International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

Hand Gesture Recognition using fusion of SIFT and HoG with SVM as a Classifier

Farah Jamal Ansari
University Polytechnic, Faculty of Engineering & Technology,
Jamia Millia Islamia, New Delhi, India

ABSTRACT-*This paper focuses on the hand gesture recognition using the various feature extraction techniques and SVM as a classifier. Her we have proposed the hybrid approach using SIFT and HoG combined as a feature extraction technique and gestures classification done using SVM linear kernel function.The accumulative multi class SVM method is employed in order to obtain a classification of the multiple gestures. In this computer age the hand gesture recognition is one of the important domain of the computer application wherein the human computer interaction is done without any contact. Various research are ongoing in order to produce the cost effective and robust system design in this field. We have also proposed our model with max 97% accuracy with 10 set of gesture.*

Keyword: SVM, HoG, SIFT, Hand Gesture recognition, Gesture, HCI

INTRODUCTION

Gestures are powerful means of communication among humans. Among different modality of thebody, thehand gesture is the most simple and natural way of communication mode. Real-time, vision-based hand gesture recognition is more feasible due to the latest advances in the field of computer vision, image processing, and pattern recognition but it has yet, to be fully explored for Human-Computer Interaction (HCI). Our system receives input from webcam images and classifies them based on features denoted by post-processing the images. For the purpose of data classification, we used two methods HOG and SIFT which extracted the feature vectors and further these feature vectors were used for machine learningwhere we used SVM for the same. Our algorithmfirst applies image processing techniques to theimages in order to cancel background and noiseeffects.It then extracts relevant features for classificationsuch as area, orientation, and normalized image pixelsdata and finally classifies the gesture features using amulticlass Support Vector Machine classifier.

LITERATURE REVIEW

In [1], the author included the algorithm in which first the video was captured and then divided into various frames and the frame with the image was extracted and further from that frames various features like Difference of Gaussian (DOG) was calculated.

In [2], the author applied a combination of (2D) shape-based and size-based features to recognize the configuration of the hand in the scene. In [3], the author gathered the tracking data over time with a 3D-camera, and the author devises an SVM-based system to classify between one finger, two fingers gestures In [4], the authors implemented Hidden Markov models, commonly used in handwriting recognition, achieve a 97%+ accuracy in classifying 40 words in American Sign Language.

In [5], a method had been developed in which gesture recognition model was developed using SVM and better outputs were obtained using Cross-Validation. In [6], the authors have classified gestures through Multiclass SVM using libSVM.

Current solutions incorporate a variety of machine learning techniques to classify hands.

In [7], the author had used a Pyramid of Histogram of Oriented Gradients as a feature for an SVM with 70- 30 cross validation; the author was able to distinguish between a hand and non-hand.

PROPOSED METHODOLOGY

Generally, vision-based hand gesture recognition consists of three basic processing stages

A. Database Creation

B. Hand Part Segmentation.

C. Gesture Feature Extraction.

D. Gesture Classification.

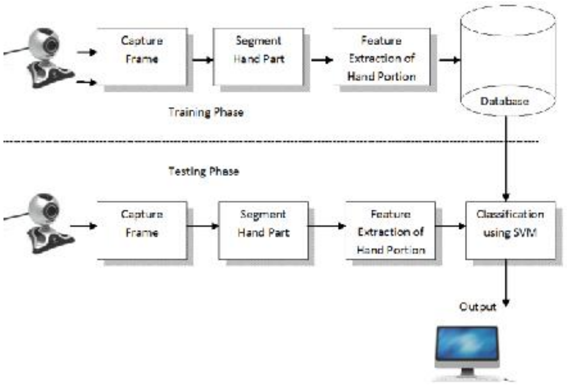


Fig 1: Block Diagram of Methodology

The above figure shows these stages in two phases namely, the Training phase and the Testing phase. The Training phase includes only the first two processing stages, whereas the Testing phase includes all three processing stages

DATABASE CREATION

Our database matches over 3 variable folders having agroup of hand gestures of different orientations for durable categorisation. The testing pitchers were taken from a normal laptop webcam at 320 x 240pix resolutions.

While takingpitchers, the hand gestures were moved slowly to avoid taking collections of pictures that were too match. By introducing fewvariance into the data collection, we satisfy that any new data we check does not need to look exactly same the training data in order to be correctly classify.

HAND PARTSEGMENTION

The main procedure of hand segmentation is to detect hand regions in the image captured of hand gesture and separate them from backgrounds, as shown inFig.1. It should be mentioned that the correctness of hand gesture recognition has a close relationship to the accuracy of hand segmentation. Many conventional methods of hand segmentation take advantages of color cues.

However, the accuracy of hand gesture segmentation tends to be affected easily by several factors such as the skin color differences between humans, the sensitivity of color to illumination, and especially the situation due to thepresence of objects with similar skin color. Here we have utilized color detection scheme where only a specific color would be detected.

GESTUREFEATURE EXTRACTION

The final stage is hand gesture recognition in whichthe output of current gesture model from the secondstage is compared with each model in hand gesturedatabase where the most matched hand gesture isselected as thefinal recognition result. This has beenillustrated in the testing phase of Fig1. Differenthand gesture modeling methods have diverserecognition approaches. The hand gesture isrecognized by counting the number of active fingers.

The hand gesture is modeled as the star skeleton,and the recognition is performed by distancesignature. Other features such as hand position anddirection, finger position and direction, and thedistancebetween the fingers are always used for establishing aspatial model of hand gesture.

SIFT

Scale-invariant feature transform (SIFT) is an algorithm in data processes eye to find out and explain local characteristic in the picture. The algorithm rule revealed by David Lowe in 1999. SIFT keynote points of objects are first extracted from a pair of the reference picture and collected in a database management system.

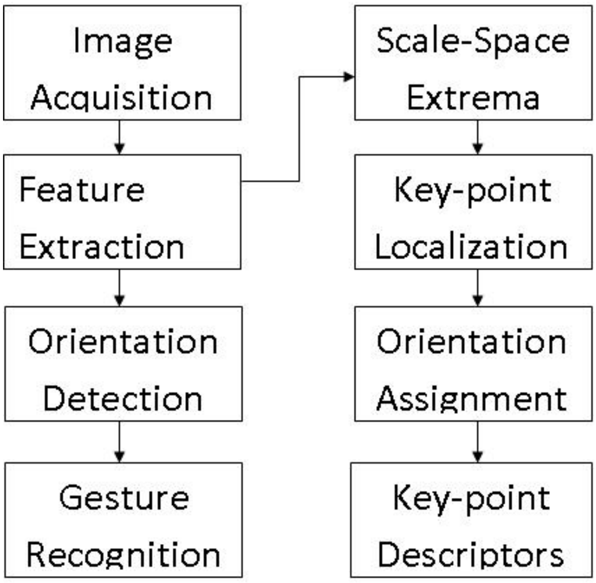


Figure 2: Flowchart of SIFT Technique

A.IMAGE ACQUISTION:

The first footprint of picture Acquisition as the name recommended is of encapsulating the picture in a period of runtime through a webcam. It is actually the formation of digital picture, typically by a physical means.

B.FEATURE EXTRACTION:

For any hand, there are many characteristics similar, the exciting instant that can be filtered to provide a "uniquecharacteristics" differentiation of the object. SIFT picture quality provide a pair of the characteristicona hand that is not affected by the complications experienced in additional methods, such as object scaling and revolution[4].

C.SCALE SPACE EXTREMA DETECTION:

This level of scale-space extrema recognition attempts to identify those scales and locations that are identifiable from distinct corner sight of the similar hand gestures [4].

D. KEYPOINT LOCALISATION:

Scale-space extrema recognition generate too many keynote point candidates, some of which are unstable. The next step in the algorithm is to perform a detailed fit to the nearby data for accurate location, scale, and the ratio of principal curvatures [4]

E.ORIENTATION DETECTION:

Histogram of Oriented Gradients (HOG) are characteristic descriptors used in electronics computer sight and pictureprocessing for the suggest of object recognition [4].The technique counts occurrences of ramp orientation in localized segment of a picture.

F.ORIENTATION ASSIGNMENT:

In this point, each keynote point is assigned 1 or more orientations based on local picture gradient directions. This is the key level in gaining invariance to therotation as the keynote pointexplanation can be represented relative to this direction and therefore gains invariance to picture movement.

G. KEYPOINT DISCRIPTOR:

To calculate a descriptor vector for singlekeynotepoint such that the descriptor is maximumunique and moderately invariant to the remaining changes such as illumination, 3D viewpoint.

H. GESTURE RECOGNITION:

Gesture recognition initiates humans to communicate with the machine (HMI) and interact simply without any mechanical devices.This is the last part of the algorithm where the arrow shown in front of thecamera will be changed to the relating text.

HISTOGRAM OF GRADIENTS (HoG)

HOG ischaracteristicdescriptors access in picture processingfor the motive of object recognition. The techniquecounts theoccurrence of gradient direction inlocalized section of a picture.

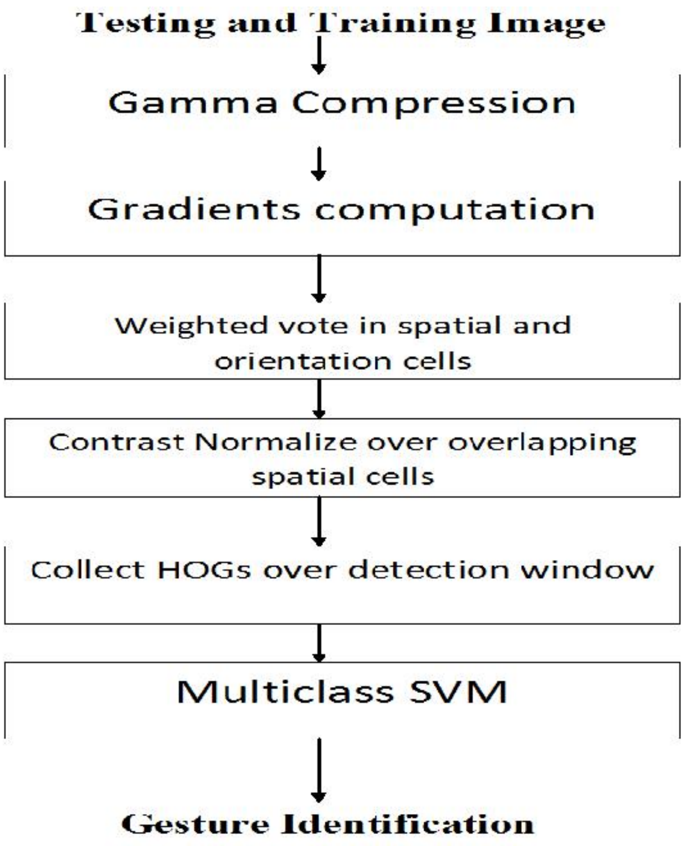

Testing and Training Image
Gamma Compression
Gradients computation
Weighted vote in spatial and orientation cells
Contrast Normalize over overlapping spatial cells
Collect HOGs over detection window
Multiclass SVM
Gesture Identification

Figure 3: Flowchart of Histogram of Oriented Gradients

A**. Test/Training Image:** Taking the pictures is the first level. Here the size is taken to be 320×240 pixels

B**.Gamma Normalization:** Gamma Normalization is a secondary method, whereeither square root of the picture logarithm of the similar can be taken.

C**.Calculating the Gradients:**Each gradient has its own magnitude value as well asdirection.

D**.Weighted vote in spatial and direction cell:**We produce a window and then dividesinto a huge grid.After divinginto a dense grid which holds ofcells, single cell gradient is considered in order tocalculate its magnitude and the w.r.t directionand histogram of its orientation is taken which isweighted by themagnitude value of thegradient. Thus, in shorthistogram of gradient direction is taken [7].

E**. Contrast Normalizing over overlapping spatialcells:** We bind many of these so-invite cells inthe neighborhood so as to produce a block. Blockallows normalizing the picture or the block per say.This gives contrast normalization.This contrast normalizations aids to get rid of variationin strength. Here we can even have differentoverlapping blocks [7].

F**.A collection of HoG's:**HOG's of these blocks is gathered in a big array of thecharacteristic vector. Further, these characteristic vectors are givento learning algorithm, in this case, SVM. Also, heresince a gesture is assumed, the positive images are labeled as 1 while the negative pictures are assignedthe value -1

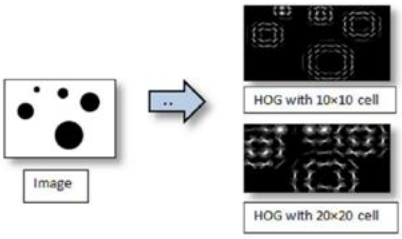


Figure 4: Visualization of HOG

D. GESTURE CLASSIFICATION

For the Classification SVM (support vector machine) is taking into consideration with RBF kernel function. The total of 10 gestures with 10 samples of each gesture is taken. Out of which 70 percent is considered for training and rest 30 percent is for testing.

SUPPORT VECTOR MACHINE:

The Supported Vector Machine (SVM) Classifier is largely used for classification and regression testing. SVM training algorithm makes a model that assumeswhether a new example deep into single category or other. The SVM classifier learns from the data points in examples when they are classified belonging to their respective categories. The SVM [2] is a very well-known learning algorithm for classification problem. The main goal of SVM is to design an optimal separating hyperplane such that it can classify training vectors into classes.

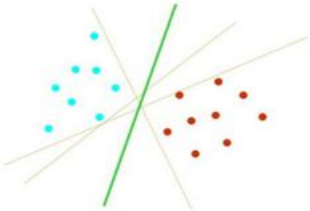


Figure 5: SVM Hyperplane

SVM then separates a given binary labeled training data with the hyperplane that is maximally distant from them also known as the maximal margin hyperplane. For cases in which no linear separation is possible, they can work in combination with the technique of kernels, which automatically realizes a non-linear mapping to a feature space.

Consider the problem of separating the set of training vector belonging to two separate classes,1,2x2………Xn Which are vectors in D.

We consider a decision function of the following form [6]

yx =wTΦx+b ……………………..(1)

Attached to each observation 'xi' is a class label,

$$t_i \in \{-1, +1\}.$$

Without loss of generality, we must build a decision function such that, y xi>0. The idea is to extend it to multi-class problems is to decompose an M-class problem into a series of thetwo-class problem. Let 'yjx' be the decision function, with the maximum margin that separates class I from the remaining classes,

$$\text{yjx} = \text{wjT}\Phi\text{x} + \text{bi} \quad \ldots\ldots\ldots\ldots\ldots.(2)$$

Here, 'Wj' is ann-dimensional vector, 'x' is mapping function which maps x into n-dimensional space. The problem can be equivalently understood in terms of projecting the input data into a higher dimensional space where they are separated using parallel hyperplanes. Now, if the classification problem is separable, the training data belonging to class 'k' satisfy yk x=0 and data belonging to other classes must satisfy yk x<1 and other data belonging to other classes satisfy yk x>-1.

An n-dimensional pattern (object)x has n coordinates, =$(x_1, x_2, \ldots\ldots\ldots\ldots\ldots, x_n)$, where each x_i is a real number, $y_j \in \{-1, +1\}$. Consider a training set T of m patterns together with their classes, $T = \{(x_1, x_2,), (x_2, y_2), \ldots\ldots\ldots\ldots\ldots, (x_m, y_m)\}$. Consider a dot product space S, in which the patterns x are embedded, $x_1, x_2, \ldots\ldots\ldots, x_m)$ ƐS. Any hyperplane in the space S can be written as [1]

$$\{x \in S | w.x + b = 0\}, w \in S, b \in R \ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

The dot product w.x is defined by:

$$W.x = \sum_{i=1}^{n} w_i x_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

A training set of patterns is linearly separable if there exist at least one linear classifier defined by the pair (w, b) which correctly classifies all training patterns. This linear classifier is represented by the hyperplane H (w.+b=0) and defines a region for class +1 patterns (w.+b>0) and another region for class -1 patterns (w.+b<0) [3].

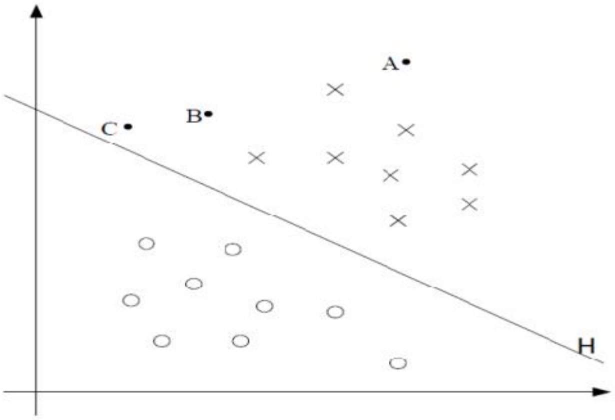


Figure 6: Linear classifier defined by the hyperplane H.

After training, the classifier is ready to predict the class membership for new patterns, different from those used in training. The class of a pattern Xk is determined with the equation[2]

$$c_{\text{l}} \quad (x_k) = \begin{cases} +1 \; i \;\; w.x_k + b > 0 \\ -1 \; i \;\; w.x_k + b < 0 \end{cases} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.. (5)$$

Therefore, the classification of new patterns depends only on the sign of the expression w.x+b.

MULTI CLASS SVM:

Multiclass SVM is useful for multiple gesture classification where a number of testing gestures is more than two.We implemented multiclass classification by using the library mentioned in libSVM supports one vs one and one vs. all types of classification. We used the technique of one vs. all method in which we take the training samples with the same label as one class and the others as the other class, and then it becomes a two-class problem[2]. libSVM supports various kernels which are required for obtaining the best model for optimum classification. RBF kernel was implemented as it provides better results. This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear [2]. Cross validation was implemented for obtaining best values of C, γ which provides the best separating hyperplane The goal is to identify good (C, γ) so that the classifier can accurately predict unknown data (i.e. testing data)[5]. In v-fold cross-validation, we first divide the training set into v subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining v-1 subsets. We implemented v=10 for parameter selection.

IMPLEMENTED ALGORITHM:

Implementing the techniques in machine learning and image processing, we hope to obtain a high level of accuracy in distinguishing between the letters in the Indian sign language alphabet.

1) Gesture Segmentation: In order to distinguish hand gesture from background image we use theblue glove to separate the gesture from thebackground by implementing colour based object detection. Steps:-

a) Input testing images were restricted to a 320x240 pixel size further

b) The blue plane of the 3D image is subtracted from the grayscale of the original image. Later median filtering is performed to reduce the salt and pepper noise while retaining the edges.

FEATURE VECTOR:

As mentioned earlier HOG as well as SIFT Algorithms were implemented to calculate feature vectors. Feature vectors provide the behavior of all the pixels in the form of vector. For HOG parameters included in the code are Cell size=8, block size=2, orientation bins=9, oriented gradients=1 feature length of fixed dimension were obtained for further calculation. These vectors were saved in a mat. File. For comparison of output, another feature vector SIFT was implemented.These two feature vectors were used to classify separately and the efficiency was tested. Similarly, feature vectors were calculated for different gestures we used the database available on Cambridge Hand Gesture Dataset to train the model for classification. Training images were segmented into different files and the feature vectors classified were stored in a matrix file with proper labelling according to the gestures.

SVM:

libSVM was used to classify the different gestures based on their feature vectors. The multiclass classification was implemented using RBF kernel. After cross-validation was performed, optimum parameters were obtained for C and γ. One vs all multiclassmethod of gesture classification used [1]. This algorithm was implemented on Matlab running on Windows platform based on Intel core i7 processor.

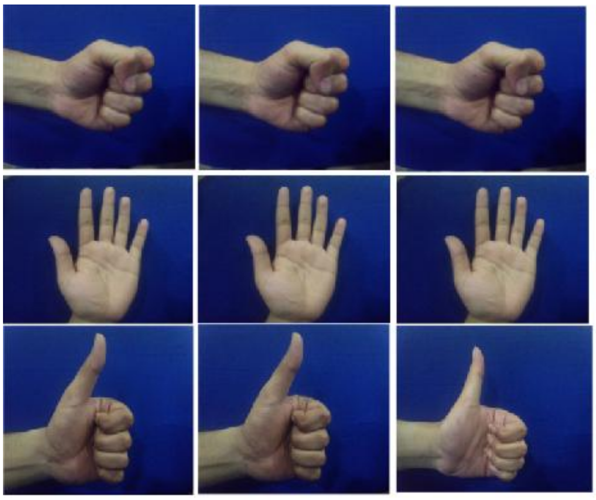


Figure 8:Custom SampleHand gesture Data Set

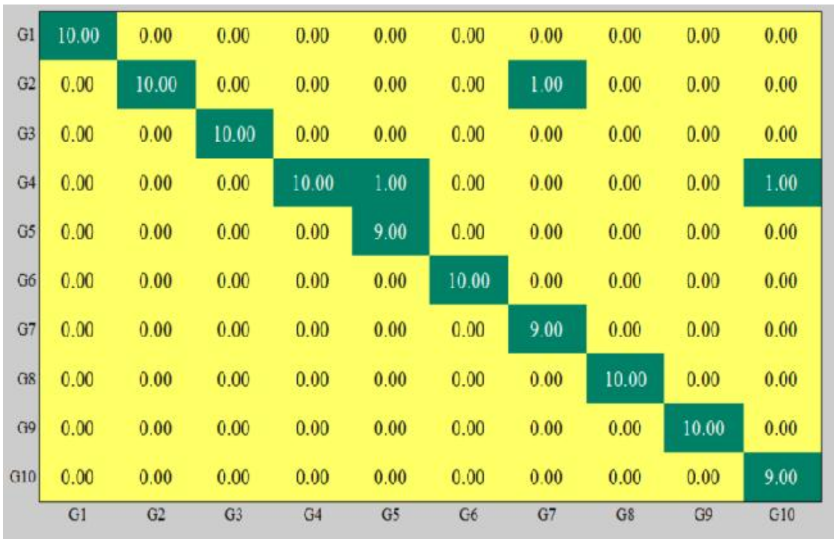
RESULTS

The experimentation is done on 10 set of different gestures. The feature extraction is done using SIFT algorithm and the HoG algorithm and then it is further classified using the support vector machine multiclass algorithm. The results are obtained by changing the ratios of the total training and the testing samples. Ist stage we considered 70% training set and 30% testing set. IInd stage we considered is 50% training and 50 percent testing. IIIrd stage 30% training and 70% testing.

Table 1: Comparative results on different training and testing sets ratios.

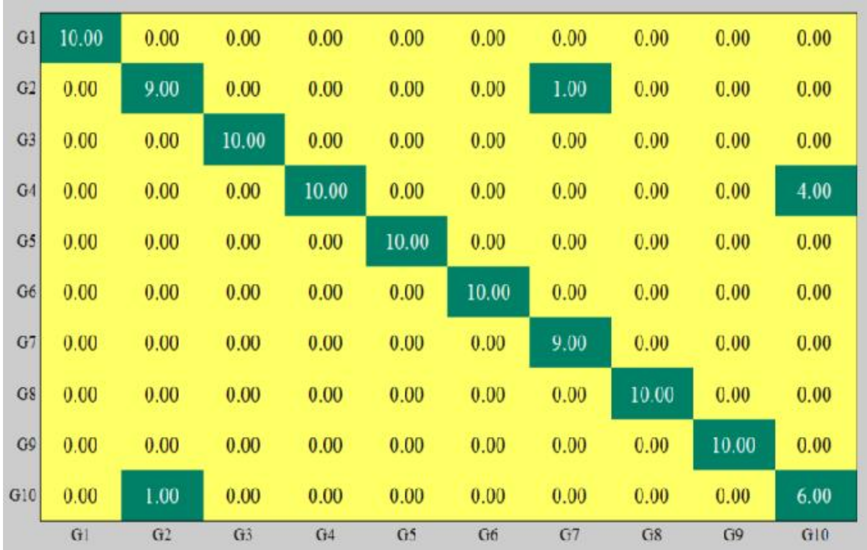
Method	Training/Testing Ratio	Testing	Recognition Rate
SIFT + HoG + SVM	70	30	97%
SIFT + HoG + SVM	50	50	94%
SVM + HoG	30	70	86%

Table 1 illustrates the comparative analysis of the gesture recognition of all the three ratios. From the table we can figure out that the more is the training dataset the better is the accuracy. The 70:30 ratio yields the 97% accuracy on the 10 gestures where as in the 30:70 ratio it is reduced and degrades to 86% recognition.



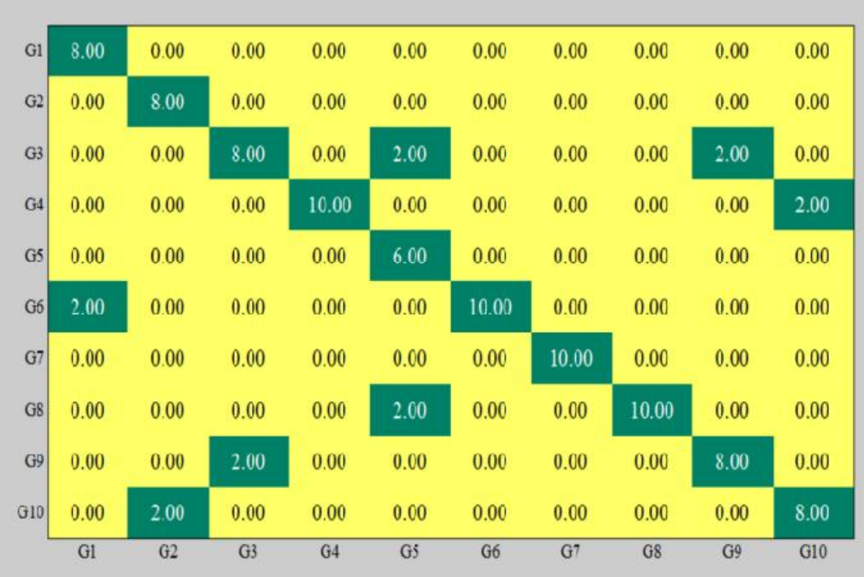
G1	10.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G2	0.00	10.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
G3	0.00	0.00	10.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G4	0.00	0.00	0.00	10.00	1.00	0.00	0.00	0.00	0.00	1.00
G5	0.00	0.00	0.00	0.00	9.00	0.00	0.00	0.00	0.00	0.00
G6	0.00	0.00	0.00	0.00	0.00	10.00	0.00	0.00	0.00	0.00
G7	0.00	0.00	0.00	0.00	0.00	0.00	9.00	0.00	0.00	0.00
G8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.00	0.00	0.00
G9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.00	0.00
G10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9.00
	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10

Figure 9: 70:30 ration confusion matrix



	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
G1	10.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G2	0.00	9.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
G3	0.00	0.00	10.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G4	0.00	0.00	0.00	10.00	0.00	0.00	0.00	0.00	0.00	4.00
G5	0.00	0.00	0.00	0.00	10.00	0.00	0.00	0.00	0.00	0.00
G6	0.00	0.00	0.00	0.00	0.00	10.00	0.00	0.00	0.00	0.00
G7	0.00	0.00	0.00	0.00	0.00	0.00	9.00	0.00	0.00	0.00
G8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.00	0.00	0.00
G9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.00	0.00
G10	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.00

Figure 10: 50:50 ration confusion matrix



	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
G1	8.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G2	0.00	8.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G3	0.00	0.00	8.00	0.00	2.00	0.00	0.00	0.00	2.00	0.00
G4	0.00	0.00	0.00	10.00	0.00	0.00	0.00	0.00	0.00	2.00
G5	0.00	0.00	0.00	0.00	6.00	0.00	0.00	0.00	0.00	0.00
G6	2.00	0.00	0.00	0.00	0.00	10.00	0.00	0.00	0.00	0.00
G7	0.00	0.00	0.00	0.00	0.00	0.00	10.00	0.00	0.00	0.00
G8	0.00	0.00	0.00	0.00	2.00	0.00	0.00	10.00	0.00	0.00
G9	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	8.00	0.00
G10	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.00

Figure 11: 70:30 ration confusion matrix

Fig 9, 10 and 11 shows the respective confusion matrix of 70:30, 50:50 and 30:70 ratio of the training and testing samples. The column of the confusion matrix represents the actual label and the row represents the output label.

CONCLUSION

The implemented proposed method for gesture recognition is robust to illumination changes as well as gesture orientation. Most of theapplications using SVMs showed SVMs-based problem-solving approach are better than one to one comparison method. The recognition that is achieved for the 10 set of gestures is 97% . In future the incorporation of the deep learning methods and the feature dimension reduction methods we can work on increasing the accuracy of the system and implement the system on the real time.

REFERENCES:

[1] David G. Lowe, “Distinctive Image Features fromScale-Invariant Key points” January 5, 2004

[2] Shariz, S. and Kulkarni, A. "Identifying HandConfigurations with Low-Resolution Depth Sensor Data", CS229 Final Project Paper, 2009.

[3] Marx, M., Fenton, M., and Hills, G. "Recognizing Hand Gestures with a 3D Camera", CS229 Final Project Paper.

[4] Starner, T. and Pentland, A. "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video", IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 20, No. 12, December 1998

[5] Ruslan Kurdyumov, Phillip Ho, Justin Ng, “Sign Language Classiffication Using Webcam Images”, December 16, 2011

[6] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen LinDepartment of Computer Science National Taiwan University, Taipei106, Taiwan http://www.csie.ntu.edu.tw/~ cjlin Initial version: 2003 Last updated: April 15, 2010

[7] Navneet Dalal and Bill Triggs “Histograms of Oriented Gradients for Human Detection”, INRIA Rh.one-Alps, 655 avenues de l'Europe, Montbonnot 38334, France Navneet.Dalal,Bill.Triggsg@inrialpes