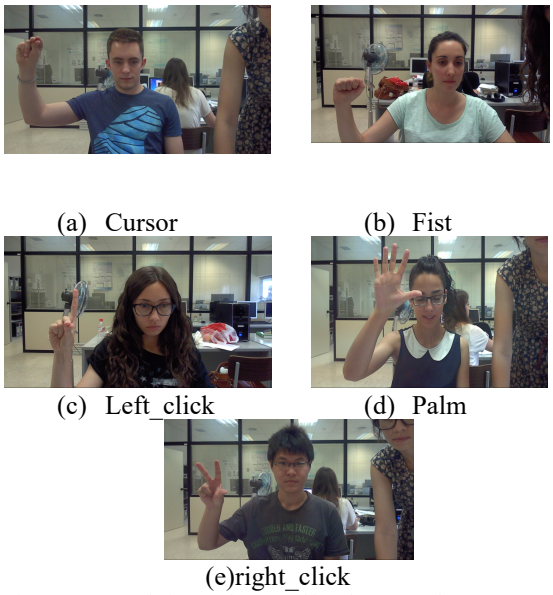
5.1. Performance Analysis

5.1.1. Performance Analysis with Hand gesture Database for HCI

The database contains a group of $1280\,x\,720$ pixel resolution color image frames that are generated for human computer interaction applications using Senz3D sensor [4]. This was generated for testing the mouse functionalities such as left and right click, curser, fist and palm. The samples were captured in realistic way with uneven environment. It composed of two sets of image frames. Five various gestures performed by distinct people are contained in Set 1 and Set 2. For testing, individual videos were created. Samples of hand gesture image frames are shown in Figure 8.

In the first stage, process of hand detection is done and the resolution (width is reduced by a factor of 5 and height is reduced by a factor of 3) of the image frames is reduced. Thus the process will lessen the computational complexity and dimension of the feature vectors and saves the computational time. The feature vectors are generated with the Dense SIFT flow algorithm and are tested with KNN and SVM classifiers.

Evaluation is done using KNN classifier with two values of K=3, K=5. The performance analysis based on recognition rate is provided in the Fig .9. It shows that the highest recognition rate of 87.8% with K = 5 is obtained for the testing video of seq 5 and the average recognition rate is 85.9%.



(a) Cursor (b) Fist

(c) Left_click (d) Palm

(e)right_click

Figure 8: Sample image frames of hand gestures for HCI

The system is tested for using KNN classifier with K =3 and K =5. The performance analysis is provided in Table. 1. The highest training and testing accuracy were achieved for k = 5.

The system is validated using SVM classifier with various kernel functions such as polynomial, linear, radial basis and quadratic in hand gesture database and the performance analysis is shown in Figure 10. The performance level shows SVM classifier with radial basis kernel produces better results than the other kernels and it gives 97.6%.

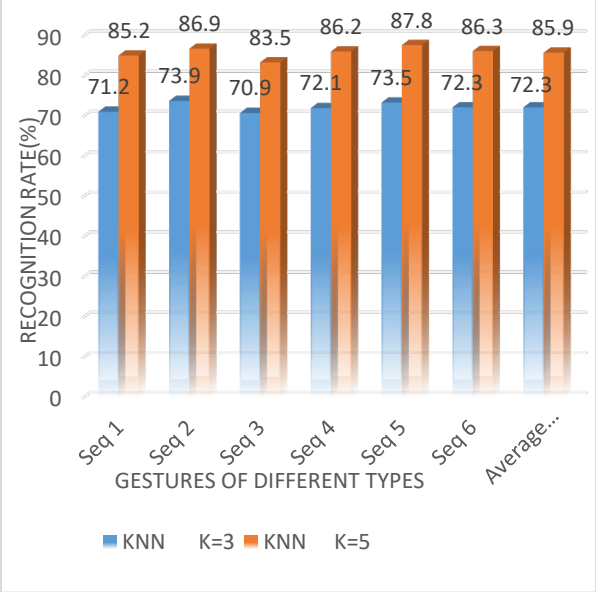


Figure 9: Performance of the system for various hand gestures on hand gesture database using KNN with K=3 and K=5

Table 1: Accuracy of algorithm on hand gesture database using KNN classifier

Expt	K = 3		K = 5	
	Training accuracy	Testing accuracy	Training Accuracy	Testing Accuracy
Subject 1	0.768	0.712	0.873	0.852
Subject 2	0.769	0.739	0.891	0.869
Subject 3	0.748	0.709	0.852	0.835
Subject 4	0.751	0.721	0.878	0.862
Subject 5	0.772	0.735	0.902	0.878
Subject 6	0.758	0.723	0.889	0.863
Mean Accuracy	**0.761**	**0.723**	**0.881**	**0.859**

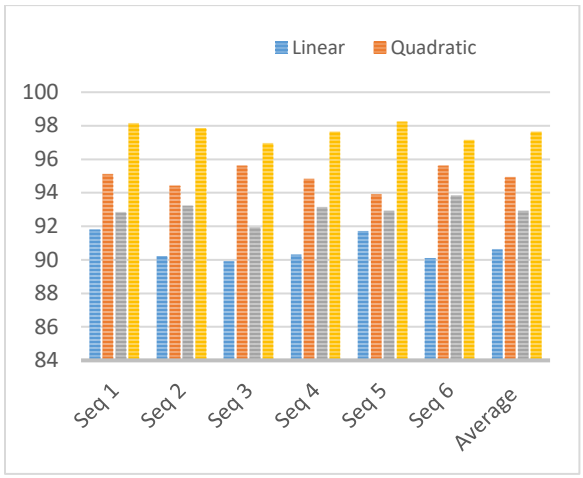


Figure 10: Performance of the system using SVM classifier with different kernels

Group of experiments were done using the multiclass SVM classifier with rbf kernel. Table. 2 shows the evaluation based on accuracy in both training and testing phase. The mean accuracy is obtained as 0.976 which is better as compared with the existing methods.

Table 2: Accuracy of algorithm on hand gesture database using SVM classifier

Gestures	Training accuracy	Testing accuracy
Subject 1	0.998	0.981
Subject 2	0.982	0.978
Subject 3	0.978	0.969
Subject 4	0.989	0.976
Subject 5	0.991	0.982
Subject 6	0.986	0.971
Mean Accuracy	**0.987**	**0.976**

5.1.2. Performance analysis with Sebastien Marcel Dynamic Hand Posture Database

The dynamic hand posture database [34] consists of image frames for four hand gestures such as Rotate, Click, No, and Stop-Grasp-Ok from 10 different persons with various resolution ranges from 50 x 50 pixels to 80 x 80 pixels. Figure 11 shows the Sample image frames.

(a)Clic (b)StopGraspOk (c)No (d)Rotate

Figure 11: Sample frames from Sebastien Marcel Dynamic Hand Posture Database

Table 3 and Table 4 show confusion matrix obtained for the four dynamic hand gestures using KNN classifier with K =3 and K = 5. We observed that the size of hand gestures ranges from 50 × 50 pixels to 80 × 80 pixels, didn't affect the efficiency of the framework in terms of recognition rate and computational time. Some of the frames 'clic' and 'no' hand gestures seem to be same that is why the recognition rate is less for these gestures while using KNN classifier.

Table. 3: Confusion matrix for hand gesture database using KNN with K=3

Gesture	Clic	No	Rotate	StopGraspOk	Recognition Rate(%)
Clic	11	3	0	1	73.3
No	3	10	1	0	71.4
Rotate	1	1	9	2	69.2
StopGraspOk	1	0	3	11	73.3

Table. 4. Confusion matrix for hand gesture database using KNN with K=5

Gesture	Clic	No	Rotate	StopGraspOk	Recognition Rate(%)
Clic	12	3	0	0	80.3
No	2	12	0	0	85.7
Rotate	1	0	11	1	84.6
StopGraspOk	1	0	2	12	80.0

For efficient classification, multi SVM classifier is used and got the recognition rate of 100%. Table. 5 shows performance matrix.

Table. 5: Confusion matrix for hand gesture database using SVM with rbf kernel

Gesture	Clic	No	Rotate	StopGraspOk	Recognition Rate(%)
Clic	15	0	0	0	100
No	0	14	0	0	100
Rotate	0	0	13	0	100
StopGraspOk	0	0	0	15	100

Figure 12 shows the system performance analysis for different types of classifiers based on true positive rate and false positive rate. From the observation, we can say that the SVM classifier with radial basis kernel gives better results. The experiments reveals that the developed framework gives good results for dynamic gestures with different illuminations and resolution levels.

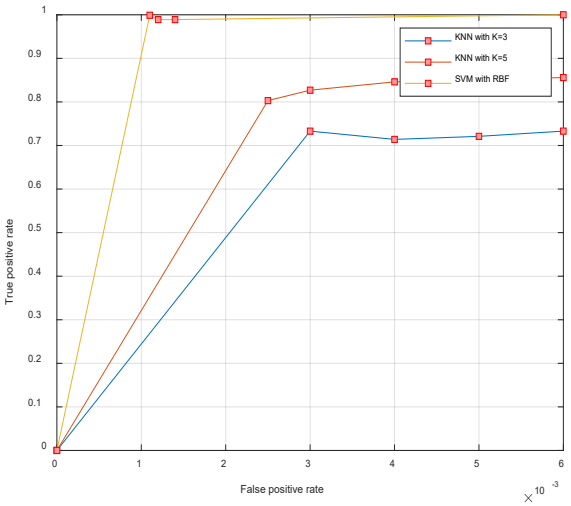


Figure 12: System performance with classifiers

5.1.3. Performance Analysis with RMTH German finger spelling Database

The database [35] consists of gestures of 35 types denoting 1 to 5 numbers, the symbols A to Z, 'SCH and the German umlauts Ä, Ö, Ü. It contains five dynamic gestures captured in uneven environment with varying camera positions. Among 1400 image sequences in the dataset, 700 videos are used for testing. The videos are captured at the rate of 25 frames per second with 320*240 pixels of resolution. The sample image sequences are shown in Figure 13.

Figure 12: Sample frames from RWTH German finger spelling database

The testing and evaluations are performed on five dynamic gestures of RWTH gesture database and confusion matrix obtained is shown in Table 6. The recognition rate achieved is 93% with SVM classifier (rbf kernel). The performance analysis in Figure 13 reveals that the dense SIFT flow-based framework provides better results for gestures from uneven background.

Table 6: Confusion matrix attained for five gestures in RMTH gesture Database

Gesture types	J	Z	Ä	Ö	Ü	Recognition Rate in %
J	18	0	0	2	0	90
Z	0	18	0	0	2	90
Ä	0	1	19	0	0	95
Ö	0	0	1	19	0	95
Ü	0	0	1	0	19	95

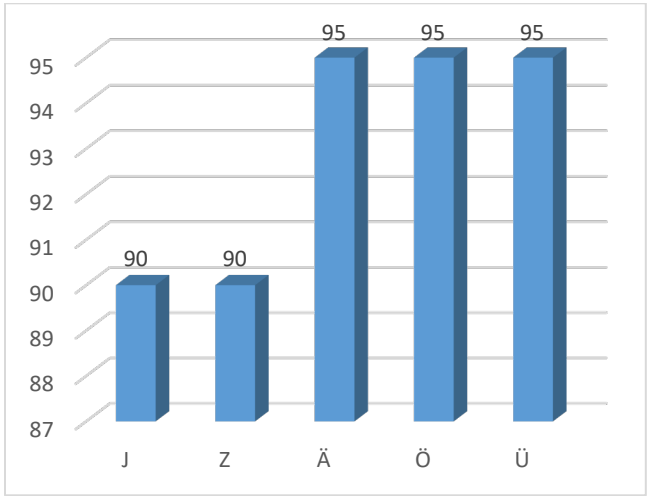


Figure 13: Performance of the system using SVM classifier with rbf kernel

Table 7: Performance Metrics of SVM (rbf) classifier on three different databases

Metric	Hand gesture Database for HCI	Sebastien Marcel Dynamic Hand Posture Database	RMTH German finger spelling Database
Accuracy(%)	97.6	100	93
Sensitivity(%)	97.6	100	93.1
Specificity(%)	98.8	100	94.7
Negative Predictive Value(%)	99.1	100	95.8
F1 Score(%)	97.8	100	92.9
Response time(Sec)	34	28	37

Table 7: Performance comparison of the proposed architecture with the state of the art methods on Hand gesture database

Method	Subject1	Subject2	Subject3	Subject4	Subject5	Subject6	Mean Accuracy
VS-LBP[4]	0.951	0.962	0.939	0.931	0.801	0.959	0.927
LBP+ Moments [7]	0.942	0.918	0.937	0.923	0.851	0.923	0.915
TPM-SLB$_s$P$_{8,1,10,8,1}$ [33]	0.974	0.961	0.976	0.960	0.926	0.975	0.965
TPM-SLBP$_{8,1,10,8,3}$ [33]	0.980	0.957	0.981	0.972	0.924	0.974	0.962
PHOG_TOP+ Optical flow based framework [39]	0.967	0.971	0.942	0.939	0.892	0.962	0.946
Dense SIFT flow based framework	**0.981**	**0.978**	**0.969**	**0.976**	**0.982**	**0.971**	**0.976**

Table 8: Comparative analysis of the proposed method with existing techniques on RMTH German Finger spelling database

Study	Technique	Error rate
Reference [35]	Thresholding with skin color probability+ camshift tracking	35.7%
Reference [38]	features from hand movements and contour shape	27.6%
Reference [39]	PHOG_TOP+ Optical flow based framework	11%
Proposed architecture	Dense SIFT flow	**7%**

From Table. 7, the accuracy of the proposed architecture on the selected datasets come out to be well (97.6% for hand gesture database for HCI, 100% for Sebastien Marcel dynamic hand posture database and 93% for RMTH German finger spelling database). The negative predictive value and specificity records higher values. Higher sensitivity is recorded. Response time of the proposed architecture is very good (< 37). So, the system can work well in real time applications.

5.2. Comparative Analysis

The developed Dense SIFT flow-based framework for hand gesture recognition differs from other methods, in terms of various preprocessing and filtering techniques. Therefore, it is hard to make a quantitative analysis with the state of the art methods. However, we have analyzed performance based on accuracy with existing approaches on the hand gesture database and tabulated on Table 7. **T**he global and local spatial information with temporal data utilizes in volumetric spatiograms of local binary pattern (VS-LBP), achieves the accuracy of 0.927 [4]. The authors of [7] developed an algorithm that fuses directional wavelet Local Binary Pattern (LBP) and moments and obtained the accuracy of 0.914. But it is sensitive to dynamic backgrounds. The Temporal Pyramid Matching of Local Binary Pattern (TPM-LBP) algorithm created by authors of [33] achieved good recognition rate with the expense of computational complexity. In one of recent work [39], hand gesture recognition based on pyramid histogram of gradients (PHOG) and optical flow achieves a recognition rate of 94.6% with the expense of high dimensional feature vectors. From this analysis, we can found that proposed architecture shows considerable improvement in recognition rate and work well in non-uniform background and varying viewpoints.

Table. 8 illustrates the comparative analysis of the proposed architecture with existing techniques on RMTH German finger spelling database based on error rate. The authors of [35] generated a hidden Markov model emission probabilities based on the appearance features to detect the hand gestures and was achieved an error rate of 35.7%. The features from contour shape (orientation, seven hu moments area and perimeter) and motion of hand (angle and velocity of movement) are used to detect hand gestures and achieved a better error rate 27.6% [38]. Another work [39] obtained 11% of error rate by using a multiple feature based framework. The proposed SIFT flow based architecture attained an error rate of 7%. The outcome of the framework is promising in the light of the fact that some of the symbols are very identical in database, for example the gestures for M, N and A.

6. Conclusions

In this paper, we have proposed a novel architecture for hand gesture recognition based on dense SIFT flow. A hand segmentation method is introduced in the pre-processing stage, which utilizes the three frame differencing and skin filtering to reduce the dimensionality of feature vectors. SIFT flow algorithm calculates the displacement between SIFT features of all pixels in

two consecutive sequences in a video. The flow vectors generated are invariant to image rotation and scale, image noise, illumination and camera view point. The KNN classifier and Support vector machine classifier is used for testing effectiveness of features. The present design can further benefit from improvements in SIFT flow estimation and hand segmentation technique. This approach considers the spatial and temporal consistency, and it is capable of accommodating itself to various illuminations. The evaluation results showed that the presented architecture can achieve high accuracies even for non-even background and image noise. As a result of comparison, we determined that our architecture has better recognition accuracy (0.976 for hand gesture database for HCI) than the state of the art methods.

However, a few aspects remain untapped, which may lead to further investigation. Even though our framework provides good results in various levels, it is computationally a little bit expensive. Hence developing faster and more accurate framework will certainly receive more attention in the future. In hand segmentation stage, adding a good key frame extraction technique may well significantly improve both speed and accuracy. And this problem is definitely worth further enhancement.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] T.H. Davenport, "From analytics to artificial intelligence", Journal of Business Analytics, **1**(2), 73 -80, 2018. https://doi.org/10.1080/2573234X.2018.1543535

[2] Y. Liu, P. Zhang, "Vision-Based Human-Computer System Using Hand Gestures", Proceedings of International Conference on Computational Intelligence and Security, Beijing, China, 2009.

[3] J. Zemgulys, V. Raudonis, Rytis Maskeliūnas and Robertas Damaševičius, "Recognition of basketball referee signals from videos using Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM)" Procedia Computer Science **130**, 953–960, 2018.

[4] I. Ana, R Carlos "Human-computer interaction based on visual hand- gesture recognition using volumetric spatiograms of local binary patterns", Computer Vision and Image Understanding, Special Issue on Posture & Gesture, **141**, 126-137, 2015.

[5] M. Azher Uddin, Mostafijur Rahman Akhond and Young-Koo Lee, "Dynamic Scene Recognition Using Spatiotemporal Based DLTP on Spark", IEEE Access, 6, 66123 – 66133, 2018.

[6] X. Fan and T. Tjahjadi, "A spatial – temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences", Pattern Recognition, **48**(11), 3405-3416, 2015. https://doi.org/10.1016/j.patcog.2015.04.025

[7] M.N. AI-Berry, A.M. Mohammed "Fusing directional wavelet local binary pattern and moments for human action recognition", IET ComputerVision, **10**(2), 1-10, 2015.

[8] T. Ojala, M. Pietikainen and Topi Maenpaa, "Multiresolution Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns", IEEE Transactions on Pattern Analysis & Machine Intelligence, **24**(7), 971-987, 2002.

[9] D. Sun, Roth, S. and Black, M.J., "A quantitative analysis of current practices in optical flow estimation and the principles behind them", International Journal of Computer Vision, **106**(2), 115–137, 2014.

[10] J. Walker, "Dense optical flow prediction from a static image", Proceedings of IEEE International Conference on Computer Vision (ICCV), 2443–2451, 2015.

[11] S.-Lara, L., Sun, D., Jampani, V. and Black, M.J., "Optical flow with semantic segmentation and localized layers", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3889–3898, 2016.

[12] S. Shantaiya, "Multiple object tracking using Kalman filter and optical flow", European Journal of Advances in Engineering and Technology, **2**(2), 34–39, 2015.

[13] T. Kroeger et al., "Fast optical flow using dense inverse search", Proceedings of European Conference on Computer Vision, 471–488, 2016.

[14] H.W. Ho et al., "Optical-flow based self-supervised learning of obstacle appearance applied to MAV landing", Robotics and Autonomous Systems, **100**, 78–94, 2018.

[15] P. Liu, Xiangxiang Li, Haiting Cui, Shanshan Li and Yafei Yuan, "Hand Gesture Recognition Based on Single-Shot Multibox Detector Deep Learning", Mobile Information Systems, Article ID 3410348, 2019.

[16] M. Abavisani, Hamid Reza Vaezi Joze and Vishal M. Patel, "Improving the Performance of Unimodal Dynamic Hand-Gesture Recognition with Multimodal Training", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1165-1174,2019.

[17] H. Tang et al., "Fast and Robust Dynamic Hand Gesture Recognition via Key Frames Extraction and Feature Fusion", Neurocomputing, **331**, 424-433, 2019.

[18] Q.D. Smedta et al., "Heterogeneous hand gesture recognition using 3D dynamic skeletal data", Computer Vision and Image Understanding, **181**, 60-72, 2019.

[19] Z. Rasyid et al., "Efficient Hand Gesture Recognition Using Modified Extrusion Method based on Augmented Reality", Test Engineering and Management **83**(5-6), 4020-4027, 2020.

[20] D. G. Lowe, "Distinctive image feautres from scale-invariant keypoints," Int. J. Comput. Vis., **60**(2), 91–110, 2004.

[21] Z. Hossein-Nejad and Mehdi Nasri, "An adaptive image registration method based on SIFT features and RANSAC transform", Computers & Electrical Engineering, **62**(C), 524-537, 2017. https://doi.org/10.1016/j.compeleceng.2016.11.034

[22] W. Zhang et al., "Gender and gaze gesture recognition for human-computer interaction", Computer Vision and Image Understanding, **149**(C), 32-50, 2016. https://doi.org/10.1016/j.cviu.2016.03.014

[23] T. Furuya and R. Ohbuchi, "Dense sampling and fast encoding for 3D model retrieval using bag-of-visual features", Proceedings of the ACM International Conference on Image and Video Retrieval, Article 26., 2009.

[24] Z. Daixian, "SIFT algorithm analysis and optimization," 2010 International Conference on Image Analysis and Signal Processing, Zhejiang, 2010, 415-419, doi: 10.1109/IAS2010.5476084.

[25] D. Simon, " A New Building Material for Artificial General Intelligence", Proceedings of the 2008 conference on Artificial General Intelligence , 414–418, 2008.

[26] C. Liu et al., " SIFT Flow: Dense Correspondence across Scenes and Its Applications", IEEE Transactions on Pattern Analysis and Machine Intelligence, **33**(5), 978 – 994, 2011. DOI: 10.1109/TPAMI.2010.147

[27] J. Singha, Amarjit Roy and Rabul Hussain Laskar, "Dynamic hand gesture recognition using vision-based approach for human–computer interaction", Neural Comput & Applic, **29**, 1129–1141 ,2018.

[28] P. Viola, "Rapid object detection using a boosted cascade of simple features", Proceedings of the IEEE conference on computer vision and pattern recognition, **1**, 511–518, 2001.

[29] P. Viola "Robust real-time face detection", International Journal of Computer Vision, **57**(2), 137 -154, 2004.

[30] D. Chai "Face segmentation using skin-color map in videophone applications", IEEE Trans Circuits and Systems for Video Technology, **9**(2), 551–564, 1999. DOI: 10.1109/76.767122

[31] Y. Liu, "Multi-focus image fusion with dense SIFT", Information Fusion, **23**, 139 -155, 2015.

[32] H. Wang, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition", International Journal of Computer Vision, **103**(1), 60–79, 2013.

[33] I. Ana et al., "Temporal Pyramid Matching of Local Binary Subpatterns for Hand-Gesture Recognition", IEEE Signal Processing Letters, .**23**(8), 1037 -1041, 2016. DOI: 10.1109/LSP.2016.2579664

[34] S. Marcel, O. Bernier, J-E. Viallet and D.Collobert, "Hand gesture recognition using Input/Output Hidden Markov Models", Proceedings of 4th International Conference on Automatic Face and Gesture Recognition (AFGR), 2000.

[35] A. Ahmed et al., "Modeling and Simulation of Office Desk Illumination Using ZEMAX," in 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 1–6, 2019. DOI: 10.1109/ICECCE47252.2019.8940756

[36] P. Baldi et al., "Assessing the accuracy of prediction algorithms for classification: an overview", Bioinformatics Review, **5**(5), 412-424,2000.
[37] T. Raghuveera, R Deepthi, R Mangalashri and R Akshaya. A depth-based Indian Sign Language recognition using Microsoft Kinect. Sadhana; 2020:45:34.
[38] M.S. Abdalla, "Dynamic Hand Gesture Recognition of Arabic Sign Language using Hand Motion Trajectory Features", Global Journals Inc. (USA), 13, 2013.
[39] S. S. Suni and K. Gopakumar, "Fusing Multimodal features for Recognizing Hand Gestures," 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), 1-6, 2019. doi: 10.1109/ICACC2019.8882910.