Journal of Future Robot Life 2 (2021) 1–21
DOI 10.3233/FRL-200010
IOS Press

Extraction of multimodal features from depth and RGB images for recognising hand gestures

S.S. Suni [a,*,**] and K. Gopakumar [b,***]

[a] *LBS Centre for Science and Technology, India*

[b] *TKM College of Engineering Kollam, India*

Abstract. In this study, we propose a multimodal feature based framework for recognising hand gestures from RGB and depth images. In addition to the features from the RGB image, the depth image features are explored into constructing the discriminative feature labels of various gestures. Depth maps having powerful source of information, increases the performance level of various computer vision problems. A newly refined Gradient-Local Binary Pattern (G-LBP) is applied to extract the features from depth images and histogram of gradients (HOG) features are extracted from RGB images. The components from both RGB and depth channels, are concatenated to form a multimodal feature vector. In the final process, classification is performed using K-Nearest Neighbour and multi-class Support Vector Machines. The designed system is invariant to scale, rotation and illumination. The newly developed feature combination method is helpful to achieve superior recognition rates for future innovations.

Keywords: Human-computer interaction, feature extraction, local binary pattern, Histogram of Gradients, support vector machine, K-Nearest Neighbour classifier, hand gesture recognition

1. INTRODUCTION

In the fast growing world of electronics, gesture based technology is an inevitable subject of discussion. The flexibility of hand gives more convenience in expressing inner thoughts and give a way to interaction with the external environment. Efficient and accurate recognition of gesture becomes the subject of challenge as well as a matter of theoretical interest. Hand gestures are congenial in the interaction with computers more realistically. It is used in virtual reality and computer games (Wachs et al., 2011), automobiles and smart devices. Traditional hand gesture recognition systems detect and segment hands based on the approaches including employing colour gloves (Wang and Popović, 2009), and skin colour detection (Bretzner et al., 2002; Argyros and Lourakis, 2004, 2006; Lee et al., 2007; Lee and Hollerer, 2009), both of them have merits and demerits. Another type of hand gesture recognition systems recognise and segment hands using marker-aided methods (Wang and Popović, 2009; Ren et al., 2011). However, these approaches are inconvenient compared with touch less vision based solutions. The issue with regard to gesture interaction is how to create hand gesture knowledge recognised by computers accurately. The multimodal gesture recognition is needed to cater the needs of customers.

However, research efforts continue to cross such limitations. Recent introduction of compact multimodal sensors like the Microsoft Kinect has led the capturing of human movements. Microsoft Kinect

[*] Corresponding author. E-mail: suni.ss@gmail.com.

[**] Research Scholar, Department of Electronics & Communication, LBS Centre for Science and Technology, Thiruvananthapuram, Kerala, India.

[***] Professor, Department of Electronics & Communication, TKM College of Engineering Kollam, Kerala, India

a depth camera, in this regard is a highly successful opening for real world visual applications. Depth camera usage is vital attempt in object recognition problem, a basic issue in computer vision and robotics. The success of visual recognition ever since have been limited to particular cases, such as faces and handwritten digits. The ultra-move in computer vision is large-scale recognition (hundreds of thousands of categories, as in ImageNet (Deng et al., 2009)). For real-world object recognition process, studies have shown that it is beneficial to robustly detect hundreds of household objects (Lai et al., 2011). Kinect has already been utilized to recognize human poses from depth images (Shotton et al., 2011). It is due to the abundant supply of raw and metadata streams by the device, besides the traditional planar RGB video (Kinect, 2016). The progressive improvements like this accelerated the efforts in the field of multimodal gesture interfaces in real-life applications. In this work, we are using RGB and depth images from the Microsoft Kinect sensor to identify the hand gestures.

Numerous methods detect gestures using machine-learning techniques from extracted features of depth data. Zhao et al. introduced interest point based approach to extract features from RGB and depth maps to perform human activity recognition (Zhao et al., 2012). Marin et al. proposed a hand gesture recognition system based on features extracted from leap motion data and depth maps and classified using multiclass support vector machine (Marin et al., 2014). In this work, we improved the method of Kumar Pisharady and Saerbeck (2015); Ullah et al. (2010) to categorize the hand gestures on RGB images and depth maps rather than using only RGB image. Accordingly, appearance or shape features from colour and depth channels are captured and integrated to form one feature vector for each hand gesture. There are two reasons to combine depth information to the RGB features: firstly, the RGB data can be strictly impacted by the variation of lighting condition, alongside with the variation of subjects clothing and viewing angle both may influence the recognition results. Compared with depth data, RGB data contains more intrinsic information for gesture representation. And secondly, because of both physical bodies and movements are presenting as four dimensions in the real world, there will be a loss of depth information. Human activities involve not only spatial-temporal axes but also the depth axis, that are constraints of the 3D scenes and activities which can be directly transposed into image/video contents. Therefore, it becomes inevitable to concentrate on colour and depth data for hand gesture recognition.

Our approach performs the hand gesture recognition in four steps: The First is the manual marking of the interest region in the image. After that, HOG features are extracted from the colour image. The second is to extract the features from depth maps to produce the discriminative feature vector using proposed gradient local binary pattern. The third is the integration of feature vectors from RGB and depth channels, which concatenates all the descriptors extracted from each image into a single descriptor. And finally, in the fourth, we use K-Nearest Neighbour and supervised learning like Support Vector Machine (SVM) for classification of the different gestures.

The paper is organized in the following way: Section 2 gives the related works in the area of multimodal feature based gesture recognition and its inference. Section 3 introduces the general architecture and explains the process of feature extraction from depth and RGB images from Kinect data. Also describes the classification algorithm. Experimental results are in Section 4 and finally Section 5 draws the conclusion.

2. RELATED WORKS

Presently, numerous different techniques are employed in the field of hand gesture recognition and they are surveyed recently in Kumar Pisharady and Saerbeck (2015). These techniques are categorised into five. They are Bag of features/SVM approaches (Ullah et al., 2010), Skelton based approaches (Bogdan Ionescu et al., 2005; De Smedt et al., 2016), texture based approaches (Yang et al., 2016) and multimodal feature based approaches (Marin et al., 2014; Sung et al., 2012). This section mostly concentrates the recent works related to hand gesture recognition using RGB and depth data.

The emergence of Microsoft Kinect devices, depth based gesture recognition has been applauded much due to its wider acceptance from the computer vision community (Uddin and Sarkar, 2014; Sung et al., 2012; Kumar Pisharady and Saerbeck, 2015; Ullah et al., 2010). Sung et al. proposed a hierarchical Maximum Entropy Markov Model (MEMM), in which a person's activity is composed of a group of sub-activities and the two-layered graph structure is inferred by using a dynamic programming method (Sung et al., 2012). The BoFs/SVM techniques are extensively used in activity recognition process because of its effectiveness and easiness (Ullah et al., 2010). Ni et al. suggested a Depth-Layered Multi-Channel STIPs (DLMC-STIPs) framework (Ni et al., 2011), in which STIPs were divided into multiple depth layered channels and resulting STIPs pooling correspondingly within different depth layers. In the end, it gives multiple depth channel histogram representation. Concurrently, Ni et al. introduced a 3D Motion History Images (3D-MHI) using depth information in the same paper. Eigen et al. attempted to estimate depth using multi-scale CNNs (Eigen et al., 2014). It regresses on the depth using CNN with two components: one evaluates the global appearance of the scene while the other one refines the structure using local features. Marin et al. introduced a hand gesture recognition network that takes the data from leap motion sensor and Microsoft Kinect. An ad-hoc feature set based on the positions and orientation of the fingertips is computed and fed into a multi-class SVM classifier in order to recognize gestures (Marin et al., 2014).

The recent popularity of depth based hand gesture recognition is the result of discriminative nature of person independent features and the surveys on hand gestures are detailed in paper Suarez and Murphy (2012). The paper deals with depth-based hand gesture recognition systems, hand position and gesture, recognition methods established and used, the applicability of these techniques and the impacts of the low-cost Kinect and OpenNI software libraries in the concerned area. In 2016, Wu et al. introduced a multimodal hand gesture segmentation and recognition system based on the deep neural networks. A semi-supervised hierarchical dynamic architecture based on a Hidden Markov Model (HMM) is applied for simultaneous gesture segmentation and recognition that utilizes features from skeleton joints, depth and RGB images (Wu et al., 2016). Roccetti et al. (2012) introduced a set of algorithms designed for gesture based interfaces for public spaces. They created the game called Tortellino X-Perience to test and implement the design. Gao et al. (2017) developed a static hand gesture recognition system using parallel CNNs for space robot interactions. Zhu et al. developed 3-D convolution and convolutional long-short term memory (LSTM) based multimodal gesture recognition system for human-robot interactions (Zhu et al., 2017). The method first learns short-term spatiotemporal features of gestures through the 3-D convolutional neural network, and then learns long-term spatiotemporal features by convolutional LSTM networks based on the extracted short-term spatiotemporal features.

In 2018, Junokas et al. came up with a multimodal learning approach through personalised gesture recognition that uses multimodal analytics enabling students to define their physical interactions with computer-assisted learning environments (Junokas et al., 2018). The method uses real time learning analytics and takes three different modes of data such as skeleton positions, kinematic features and internal modal parameters. Even though model gives good results, but the sample size was fairly small and not very diverse.

In 2018, Zhang et al. proposed a new system for recognizing hand gestures based on a combined RGB and depth cameras to improve the fine-grained action descriptors as well as preserve the ability to perform general action recognition (Zhang et al., 2018). The system utilized two interconnected 3D convolutional neural networks to extract the spatio-temporal features from depth and RGB images and detected with Support Vector Machine. Duan et al. introduced a spatial-temporal network architecture based on consensus-voting has been proposed to explicitly model the long-term structure of the video sequence and to reduce estimation variance when confronted with comprehensive inter-class variations (Duan et al., 2018). In addition, a three-dimensional depth-saliency convolutional network is aggregated in parallel to capture subtle motion characteristics. Wei et al build a multi-stream convolutional neural network to improve the accuracy of gesture recognition by learning the correlation between individual muscles and specific gestures with a "divide-and-conquer" strategy (Wei et al., 2019). Cardenas and Chavez integrated multimodal features from Kinect sensor to recognize hand gestures (Escobedo Cardenas and Camara Chavez, 2020). The system extract features from RGB_D data by taking the advantages of convolutional neural networks and histogram of cumulative magnitudes. Deng created a model parameter learning during training to avoid overfilling of gesture labels in the hand gesture recognition process (Deng, 2020). This method is effective to preserve the key information in a small scale and reduce the information loss.

The major constraints from the literature are

- The vital issue performing gesture recognition is in dealing with the matching of position and angular pose.
- The problem of accuracy and efficiency in fingertips shape motion feature extraction should be addressed.
- Limitation of Person dependent descriptors exists.

The solution to the above constraints are the use of depth maps and RGB images (multimodal inputs) to extract the powerful discriminative features. A newly refined descriptor called gradient local binary pattern is utilised to extract the features from depth images, that will help to extract fingertip shape features. HOG features extracted from RGB image to preserve the local appearance features. These extracted feature histograms concatenated together to form the final feature histogram. The main novelties and contributions in this work are

- Framework that incorporates multimodal features from depth and colour images.
- Using Gradient – LBP, we can develop the person independent descriptor.
- Using depth and RGB data, we are able to capture the shape and micro scale movements of fingertips. That helps to address the problem of accuracy in fingertips shape and motion feature extraction.

3. PROPOSED SYSTEM OVERVIEW

The two challenges in the area of hand gesture recognition problems are hand detection which is to detect hand robustly and how to express the features effectively to accurate recognition of gestures. Here we are concentrating on the second stage to extract features productively for getting successful recognition. To gather more discriminative features, we have the colour image data as well as depth maps. Depth images can accumulate inter-object variations and exclude intra-object variations. It is essential for understanding scenes. Figure 1 represents the framework of gesture recognition system with RGB and depth images.

In the design, there are two channels, colour channel and depth channel. From the colour image the HOG features were extracted after manually selecting the hand region. The same process is repeated for depth image with the proposed Gradient Local Binary Pattern (GLBP). The features from two channels concatenated together to form the feature vector of the image. The KNN and SVM classifier is used to identify the gestures.

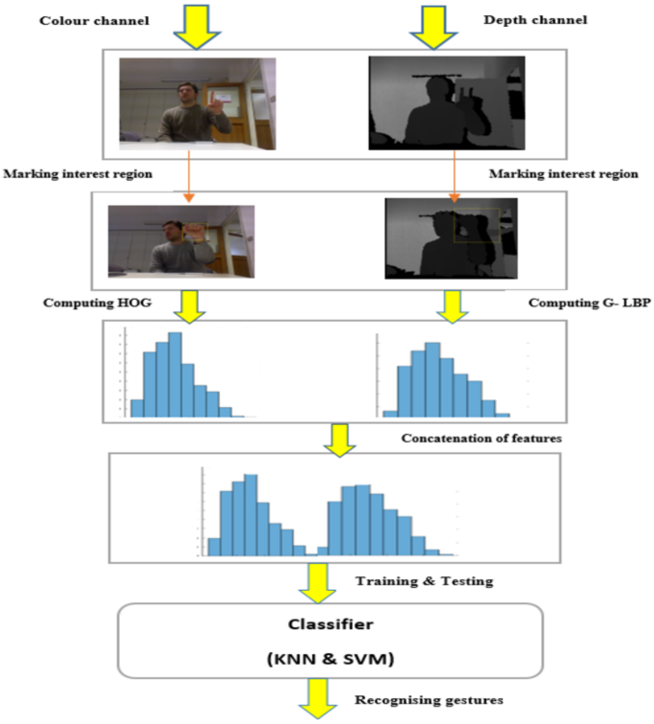


Fig. 1. Proposed framework for gesture recognition system.

3.1. Gradient – LBP features from depth image

In order to improve the drawbacks of traditional gradient and texture based approaches (Yang et al., 2016; Klaser et al., 2008), a newly refined GLBP method is introduced for extracting features from depth maps. In this method of feature extraction, original 256 local binary patterns (LBP) from 16 × 16 image patch are reduced to 56 patterns (Zhao and Pietikainen, 2007). These 56 patterns called uniform patterns are used for creating a 56-bin histogram. The gradient value of each pixel is put as the weight that is always same in LBP based features to computing the values in 56 bins for histogram calculation. In our work, the window size is fixed. The computation of GLBP feature is parallel, that make it easy for hardware realisation. These interesting aspects make GLBP feature possible for real-time hand gesture detection. The flow of GLBP feature extraction is shown in Fig. 2.

After the feature extraction, we get 190 × 395 dimensional feature vector for each depth image. There are mainly four steps to calculate features using Gradient-LBP. In the first step, from input image of sample 640 × 480, manually select the region of interest 200 × 420. The image is divided into blocks of size 16 × 16. We get 395 blocks in this level. The stride between two neighbouring blocks is 8.

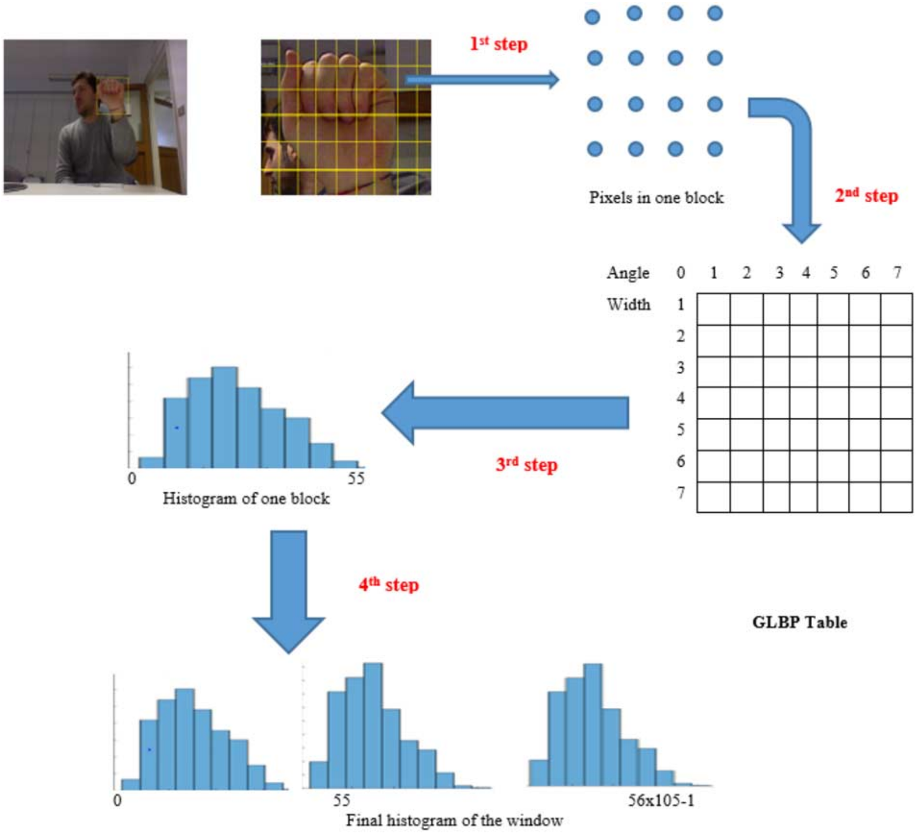


Fig. 2. Flow of GLBP feature extraction.

In the second step, we evaluate the gradient value for each pixel in one block and also find GLBP table position value (width and angle value). The flow of extraction of features in pixel level is shown in Fig. 3.

Here the original pixel value and its eight neighbouring pixels are read first. Calculate the eight bit binary code by comparing the centre pixel value with values of eight neighbouring pixels one by one. The binary value is ‘0’ for the pixel when the centre pixel is greater and ‘1’ when the value of neighbouring pixel is bigger. Then we get one pattern from $2^8 = 256$ pattern for each pixel. Then we put the same value to the neighbour pixel together to obtain several ‘1’ area and ‘0’ area. The ‘0’ area and ‘1’ area appear only once in pattern, that pattern is called as uniform pattern. The binary pattern ‘00111000’ is an example of a uniform pattern. Patterns like ‘00000000’ and ‘01110101’ are non-uniform patterns. From the uniform pattern checking, 56 patterns are identified from total of 256 patterns as uniform patterns. We determine the binary pattern of the pixel if it is uniform. Otherwise, we neglect that pixel and evaluate for the next pixel.

After find out the pixel with uniform pattern, we determine the angle value and width value for the particular pixel. Width value is the number of ‘1’s in the binary code of that pixel. Eight direction codes with 0 to 7 are assigned in the direction of eight neighbouring pixels. The direction code of the centre pixel in ‘1’ area of its binary code is taken as the angle value. If the width value is an even number, then we put the angle value at the smaller value of these two direction values except the middle direction of ‘1’ area is between direction ‘7’ and direction ‘0’. In this case, we kept the angle value at ‘7’. After that, we evaluate the gradient value by the value of original pixel and the values of

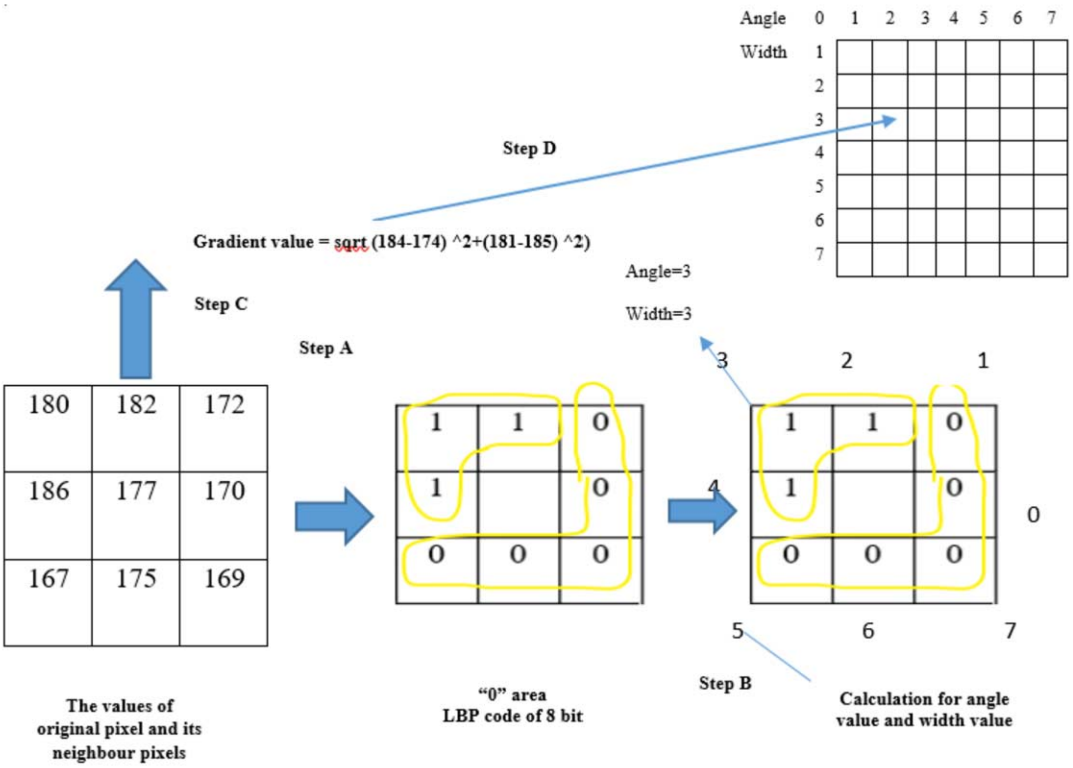


Fig. 3. Pixel level feature extraction process using GLBP method.

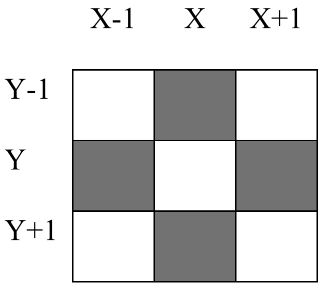


Fig. 4. Feature extraction using gradient technique.

its 4 neighbour pixels by the following formula in Fig. 4.

$$dx = I(x+1, y) - I(x-1, y) \tag{1}$$

$$dy = I(x, y+1) - I(x, y-1) \tag{2}$$

$$m(x, y) = \sqrt{dx^2 + dy^2} \tag{3}$$

Finally, width value and angle value from earlier step are used for mapping the position of bin in GLBP Table. Then we write the gradient value from second step into this bin of the GLBP Table. Last, we get a 56-dimensional vector by GLBP table. Only the value of one element in this 56-dimensional vector is non-zero. The values of other 55 elements are zero.

In third step, we evaluate the 56 bins histogram by adding the 56-dimensional vectors from 16 × 16 pixels by second step in each block. After getting 56 bin histogram for each block, normalization has been done on 56 values in the histogram. The types of normalization methods used here are:

$$L1\ Normalisation: \quad \nu = \frac{\nu}{\|\nu\|_1 + C} \tag{4}$$

$$L2\ Normalisation: \quad \nu = \frac{\nu}{\|\nu\|_2^2 + C^2} \tag{5}$$

In the experiments, we set C to '0'. We trained the classifier with three scenarios, 1.no Normalization 2. L1 normalization 3. L2 normalization. After L1 normalization, the sum of 56 elements is '1' in each block. After L2 normalization, the sum of squares of 56 elements is '1' in each block. The experiment is done in Section 4.3 to find the best normalization type for GLBP method.

In final step, we get a 190 × 395 = 76050 bins histogram by put 395 blocks together. The final 76050 dimensional vector from this 76050 bin histogram is the final Gradient LBP.

3.2. HOG features from RGB image

Histogram of Gradients (HOG) is a widely accepted feature extraction method that is mostly used in computer vision applications for object detection. HOG concentrates on the shape or appearance of the object. This feature evaluates the occurrences of gradient orientation of the localised parts of an image. From the required image we have selected the required portions of 256 × 256 pixels manually. Before taking gradients we are dividing the image into 16 × 16 patches. Each block spitting into 4 cells of size 8 × 8. For each cell, gradient information is evaluated for all 64 pixels using equation (1)–(3) and (6).

$$\theta(x, y) = \tan^{-1}(dy/dx) \tag{6}$$

For every pixel in a cell, we utilize $\theta(x, y)$ to select the histogram and $m(x, y)$ for weight to voting in the histogram. The area from 0 to 2 is divided into 9 histograms. We can get a 9-dimensional vector for each cell after the voting. The total length for HOG feature vector is $9 * 4 * 105$.

3.3. Weighted combination of gradient – LBP and HOG features

Since the contribution of our proposed descriptor for depth images and HOG feature for grayscale images is unbalanced, a weighted combination scheme is applied to concatenate the feature vectors, as grayscale images are more discriminative than depth images. The method is then evaluated with individual and combined feature vectors to analyse the behaviour of the approach on different sources of RGB and Depth data. Support Vector Machines is chosen to perform the classification task considering its higher performance in gesture classification as has been proven in Klaser et al. (2008). More specifically, the classification is first performed separately for grayscale and depth images using SVM, resulting the probabilities of classes belonging to each hand gesture. For the combination of feature vectors, a weighting strategy is introduced.

$$f_{com} = \{\omega_g . f_{GLBP,} \omega_d . f_{HOG}\} \tag{7}$$

ω_g and ω_d are the weighting factor for the grayscale and depth information respectively, they are the free parameters and could be adjusted according to the contribution of each part to the final decision. In our experimentation, we use $\omega_g = 1$ and $\omega_d = 2$ to use these parameters as the resulting accuracy returned by SVM when using each source of information (grayscale or depth) separately for training and validating.

3.4. Recognizing hand gestures

3.4.1. *KNN classifier*

The independent and concatenated feature vectors generated from depth and RGB images in the feature extraction stage is used to train the classifiers to evaluate the performance of the system. KNN classifier is selected here, because it is capable of handling training data that are too large to fit in memory is required. Training was done in the system for different 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 scheme is used.

3.4.2. *Support vector machine classifier*

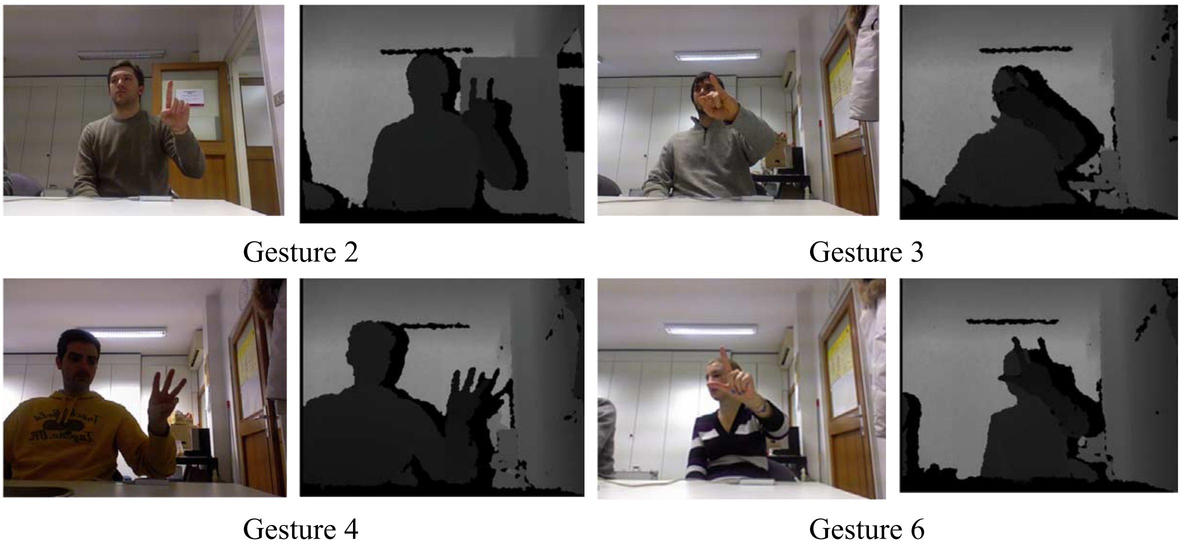
The aim of this phase was to build a SVM classifier on the extracted feature vector of all the data samples. SVM is supervised learning classifier that separates the data samples of classes by computing a maximum-margin boundary between them. The system was trained using feature vectors of images from selected database with different kernel functions: linear, quadratic, polynomial and radial basis function. A support vector classifier was chosen since it has been successfully applied in various object detection and texture classification tasks in computer vision (Suni and Gopakumar, 2020). SVM with four kernels is trained with different training samples of hand gestures from selected database. From the observation, SVM with radial basis function (RBF) kernel performs better than others. Hence SVM with RBF kernel is used in testing phase. For evaluation 'leave one group out' cross- validation scheme is used.

4. EXPERIMENTS AND EVALUATION

4.1. Experiments

Two public datasets are used to evaluate the proposed framework. The Microsoft Kinect hand gesture Dataset of American Sign Language (Marin et al., 2014) is used for experimentation, having colour and depth images. The depth images have the resolution of 640 × 680 pixels, with a frame rate of ranging from 6 to 60. The dataset contains 10 different gestures performed by 14 different people. The person performs the same gesture in 10 different ways. The dataset helps to find out the promising discriminative descriptor to train the classifiers for efficient detection. The sample colour and depth images of gestures are shown in Fig. 5.

The Cornell Activity Dataset (CAD-60) (Sung et al., 2012) comprises RGB-D sequence of different human activities at a frame rate of 15 fps. The dataset consists of RGB, depth, and skeleton data, with 15 joints. Four different persons: two males and two females, one of which is left-handed perform 12 activities in five different environments such as bathroom, bedroom, kitchen, living room and office. The activities are talking on the phone, rinsing mouth, brushing teeth, wearing contact lens, drinking water, opening pill container, cooking-chopping, cooking-stirring, talking on couch, relaxing


Gesture 2

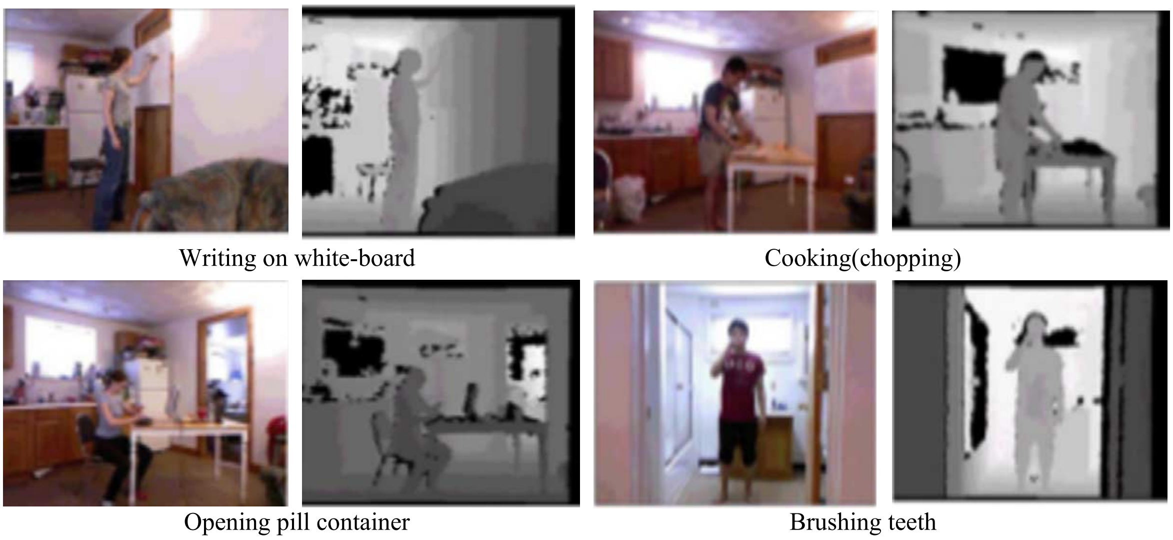
Gesture 3

Gesture 4

Gesture 6

Fig. 5. Sample colour and depth images of hand gesture dataset.

on couch, writing on whiteboard, and working on computer. The sample colour and depth images of activities are shown in Fig. 6.


Writing on white-board

Cooking(chopping)

Opening pill container

Brushing teeth

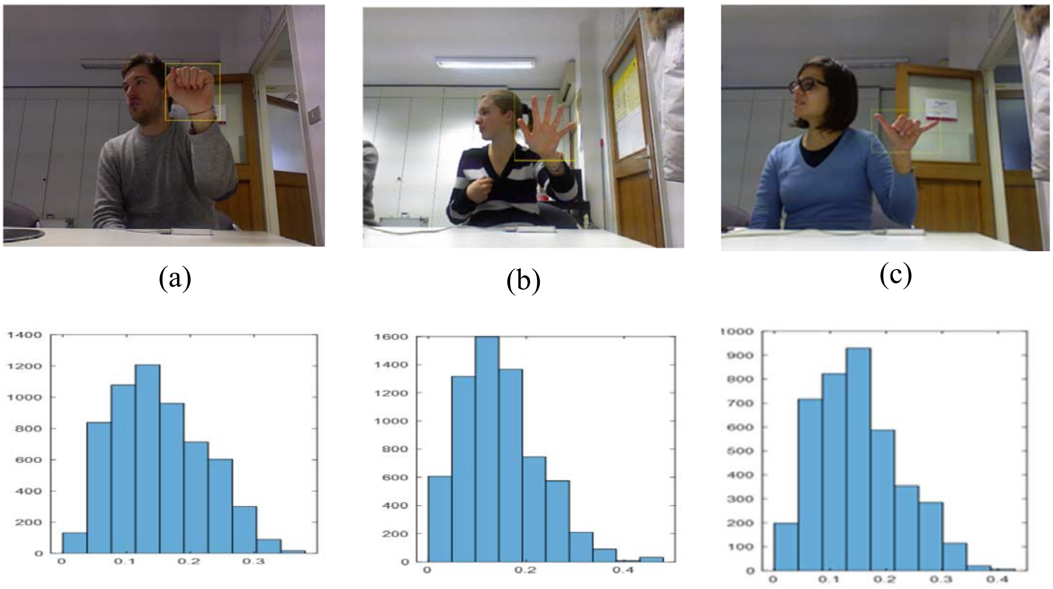
Fig. 6. Sample colour and depth images of CAD-60.

4.2. Settings

All experiments are conducted using Matlab 2014a on a standard i5 2.7 GHz computer with 8.00GB RAM. The colour and depth images are divided into 16 × 16 blocks. The HOG and Gradient-LBP are extracted for each block from RGB and depth images and then concatenated to form a spatially enhanced feature for evaluation. The individual features and combination of feature descriptors are experimented to obtain the in-depth evaluations and performance on classifiers.

4.3. Evaluation

The first set of experiments is done with hand gesture dataset of ASL. We use 14 capture sessions, and 1400 image and depth samples belonging to 10 static gesture categories for evaluating the proposed approach. In our experiments, in order to reduce the dimensions of HOG feature from RGB images we have selected hand region of resolution 256 × 256 pixels. Histogram analysis is performed to identify the variation between different gestures and effectiveness of the features. Figure 7 shows the different gestures and its histogram. From the observation, we can see that histograms are very discriminative in nature for each gesture.



(a) (b) (c)

(d) (e) (f)

Fig. 7. (a), (b) and (c) are the samples of RGB images and (d), (e) and (f) are the corresponding histograms.

From the depth image select the interest region of 200 × 420 pixels and extract the Gradient LBP. After GLBP we get 76050 dimensional feature vector. The experiments are conducted for HOG features from RGB image, GLBP features from depth image and multimodal features from both RGB and depth image. Moreover, we use a multi-class support vector machine and KNN for classification and a leave-one out strategy is used to evaluate the generalization capability of the proposed approach. The performance of the proposed system is evaluated on the basis following statistical indices, including classification accuracy, sensitivity and specificity considering to each image. The parameters are defined as follows.

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

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \tag{9}$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \tag{10}$$

$$\text{False discovery rate} = \text{FP}/(\text{FP} + \text{TP}) \tag{11}$$

$$\text{Negative predictive value} = \text{TN}/(\text{TN} + \text{FN}) \tag{12}$$

$$\text{F1 Score} = 2\text{TP}/(2\text{TP} + \text{FP} + \text{FN}) \tag{13}$$

where TP, TN, FP, and FN are true positives, true negatives, false positives and false negatives respectively.

In GLBP method of feature extraction, the gradient values are used in histogram voting as weights. The voting approach provides the big difference between each block in the image as well as the same block position in two images. Thus normalization technique enhances the values of block when the values are small in the block and reduces the values of block when the values are big in the block. For evaluation, we trained three linear SVM detectors and observed the effects of normalization. Figure 8 shows the results of detectors for different normalization types. From the observation, we can write L2 Normalization $>$ L1 Normalization $>$ No normalization. The experiments reveal that the L2 normalization is the best suit for gradient–LBP.

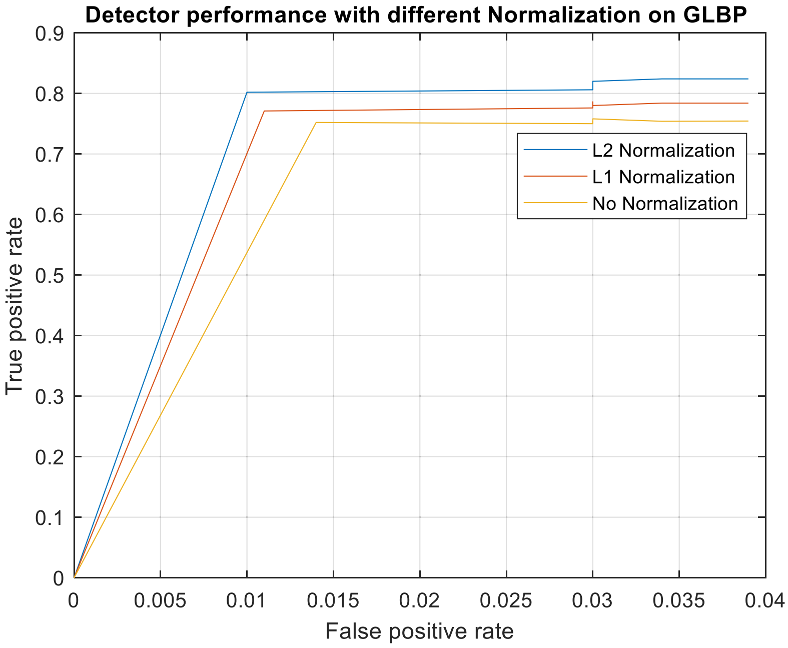


Fig. 8. Detector performance of GLBP with different normalization.

Table 1 reports performance of the proposed multimodal features in terms of accuracy with KNN classifier in different values of $K = 3$ and $K = 5$. KNN classifier with $k = 3$ gives the accuracy of 81.3% and $k = 5$ gives 88.6%. Since, we are getting more accuracy in KNN classifier with $K = 5$, the experiment is further carried out for both in individual and multimodal features for macro analysis (Fig. 9). From the figure, it can be observed that the multimodal features from depth and colour image gives better accuracy than the individual features from both images. The algorithm gets slower when the number of predictors increases. The average execution time is 8.7 miniutes, which is not suitable for real time applications.

Even if we use multimodal features from grayscale and depth images, the classifier limits the performance of the system. So finding out the good classifier increases the detection rate of the system. We trained four SVM detectors with different kernel functions such as linear, polynomial, sigmoid and radial basis. The training results of these detectors are shown in Table 2 and the time consumption result is shown in Fig. 10.

The detector performance gives the inference that polynomial $\approx$ Sigmoid $>$ Radial basis function $>$ Linear. We get the time consumption result as Sigmoid $>$ polynomial $\approx$ RBF $>$ linear.

Table 1
Recognition accuracy comparison with KNN classifier for $K = 3$ and $K = 5$

Gesture Types	**Accuracy ($K = 3$)**		**Accuracy ($K = 5$)**	
	Testing	**Training**	**Testing**	**Training**
G1	0.831	0.823	0.878	0.861
G2	o.834	0.831	0.881	0.878
G3	0.821	0.810	0.901	0.892
G4	0.828	0.828	0.897	0.881
G5	0.803	0.798	0.897	0.893
G6	0.813	0.802	0.891	0.887
G7	0.826	0.813	0.898	0.891
G8	0.815	0.803	0.901	0.897
G9	0.823	0.815	0.896	0.878
G10	0.815	0.809	0.898	0.898
Average accuracy	0.821	0.813	0.894	0.886

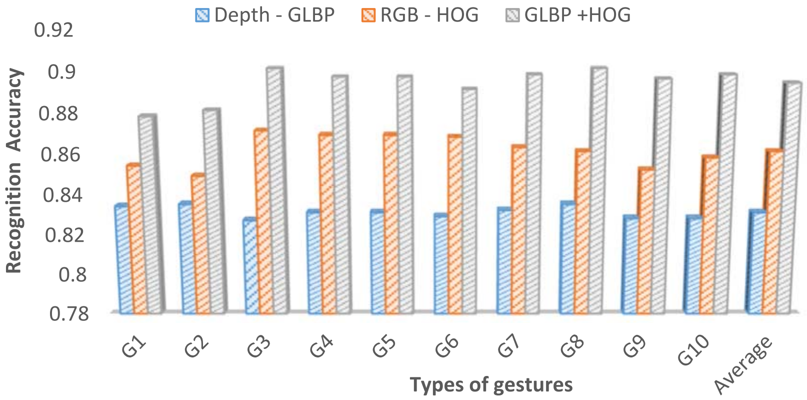


Fig. 9. Recognition accuracy of hand gestures using KNN classifier with $K = 5$.

Table 2
Performance of different kernels with multimodal features from depth and RGB channel on hand gesture dataset

Gesture Types	**SVM–Sigmoid**	**SVM–Polynomial**	**SVM–RBF**	**SVM–Linear**
G1	0.986	0.993	0.976	0.913
G2	0.979	0.993	0.969	0.921
G3	1.00	0.979	0.971	0.913
G4	0.979	0.993	0.966	0.921
G5	0.986	0.986	0.969	0.938
G6	0.979	0.979	0.971	0.921
G7	0.986	1.00	0.963	0.913
G8	0.986	0.971	0.961	0.938
G9	0.979	0.986	0.956	0.921
G10	0.986	0.979	0.963	0.913
Average Accuracy	0.992	0.986	0.966	0.921

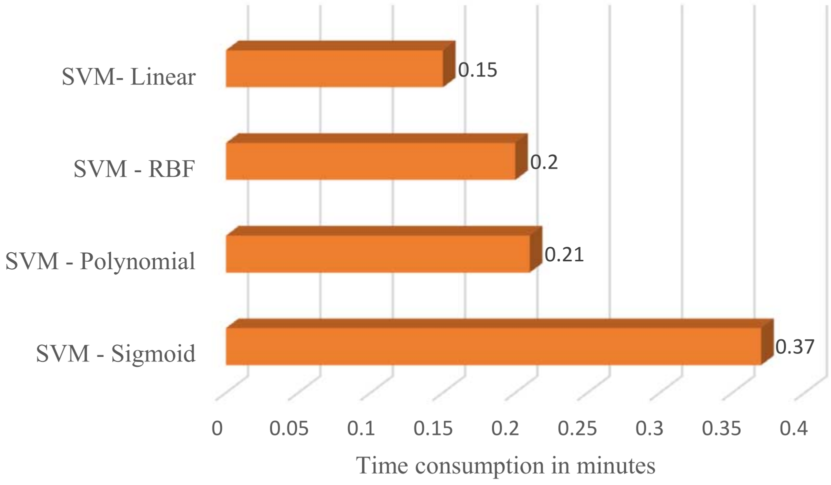


Fig. 10. Detection time of different kernels.

From the results of accuracy and time consumption, we came to conclusion that the SVM with polynomial kernel is suitable for our framework.

Confusion matrix is also used as evaluation measure for a classification problem. Table 3 shows the confusion matrix produced for multimodal feature based framework with multiclass support vector machine with polynomial kernel.

Table 3

Confusion matrix for multimodal feature based framework with SVM on hand gesture dataset

Actual	**Predicted**									
	G1	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**
G1	**139**	0	0	0	1	0	0	0	0	0
G2	0	**139**	1	0	0	0	0	0	0	0
G3	0	1	**137**	0	0	0	0	2	0	0
G4	0	0	0	**139**	0	0	0	0	1	0
G5	1	1	0	0	**138**	0	0	0	0	0
G6	0	0	0	0	0	**137**	0	0	3	0
G7	0	0	0	0	0	0	**140**	0	0	0
G8	0	0	2	0	0	0	0	**136**	0	2
G9	0	0	0	0	0	0	1	0	**138**	1
G10	2	0	1	0	0	0	0	0	0	**137**

Figure 11 shows the performance curve of the multimodal feature based framework with multiclass support vector machine.

In order to reveal the discriminating capability gained by combining RGB and depth image features, we try to evaluate the performance of the individual features and multimodal features. We compare three of them and shown in the Fig. 12. From that, it can observe that RGB-HOG performs better than depth – GLBP. But when we combine the depth – GLBP and RGB HOG features, which gives better performance than the others with average accuracy of 98.6%.

From Table 4, it is evident that the accuracy of the proposed method on the chosen dataset comes out to be 98.6%. The NPV and specificity record higher values, whereas the FDR is quite low. In this

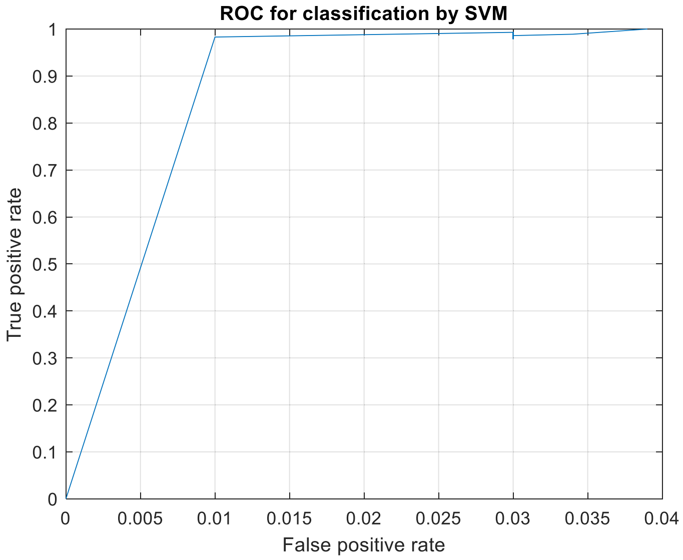


Fig. 11. ROC curve for multimodal feature based framework on hand gesture dataset.

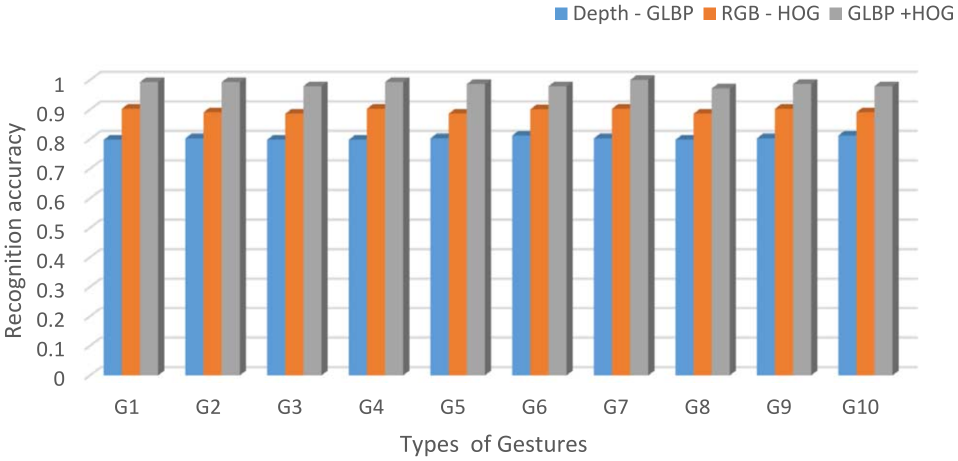


Fig. 12. Recognition rates of hand gestures from individual features and their combination.

Table 4

Performance metrics

Metric	Value
Precision	98.8%
Accuracy	98.6%
Sensitivity	98.6%
Specificity	99.8%
Negative Predictive Value	99.9%
False Discovery Rate	1.2%
F1 Score	98.7%
Response time	0.21 minutes

case, we look for interpretations with minimal false positives. Hence, we are more concerned about higher sensitivity.

The second set of experiments is done on Cornell activity dataset CAD-60 to test performance of proposed algorithm on dynamic actions and cluttered backgrounds. For the qualitative results, we extract the features from RGB and depth images and then use them to recognize the activities with SVM (polynomial kernel). Experiments are performed to identify the activities with individual features from depth and RGB image in order to visualise the contribution of each features. From the confusion matrix (Table 5), we can see that two activities such as writing on white board and wearing contact lens in CAD60 dataset have achieved the highest accuracy of 100%. Moreover, in various scenarios the talking action was predicted as a drinking action and vice versa due to similarity of activities. The average accuracy of recognizing activities achieved 96.2%.

Table 5

Confusion matrix on CAD-60 dataset irrespective of different environments

	Talking on the phone	**Writing on white-board**	**Drinking water**	**Rinsing mouth with water**	**Brushing teeth**	**Wearing contact lenses**	**Talking on couch**	**Relaxing on couch**	**Cooking (chop-ping)**	**Cooking (stir-ring)**	**Opening pill con-tainer**	**Working on com-puter**
Talking on the phone	0.93	0.00	0.06	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
Writing on whiteboard	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Drinking water	0.06	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Rinsing mouth with water	0.00	0.00	0.00	0.97	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00
Brushing teeth	0.00	0.00	0.00	0.02	0.96	0.00	0.01	0.01	0.00	0.00	0.00	0.00
Wearing contact lenses	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Talking on couch	0.01	0.00	0.00	0.00	0.00	0.00	0.96	0.03	0.00	0.00	0.00	0.00
Relaxing on couch	0.01	0.00	0.00	0.00	0.00	0.00	0.04	0.95	0.00	0.00	0.00	0.00
Cooking (chopping)	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.93	0.06	0.00	0.00
Cooking (stirring)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.94	0.00	0.00
Opening pill container	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.98	0.00
Working on computer	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99

The Fig. 13 shows the contribution of HOG features from RGB image, GLBP features from depth image and combination of both. HOG on RGB images are capable of extracting powerful information about a human. However, when looking a new person, changes in clothing and environment can cause confusion especially in uncontrolled and cluttered backgrounds. GLBP feature from depth image extracts magnitude and gradient information effectively, its contribution is notable in the performance of the system. We observed that the RGB feature and depth feature contribution is comparable in the performance of a system on CAD-60 dataset. When we combined both the features the depth information lifts the accuracy of the system. The experiments verified that our framework has strong generalization capacity-it is not only used on gesture identification but also on some other video-based detection tasks.

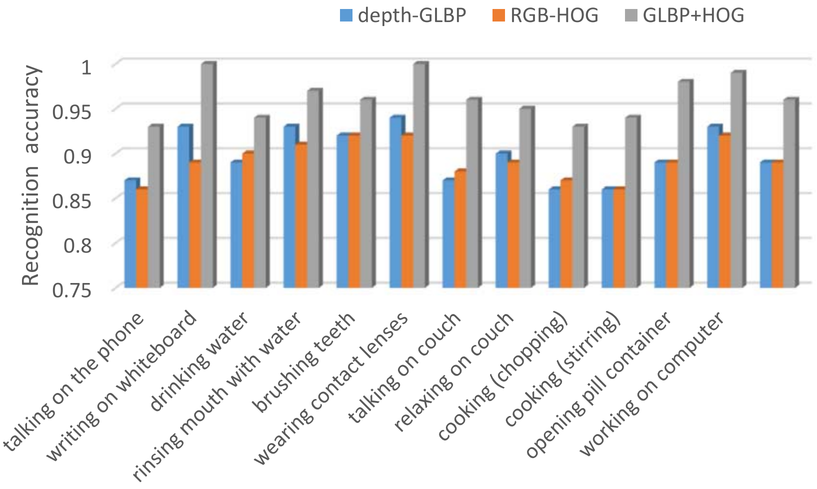


Fig. 13. Accuracy of activities on CAD-60 from depth and RGB features and its combination.

4.4. State-of-the-art comparison

In this section, we compare the proposed multimodal framework performance with the state-of-the-art. Our method outperforms the existing methods on the hand gesture dataset for ASL and CAD-60. The comparison based on recognition accuracy is reports in the Table 6 on hand gesture dataset.

The recognition accuracy of Marin et al. approach is only 89.71% that uses features like curvature and correlation from colour and depth images (Marin et al., 2014). Zhao et al. selected interest points from RGB and depth images and features such as Histogram of Flow (HOF) and Local Depth Pattern(LDP) are used to detect actions. The same is tested for hand gestures and give the recognition accuracy of 90.2% (Zhao et al., 2012). Raguveera et al. introduced an algorithm that uses Speeded Up Robust Features (SURF), Histogram of Gradients (HOG) and Local Binary Pattern (LBP) to convert hand gestures into Indian Sign Language from the depth and colour images captured through Kinect sensor (Raghuveera et al., 34). The system gives the accuracy of 71.85% in response time of 35 seconds and is limited by number of signs. Gangrade et al. applied Oriented FAST and Rotated BRIEF features from depth image to identify hand gestures and achieved the accuracy of 93.26% (Gangrade et al., 2020). The proposed multimodal feature based approach achieves the accuracy of 98.6% in response time of 12.6 seconds (0.21 minutes) and which is better than state of art methods.

Table 6

Performance evaluation of the proposed technique with state of art methods on hand Gesture dataset of ASL

Article	Methods used	Dataset	Accuracy	Precision	F1Score
Marin et al. (2014)	Curvature + Correlation	Hand gesture Dataset of ASL	89.71%	89.9%	90.2%
Zhao et al. (2012)	RGB-IP, RGB-HOGHOF, RGB-IP, Depth-LDP	Hand gesture Dataset of ASL	90.2%	89.2%	90.1%
Raghuveera et al. (34)	SURF+HOG+LBP	Hand gesture Dataset of ISL	71.85%	73.61%	71.45%
Gangrade et al. (2020)	Oriented FAST and Rotated BRIEF	Hand gesture Dataset of ISL	93.26%	94.5%	93.42%
Proposed method	**G-LBP(depth)+ HOG(RGB)**	**Hand gesture Dataset of ASL**	**98.6%**	**98.8%**	**98.7%**

CAD-60 is a relatively good dataset for showing the robustness of our framework. The images are captured in different environments. According to tables results (Table 7), a considerable improvement was obtained with our approach, reaching to 96.2%. The improvement of our algorithm over the other descriptors is good. The reason for this is that our feature is capturing important information in terms of magnitudes and gradients (GLBP). Das et al. applied a skeleton detection method from depth map and RGB image with a fusion of classifiers and achieved the accuracy of 84.37% (Das et al., 2017). Al-Akam et al. extracts Haralick features from gray level co-occurrence matrices to describe the flow pattern of RGB and depth videos to get the 95.45% accuracy (Al-Akam and Paulus, 2018). Generalisation is not done in this method. Arzani et al. formulated human activities using probabilistic graphical models (PGM) and achieved the accuracy of 93.33% (Agahian et al., 2019). Agahian et al. introduced a skeleton based key pose descriptor based framework for recognizing activities and reached the accuracy of 98.5% (Arzani et al., 2020). However, the increase in number of poses increases the noise in the recognition process. Tuning the number of key poses is a concern in the process. From the comparison, we demonstrate that method outperforms state-of-art methods in literature on two well-known gesture recognition and activity datasets.

Table 7

Comparative analysis of the proposed approach with state of art methods on CAD-60

Article	Methods used	Accuracy
Das et al. (2017)	RGB-D and Pose machines skeleton	84.37
Al-Akam et al. (2018)	Dense 3D Optical Flow Co-occurrence Matrix	95.45
Agahian et al. (2019)	Skeleton based key pose descriptor	98.5%
Arzani et al. (2020)	probabilistic graphical models	93.33
Proposed method	**G-LBP(depth)+HOG(RGB)**	**96.2%**

5. CONCLUSION

We developed an effective method for hand gesture recognition from depth and RGB images and showed that the best performance is achieved when we extract the features from both the channels. The proposed Gradient-Local Binary Pattern (G-LBP) approach is used to extract the features from depth images is concatenated with Histogram of Gradients (HOG) features from RGB images to produce the efficient feature vector. By means of extensive evaluation, we demonstrated that the combination of depth and colour image features improves classification accuracy considerably. We further demonstrated that the proposed multimodal feature fusion technique plays an important role in achieving superior performance. The system achieved a recognition rate of 98.6% on hand gesture dataset and 96.2% on Cornell Activity Dataset. Our future work is for investigating the framework for higher level dynamic gestures including activities and motion contexts. Extensive experimental results demonstrate accuracy, efficiency and robustness of our method.

REFERENCES

Agahian, S., Negin, F. & Köse, C. (2019). Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition. *Vis Comput.*, *35*, 591–607. doi:10.1007/s00371-018-1489-7.

Al-Akam, R. & Paulus, D. (2018). Dense 3D optical flow co-occurrence matrices for human activity recognition. In *Proceedings of the 5th International Workshop on Sensor-Based Activity Recognition and Interaction* (Vol. 16, pp. 1–8).

Argyros, A. & Lourakis, M. (2004). Real-time tracking of multiple skin-colored objects with a possibly moving camera. In *European Conference on Computer Vision* (pp. 368–379). Springer.

Argyros, A. & Lourakis, M. (2006). *Vision-Based Interpretation of Hand Gestures for Remote Control of a Computer Mouse. Computer Vision in Human-Computer Interaction* (pp. 40–51). Springer.

Arzani, M.M., Fathy, M., Azirani, A.A., et al. (2020). Skeleton-based structured early activity prediction. *Multimed Tools Appl*. doi:10.1007/s11042-020-08875-w.

Bretzner, L., Laptev, I. & Lindeberg, T. (2002). Handgesture recognition using multi-scale colour features, hierarchical models and particle filtering. In *IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 423–428). doi:10.1109/AFGR.2002.1004190.

Das, S., Koperski, M., Bremond, F. & Francesca, G. (2017). Action recognition based on a mixture of RGB and depth based skeleton. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance* (pp. 1–6).

De Smedt, Q., Wannous, H. & Vandeborre, J.-P. (2016). Skeleton-based dynamic hand gesture recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (pp. 1–9).

Deng, J., Dong, W., Socher, R., Li, L., Li, K. & ImageNet, L.F. (2009). A large-scale hierarchical image database. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Deng, M. (2020). Robust human gesture recognition by leveraging multi-scale feature fusion. *Signal Processing: Image Communication.*, *83*, 115768.

Duan, J., Wan, J., Zhou, S., Guo, X. & Li, S.Z. (2018). A unified framework for multi-modal isolated gesture recognition. *ACM Trans. Multimedia Comput. Commun.*, *2018*, 14. doi:10.1145/3131343.

Eigen, Puhrsch, C. & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *Adv. Neural Inf. Process. Syst.*, 2366–2374.

Escobedo Cardenas, E. & Camara Chavez, G. (2020). Multimodal hand gesture recognition combining temporal and pose information based on CNN descriptors and histogram of cumulative magnitudes. *Journal of Visual Communication and Image Representation.*, *71*, 102772. doi:10.1016/j.jvcir.2020.102772.

Gangrade, J., Bharti, J. & Mulye, A. (2020). Recognition of Indian sign language using ORB with bag of visual words by kinect sensor. *IETE Journal of Research*. doi:10.1080/03772063.2020.1739569.

Gao, Q., Liu, J., Ju, Z., Li, Y., Zhang, T. & Zhang, L. (2017). Static hand gesture recognition with parallel CNNs for space human-robot interaction. In *Intelligent Robotics and Applications*. Lecture Notes in Computer Science (Vol. 10462, pp. 462–473). doi:10.1007/978-3-319-65289-4_44.

Ionescu, B., Coquin, D., Lambert, P. & Buzuloiu, V. (2005). Dynamic hand gesture recognition using the skeleton of the hand. *EURASIP Journal on Advances in Signal Processing.*, *13*, 2101–2109.

Junokas, M.J., Lindgren, R., Kang, J. & Morphew, J.W. (2018). Enhancing multimodal learning through personalized gesture recognition. *Journal of Computer Assisted Learning.*, *34*, 350–357. doi:10.1111/jcal.12262.

Kinect (2016). https://developer.microsoft.com/en-us/windows/kinect.

Klaser, A., Marszalek, M. & Schmid, C. (2008). A spatio – temporal descriptor based on 3D gradients'. In *Proc. 19th Brit. Mach. Vis. Conf.* (pp. 275-1–275-10).

Kumar Pisharady, P. & Saerbeck, M. (2015). Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding.*, *141*, 152–165. doi:10.1016/j.cviu.2015.08.004.

Lai, K., Bo, L., Ren, X. & Fox, D. (2011). A large-scale hierarchical MultiView RGB-D object dataset. In *IEEE International Conference on Robotics and Automation*.

Lee, T. & Hollerer, T. (2009). Multithreaded hybrid feature tracking for markerless augmented reality. *IEEE Transactions on Visualization and Computer Graphics.*, *15*(3), 355–368. doi:10.1109/TVCG.2008.190.

Lee, T., Hollerer, T. & Handy, A.R. (2007). Marker less inspection of augmented reality objects using fingertip tracking. In *IEEE International Symposium on Wearable Computers* (pp. 83–90).

Marin, G., Dominio, F. & Zanuttigh, P. (2014). Hand gesture recognition with leap motion and kinect devices. In *Proceedings of IEEE International Conference on Image Processing, ICIP*.

Microsoft Kinect. http://www.xbox.com/en-us/kinect.

Ni, B., Wang, G. & Moulin, P. (2011). RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *IEEE ICCV Workshops*.

Raghuveera, T., Deepthi, R., Mangalashri, R. & Akshaya, R. (2020). A depth-based Indian sign language recognition using Microsoft kinect. *Sadhana*, *34*, 45.

Ren, Z., Yuan, J. & Zhang, Z. (2011). Robust hand gesture recognition based on finger-Earth mover's distance with a commodity depth camera. In *ACM International Conference on Multimedia ACM* (pp. 1093–1096). doi:10.1145/2072298.2071946.

Roccetti, M., Marfia, G. & Semeraro, A. (2012). Playing into the wild: A gesture-based interface for gaming in public spaces. *Journal of Visual Communication and Image Representation*, *23*(3), 426–440. doi:10.1016/j.jvcir.2011.12.006.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A. & Blake, A. (2011). Real-time human pose recognition in parts from a single depth image. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Suarez, J. & Murphy, R.R. (2012). Hand gesture recognition with depth images: A review. In *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication*.

Sung, J., Ponce, C., Selman, B. & Saxena, A. (2012). Unstructured human activity detection from RGBD images. In *IEEE ICRA*.

Sung, J., Ponce, C., Selman, B. & Saxena, A. (2012). Unstructured human activity detection from RGBD images. In *IEEE International Conference on Robotics and Automation* (pp. 842–849).

Suni, S.S. & Gopakumar, K. (2020). Fusing pyramid histogram of gradients and optical flow for hand gesture recognition". *Int. J. Computational Vision and Robotics.*, *10*(5), 449–464. doi:10.1504/IJCVR.2020.109396.

Uddin, M.Z. & Sarkar, A.M.J. (2014). A facial expression recognition system from depth video. In *Proc. WorldComp* (pp. 1–6).

Ullah, M., Parizi, S. & Laptev, I. (2010). Improving bag-of-features action recognition with non-local cues. In *BMVC*.

Wachs, J.P., Kolsch, M., Stern, H. & Edan, Y. (2011). Vision – based hand gesture appications. *Communications of the ACM.*, *54*, 60–70. doi:10.1145/1897816.1897838.

Wang, R.Y. & Popović, J. (2009). Real-time hand-tracking with a color glove. *ACM Transactions on Graphics, TOG*, *28*(63), 1–8.

Wei, W., Wong, Y., Du, Y., Hu, Y., Kankanhalli, M. & Geng, W. (2019). A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface. *Pattern Recognition Letters.*, *119*, 131–138. doi:10.1016/j.patrec.2017.12.005.

Wu, D., et al. (2016). Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, *38*, 583–1597.

Yang, F., Xia, G.-S., Liu, G., Zhang, L. & Huang, X. (2016). Dynamic texture recognition by aggregating spatial and temporal features via ensemble SVMs. *Neurocomputing.*, *173*, 1310–1321. doi:10.1016/j.neucom.2015.09.004.

Zhang, Z., Tian, Z. & Zhou, M. (2018). HandSense: Smart multimodal hand gesture recognition based on deep neural networks. *J Ambient Intell Human Comput.* doi:10.1007/s12652-018-0989-7.

Zhao, G. & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, *29*(6), 915–928. doi:10.1109/TPAMI.2007.1110.

Zhao, Y., Liu, Z., Yang, L. & Cheng, H. (2012). Combing RGB and Depth Map Features for human activity recognition. In *Proceedings Asia Pacific Signal and Information Processing Association Annual Summit and Conference*.

Zhu, G., Zhang, L., Shen, P. & Song, J. (2017). Multimodal gesture recognition using 3-D convolution and convolutional LSTM. *IEEE Access*, *5*, 4517–4524. doi:10.1109/ACCESS.2017.2684186.