An Analytical Approach towards Conversion of Human Signed Language to Text Using Modified Scale Invariant Feature Transform (SIFT)

Submitted in the partial fulfillment for the award of

MASTER OF TECHNOLOGY

IN

SOFTWARE TECHNOLOGY

by

Ved Prakash

Roll No: 2K12/SWT/14

Under the Guidance of

Prof. Anil Singh Parihar

Assistant Professor

Department of Computer Engineering

Delhi Technological University

New Delhi

2015

DECLARATION

I hereby declare that the thesis entitled "**An Analytical Approach towards Conversion of Human Signed Language to Text Using Modified Scale Invariant Feature Transform (SIFT)"** which is being submitted to the **Delhi Technological University**, in partial fulfillment of the requirements for the award of degree of **Master of Technology in Software Technology** is an authentic work carried out by me. The material contained in this thesis has not been submitted to any university or institution for the award of any degree.

Ved Prakash

Department of Computer Engineering

Delhi Technological University,

Delhi.

CERTIFICATE

DELHI TECHNOLOGICAL UNIVERSITY

Date: ____________________

This is to certify that the thesis entitled " **An Analytical Approach towards Conversion of Human Signed Language to Text Using Modified Scale Invariant Feature Transform (SIFT)"** submitted by **Ved Prakash (Roll Number: 2K12/SWT/14),** in partial fulfillment of the requirements for the award of degree of Master of Technology in Software Technology, is an authentic work carried out by her under my guidance. The content embodied in this thesis has not been submitted by her earlier to any institution or organization for any degree or diploma to the best of my knowledge and belief.

Project Guide

Prof. Anil Singh Parihar

Assistant Professor

Department of Computer Engineering

Delhi Technological University, Delhi-110042

ACKNOWLEDGEMENT

I take this opportunity to express my deepest gratitude and appreciation to all those who have helped me directly or indirectly towards the successful completion of this thesis.

Foremost, I would like to express my sincere gratitude to my guide **Prof. Anil Singh Parihar**, **Assistant Professor, Department of Computer Engineering**, **Delhi Technological University, Delhi** whose benevolent guidance, constant support, encouragement and valuable suggestions throughout the course of my work helped me successfully complete this thesis. Without his continuous support and interest, this thesis would not have been the same as presented here.

Besides my guide, I would like to thank the entire teaching and non-teaching staff in the Department of Computer Engineering, DTU for all their help during my course of work.

Ved Prakash

2K12/SWT/14

Master of Technology (Software Technology)

Delhi Technological University

Bawana Road, Delhi - 110042

Table of Contents

List of Figures

List of Tables

List of Equations

Abbreviations

1. SLR : Sign Language Recognition
2. CRT: Color Recognition Technique
3. ASL: American Sign Language
4. ISL: Indian Sign Language
5. SIFT: Scale Invariant Feature Transform
6. BoG: Bag of Features
7. DOF: Degree of Freedom
8. HCI: Human Computer Interface
9. PCA: Principal Component Analysis
10. SURF: Speeded Up Robust Feature
11. KNN: K Nearest Neighbors
12. HGR: Hand Gesture Recognition
13. SL: Sign Language
14. CRT: Color Recognition Technique
15. SVM: Support Vector Machine

ABSTRACT

Sign language is used as a communication medium among deaf & dumb people to convey the message with each other. A person who can talk and hear properly (normal person) cannot communicate with deaf & dumb person unless he/she is familiar with sign language. Same case is applicable when a deaf & dumb person wants to communicate with a normal person or blind person. In order to bridge the gap in communication among deaf & dumb community and normal community, researchers are working to convert hand signs to voice and vice versa to help communication at both ends. A lot of research work has been carried out to automate the process of sign language interpretation with the help of image processing and pattern recognition techniques.

The approaches can be broadly classified into “Data -Glove based” and “Vision-based” .Tracking bare hand and operations to detect hand from image frames. The main drawback of this method lies in its huge computational complexity which is further handled with the concept of integral image. The use of integral image for hand detection in viola-Jones method reduces computational complexity and shows satisfactory performance only in a controlled environment. To detect hand in a cluttered background, many researchers used color information and histogram distribution model. Some Local orientation histogram technique is also used for static gesture recognition. These algorithms perform well in a controlled lighting condition, but fails in case of illumination changes, scaling and rotation. To resist illumination changes, Elastic graphs are applied to represent different hand gestures

with local jets of Gabor Filters. Adaboost for wearable computing is insensitive to camera movement and user variance. Their hand tracking is promising, but segmentation is not reliable. Fourier descriptors of binary hand blobs used as feature vector to Radial Basis Function (RBF) classifier for pose classification and combined HMM classifiers for gesture classification. Even though their system achieves good performance, it is not robust against multi variations during hand movement. To overcome the problem of multi variations like rotation, scaling, translation some popular techniques like SIFT, Haar-like features with Adaboost classifiers, Active learning and appearance based approaches are used. However, all these algorithms suffer from the problem of time complexity. To increase the accuracy of the hand gesture recognition system, combined feature selection approach is adopted.

My thesis proposes new approach of hand gesture recognition which will recognize sign language gestures in a real time environment. A hybrid feature approach, which combines the advantages of SIFT, Principal Component Analysis, Histogram and they are used as a combined feature set to achieve a good recognition rate. To increase the recognition rate and make the recognition system resilient to view-point variations, the concept of principal component analysis introduced. K-Nearest Neighbors (KNN[11]) is used for hybrid classification of single signed letter. In addition, integration of color detection method is under progress to increase the accuracy further. The performance analysis of the proposed approache is presented along with the experimental results. Comparative study of these methods with other popular techniques shows that the real time efficiency and robustness are better.

Chapter 1

INTRODUCTION

Sign Language Recognition (SLR[1]) could be a comparatively new field of analysis in image process and computer vision that has gained increasing attention within the recent years attributable to its revolutionary vision in terms of major technology changes. Sign languages is thought-about natural languages with identical communicative power as spoken languages. Automatic recognition of gestures normally is that the next step towards computer-human interaction as signs will boost or complement speech. A helpful application of SLR[1] is associate degree interactive learning setting for hearing impaired individuals. Though human's area unit typically ready to interpret the signs from visual supply with relative ease, for a computer this can be troublesome. Vision based SLR[1] techniques are complex and dependent on the sensor data. Normally low-level options area unit extracted from video frames or from static pictures and used for classification. Knowing wherever humans focus their attention to sign recognition advantages SLR[1]. Contrary to widespread belief, linguistic communication isn't solely composed of manual signs. Head movement, brow configuration, facial expressions area unit important cues additionally except for our analysis work we'll target recognition of alphabet and numbers from hand gesture. the fundamental technology can stay same for any reasonably gesture recognition, therefore instead of that specialize in additional varieties we must always target accuracy of designed formula in order that in future we will improve it additional and might be commercialized for public use. Many alternative strategies for SLR are developed however still those all strategies area unit beneath analysis solely. Till date we don’t have any method which can be commercialized since researchers are struggling with the accuracy of the implemented algorithm. Industry will only pay attention on this once any researcher can guaranty about the reliability of the developed algorithm.

After going through the development trend and the results from many researchers, I have thought to improve the accuracy of existing method and algorithm so that it can be helpful in future for commercialization. In my thesis I have used the cascade of features to improve the accuracy of result. Currently I am not much bothered about the time complexity of algorithm, as for any researcher the first aim should be, achieve the result, later it can be optimized.

1.1 Motivation of the Work

Vision-based Gesture Recognition recently became a highly active research area with motivating applications such as Sign Language Recognition (SLR), Socially Assistive Robotics, and Directional Indication through Pointing, Control through Facial Gestures, Human Computer Interaction (HCI[8]), Immersive Game Technology, Virtual Controllers, Affective Computing and Remote Control. Within the broad range of application scenarios, hand gestures can be categorized into at least four classes: controlling gestures, conversational gestures, communicative gestures, and manipulative gestures. Hand gestures are powerful human interface components. However, their fluency and intuitiveness have not been utilized as computer interface. Recently, hand gesture applications have begun to emerge, but they are still not robust and are unable to recognize the gestures in a convenient and easily accessible manner by the human. Several advanced techniques are still either too fragile or too coarse grained to be of any universal use for hand gesture recognition.

Especially, techniques for hand gesture interfaces should be developed beyond current performance in terms of speed and robustness to attain the needed interactivity and usability.

It is a difficult task to recognize hand gestures automatically from camera input. It usually includes numerous phases such as signal processing, detection, tracking, shape description, motion analysis, and pattern recognition. The general problem is quite challenging because of several problems such as the complex nature of static and dynamic hand gestures, cluttered backgrounds, transformations, lighting changes, and occlusions. Trying to solve the problem in its generality needs elaborate techniques that require high performance against these issues.

Hand gesture recognition from video frames is one of the most main challenges in image processing and computer vision because it provides the computer the capability of detecting, tracking, recognizing and interpreting the hand gestures to control various devices or to interact with several human machine interfaces.

The objective of this thesis is to develop a approach to the current problem of hand gesture recognition. By proposing a cascade of features method to an existing problem, it is anticipated that this technique is a step forward in practical applications of hand gesture recognition for everyday use. The cascade of features should satisfy numerous conditions:

- The first requirement is real-time performance. This is critical for the universal acceptance of the method. This is measured in terms of matching result of the test image with respect to the training image. Classification result of testing image should be accurate and reliable so that the method can be further included in any commercial project.
- The second required condition is flexibility, and how well it combines with new applications and existing applications. The algorithm should be able to accommodate

external programs easily to be a candidate for practical applications. This will be an advantage for the application developer and the user.

- Third, the approach should be practically precise enough to be used. The approach should correctly recognize the defined gestures 90%-100% of time to be successful and of practical use.
- The fourth needed condition is robustness in which the system should be able to detect, track, and recognize different hand gestures successfully under different lighting conditions and cluttered backgrounds. The system should also be robust against scale and rotations.
- The fifth needed condition is scalability. A large gesture vocabulary can be involved with a small number of primitives. The user interacts easily with an application by building different gesture commands.
- Finally, the approach should be user-independent in which the system must be able to work for various persons rather than a particular person. The system must recognize hand gestures for different human hands of different scales and colors.

This thesis proposes a cascade of features Sign Language Recognition technique, which uses hand gestures as input for communication and interaction. The system is starts with capturing images from a webcam or a pre-recorded video file. Several systems in the literature have strict restrictions such as using particular gloves, uniform background, long-sleeved user arm, being in particular lightning conditions and using particular camera parameters. These restrictions destroy the recognition rate and naturalness of a hand gesture recognition system. The performances of those systems are not strong enough to be used on a real-time HCI [8] system. This thesis aims to design a vision based hand gesture recognition system with a high recognition rate along with real-time performance. The system is invariant against previous strict restrictions on the human environment and can be used for real-time HCI[8] systems.

Usually, these interaction systems have two challenges: hand detection and hand gesture recognition. Hand detection must be done before gesture recognition. Once the hand is detected clearly in the current image, the gesture recognition process is started around the detected hand.

Scale Invariance Feature Transform (SIFT) features, proposed by Lowe, are features extracted from images for helping in reliable matching between different views of the same object, image classification, and object recognition. The extracted key points are invariant to scale, orientation and partially invariant to illumination changes, and are highly distinctive of the image. Therefore, the SIFT is adopted in this thesis for the bare hand gesture recognition. However, SIFT features are too high dimensionality to be used efficiently. We propose to solve this problem by the cascade of features approach to reduce the dimensionality of the feature space.

A hand gesture is an action, which consists of a sequence of hand postures. Gestures are restricted to a discrete set of postures recognized in static form. In this thesis, we will highlight the dynamic aspect of hand gestures and how to recognize different hand gestures that differ solely in their timing and space aspects. Our goal is to recognize dynamic gestures in real-time with high recognition rate.

1.2 Objective of the project work

The goal of the work in this thesis is summarized below:

The main aim of this research work is to convert sign languages into text or speech so that it can be helpful for differently abled people. Deaf and dumb people can communicate with other person with ease without facing any difficulty.

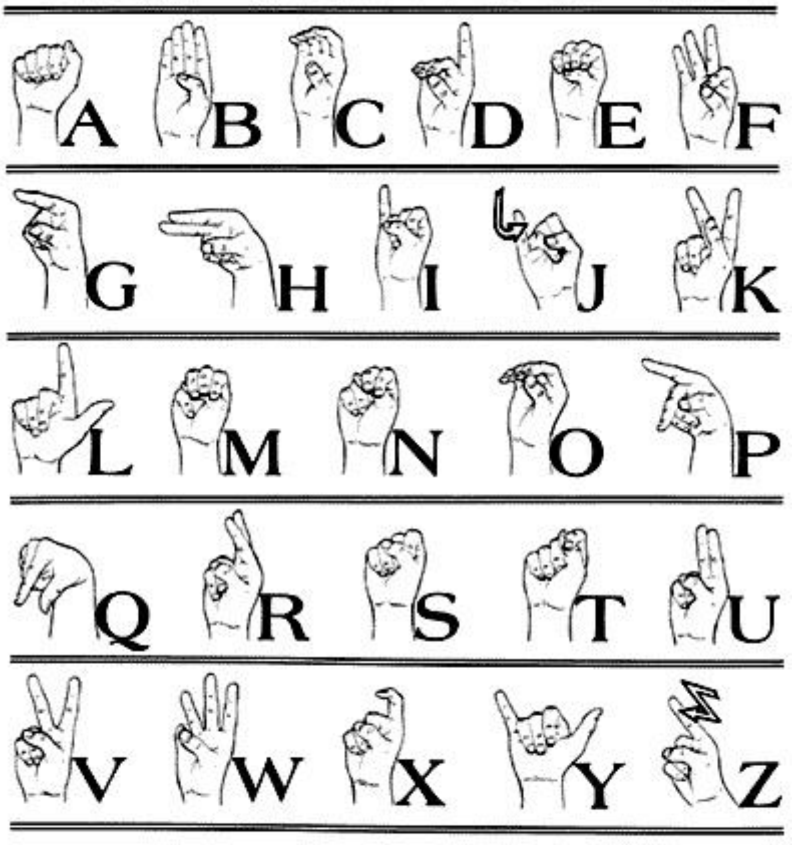


Figure 1: American Sign Language [ASL][2]

Sign languages are the most raw and natural form of languages could be dated back to as early as the advent of the human civilization, when the first theories of sign languages appeared in history. It has started even before the emergence of spoken languages. Since then the sign language has evolved and been adopted as an integral part of our day to day communication process. Now, sign languages are being used extensively in international sign use of deaf and dumb, in the world of sports, for religious practices and also at work places Gestures are one of the first forms of communication when a child learns to express its need for food, warmth and comfort. It enhances the emphasis of spoken language and helps in expressing thoughts and feelings effectively. A simple gesture with one hand has the same meaning all over the world and means either 'hi' or 'goodbye'. Many people travel to foreign countries without knowing the official language of the visited country and still manage to perform communication using gestures and sign language. These examples show that gestures can be considered international and used almost all over the world. In a number of jobs around the world gestures are means of communication. In airports, a predefined set of gestures makes people on the ground able to communicate with the pilots and thereby give directions to the pilots of how to get off and on the run-way and the referee in almost any sport uses gestures to communicate his decisions. In the world of sports gestures are common. The pitcher in baseball receives a series of gestures from the coach to help him in deciding the type of throw he is about to give. Hearing impaired people have over the years developed a gestural language where all defined gestures have an assigned meaning. The language allows them to communicate with each other and the world they live in.

Apart from signed language gesture recognition, the basic method developed can be used for other real time applications like given below:

3D Design: Computer aided design is an HCI[8] which provides a platform for interpretation and manipulation of 3-Dimensional inputs which can be the gestures. Manipulating 3D inputs with a mouse is a time consuming task as the task involves a complicated process of decomposing a six degree freedom task into at least three sequential two degree tasks. MIT has come up with the 3DRAW technology that uses a pen embedded in polhemus device to track the pen position and orientation in 3D.A 3space sensor is embedded in a flat palette, representing the plane in which the objects rest .The CAD model is moved synchronously with the users gesture movements and objects can thus be rotated and translated in order to view them from all sides as they are being created and altered.

Tele presence: There may raise the need of manual operations in some cases such as system failure or emergency hostile conditions or inaccessible remote areas. Often it is impossible for human operators to be physically present near the machines. Tele presence is that area of technical intelligence which aims to provide physical operation support that maps the operator arm to the robotic arm to carry out the necessary task, for instance the real time ROBOGEST system constructed at University of California, San Diego presents a natural way of controlling an outdoor autonomous vehicle by use of a language of hand gestures. The prospects of tele presence includes space, undersea mission, medicine manufacturing and in maintenance of nuclear power reactors.

Virtual reality: Virtual reality is applied to computer-simulated environments that can simulate physical presence in places in the real world, as well as in imaginary worlds. Most current virtual reality environments are primarily visual experiences, displayed either on a computer screen or through special stereoscopic displays. There are also some simulations include additional sensory information, such as sound through speakers or headphones. Some advanced, haptic systems now include tactile information, generally known as force feedback, in medical and gaming applications.

1.3 Organization of the Thesis

This thesis includes 6 chapters:

- **Chapter 1** introduces the background, vision-based hand gesture processing stages, and motivations of this work. The objectives to be accomplished by the work are also given.

- **Chapter 2** provides a comprehensive literature review based on different categories for vision-based hand gesture detection methods.
- **Chapter 3** presents our algorithm to detect sign language, which includes the training of images using cascade of features and then applying the method for our test images.
- **Chapter 4** proposes the overall architecture for the system: a multi stage architecture which decouples the hand gesture recognition system into a training stage to build a k-means clustering model to use them in the testing stage to recognize hand gesture using Cascade of Features.
- **Chapter 5** presents a method for recognizing hand gesture using Color Recognition Technique [CRT][2]which is proposed for integration with Cascade of Features method in future. Development work is under progress.
- **Chapter 6** provides conclusions and outlines future work.

Chapter 2

LITERATURE SURVEY

Human Computer Interaction moves forward in the field of sign language interpretation. Sign Language (ISL[4]) Interpretation system is a good way to help the hearing impaired people to interact with normal people with the help of computer. Vision based hand gesture recognition system have been discussed as hand plays vital communication mode. Considering earlier reported work, various techniques available for hand tracking, segmentation, feature extraction and classification are listed. Vision based system have challenges over traditional hardware based approach; by efficient use of computer vision and pattern recognition, it is possible to work on such system which will be natural and accepted, in general Sign languages (SL) are known as Deaf and Dumb languages. SLs are gestural languages which contain symbolic encoded message for communication without speech channel. They are unique in some ways in that they cannot be written like spoken language. Sign language varies from country to country with its own vocabulary and grammar. Even within one country, sign language can vary from region to region like spoken languages. Gestures are powerful means of communication among humans. Among different modality of body, hand 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[8]). With the wide applications of HCI[8], now days, it becomes active focus of research. To have an interaction with computer, vision based system is more suitable than traditional data glove based system, as sensors are attached to the data glove and data suit where, user has to wear these cumbersome devices . In this paper, Vision based approach have been discussed for interpreting the Indian sign language using hand cascade of different methods algorithm. A Typical Hand Gesture Recognition system consists of mainly four modules: Gesture acquisition, Tracking and segmentation, Feature extraction and description, Classification and recognition. This paper focuses on a study of sign language interpretation system with reference to vision based hand gesture recognition. An attempt has also been made to explore about the need and motivation for interpreting SL, which will provide opportunities for hearing impaired peoples in Industry Jobs, IT sector Jobs, and Government Jobs.

The Sign language to Speech/Text & Speech/Text to Sign Language, these should be both way communication to solve the problem of hearing impaired people. The second part of this is not complex as we just need to map the animation frames according to the grammar. All the research work is ongoing mainly for the first part i.e. from Sign Language to Text/Speech conversion which is the complex part and requires robust algorithms.

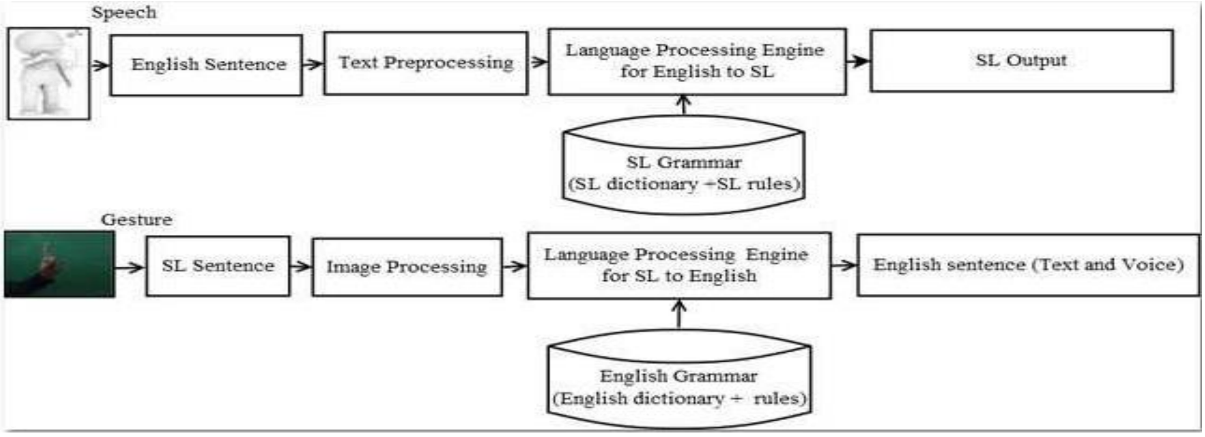

Speech
English Sentence
Text Preprocessing
Language Processing Engine for English to SL
SL Output
SL Grammar (SL dictionary +SL rules)
Gesture
SL Sentence
Image Processing
Language Processing Engine for SL to English
English sentence (Text and Voice)
English Grammar (English dictionary + rules)

Figure 2: Speech to SL and SL to Speech Conversion System [1]

Linguistic work on sign language began and with contribution of a team of researcher from America in 1970. It was found that SL is a language in its own right and is indigenous to the different subcontinent and resulted in many dictionaries.

Major research work is going on awareness and multilingual sign language dictionary tool so there is a need for sign language interpretation tool. Following may be the major advantages of sign language interpretation:

- Use and awareness of computer interface through SL interpretation.
- Education and training will be easier through SL interpretation/visualization for deaf and dumb people.
- Serving the mankind by use of technology.
- Social aspect like humanity can increase in individual mind by involving physically impaired people in our day to day life.
- Blind people can also use the same system by extending it for voice interface.

Sign language is not a universal language. Sign language recognition is a multidisciplinary research area involving pattern recognition, computer vision, natural language processing and psychology. Figure 2 shows the typical architecture for sign language interpretation system. It broadly divides into two modules. First module is for converting normal English sentences in to SL (to be understood by deaf people) and another module is for converting SL into English text (to be understood by normal people). For literate hearing impaired people, those who can read English, first module is not required. But for illiterate deaf and dumb people, both modules are essential. In both the module, language processing engine is required which is based on a particular language rules. Conversion of sign to text includes the area of computer vision, image processing, pattern recognition and language processing with linguistic study.

After study and investigation, it was found that there is a relation between human gesture and speech. Speech expression can be replaced by signs going from gesticulation to sign language. SL is a visual-spatial language. It is having linguistic information in the form of hands, face, arms and head/body posture and movements. Visual channel is active in sign language like speech channel in spoken language. Figure 3 illustrates the Indian Sign Language set:

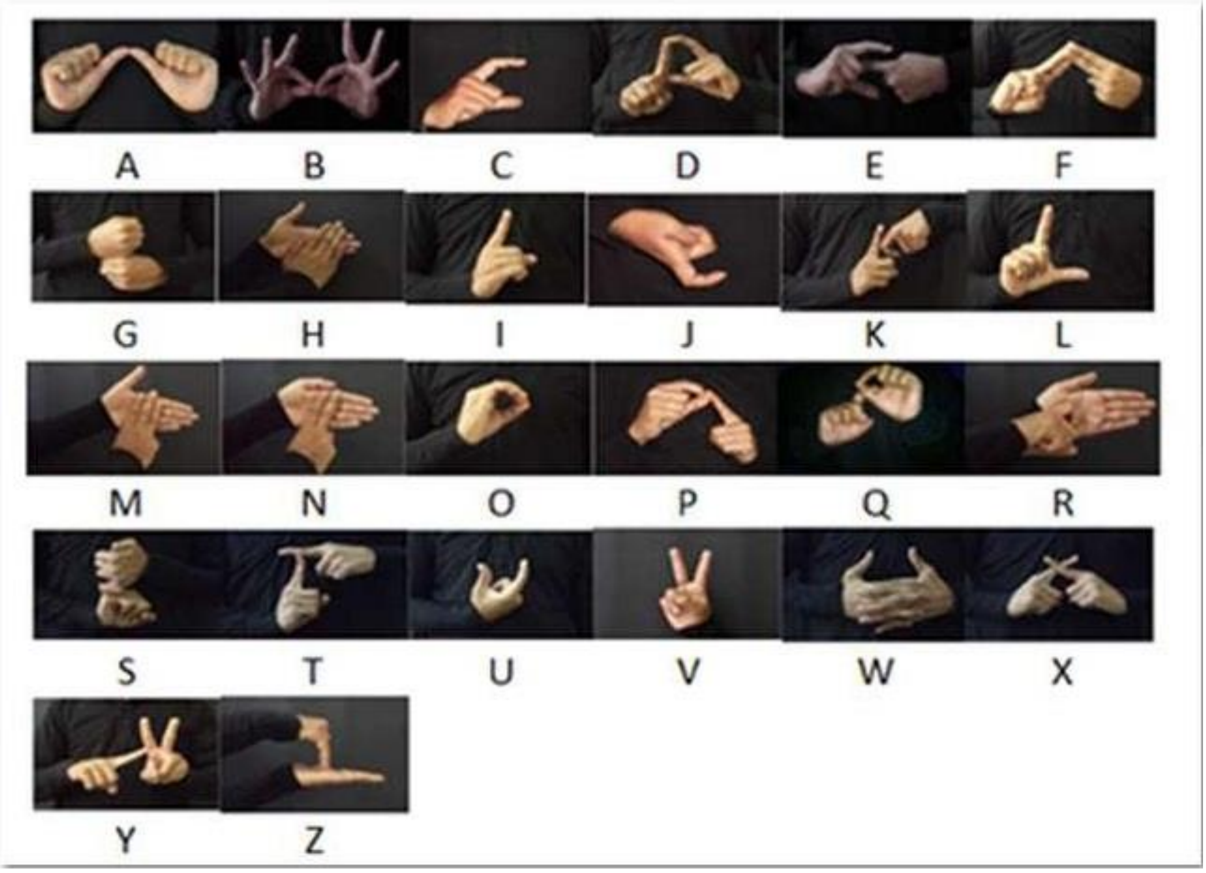


Figure 3: Indian Sign Language [ISL][4]

Though it is found that hand plays active role in sign language but due to its complex articulated structure consisting of many connected links and joints, hand gesture recognition becomes a very challenging problem. Figure 4 shows skeleton structure and the joints of the human hand, with total 27 degree of freedom (DOF[6]) considering hand wrist. There are widely two terms used in hand gesture recognition system: 1) Hand posture (static hand gesture) and

Hand gesture (Dynamic hand gesture). In hand posture, no movements are involved whereas; hand gesture is a sequence of hand posture connected by movement over a period of time In dynamic hand gesture, again two aspects are considered such as local finger motion without changing hand position or orientation and global hand motion where, position or orientation of hand gets changed. Study of hand skeleton model is very essential for developing any hand gesture recognition system.

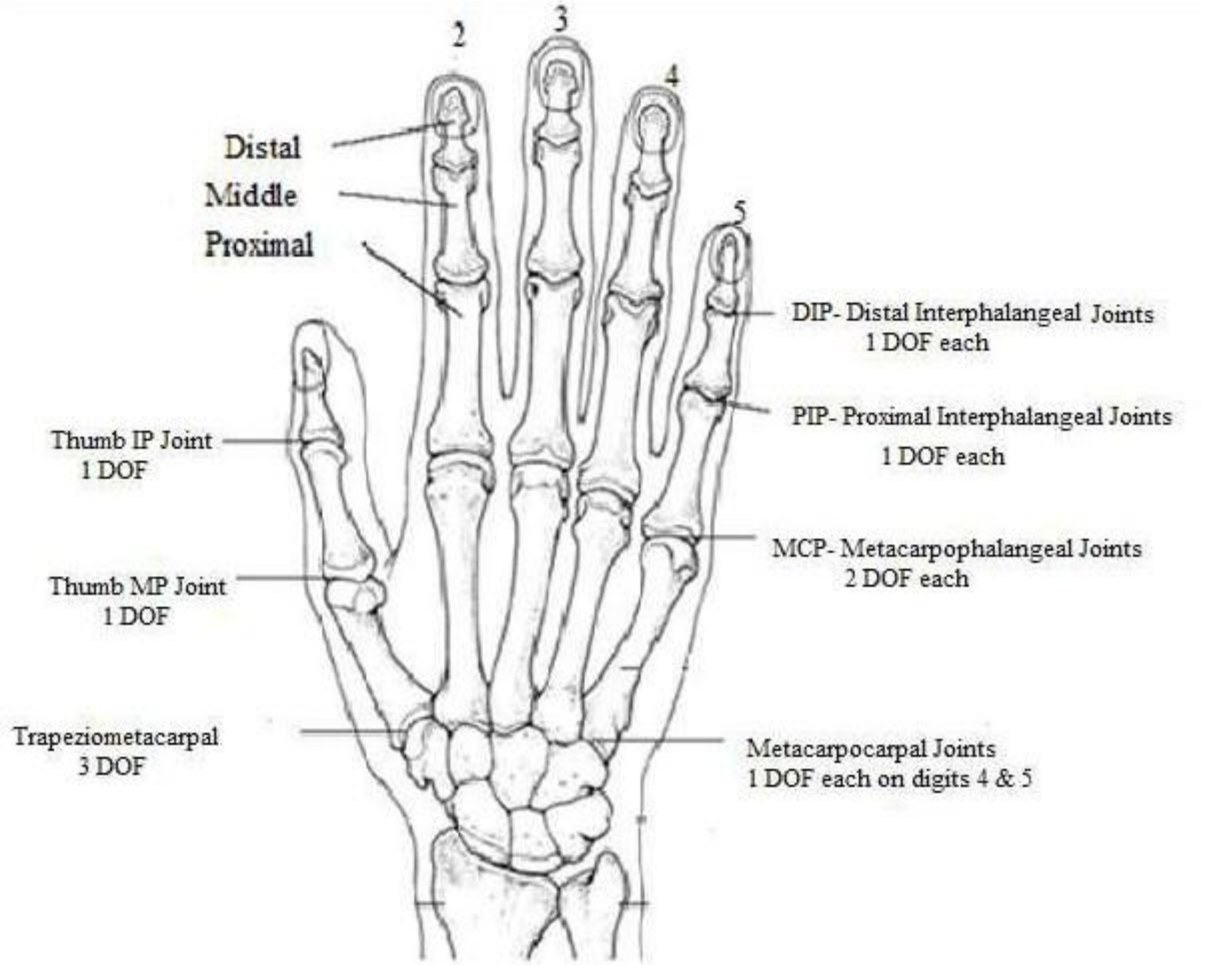


Figure 4: Degree of Freedoms of Hand Skeleton [6]

In any typical hand gesture recognition system, a good and strong set of features, description and representation are required. In the current state of the art, due to the limitation of data glove/sensor based approaches, vision based and 3-D hand model based approaches are being used. One of the major tasks in hand gesture recognition is the description of the gesture. Various methodologies are found in the literature such as, statistical and synthetic based approaches. In statistical representation, one can represent it, in the form of feature vector and then apply classification and recognition algorithm; whereas synthetic gesture recognition gesture can be represented in the form of tree, string or graph and decision rule such as graph matching, decision tree and string matching. Now days in the field of Human Computer Interaction, Hand Gesture recognition [HGR][12] is an active research topic. In this section, various approaches and techniques have been explored related to hand gesture recognition. Recognizing gesture is a complex task which involves many aspects such as object detection, object description, motion modeling, motion analysis, pattern recognition and machine learning even psycholinguistic studies also required.

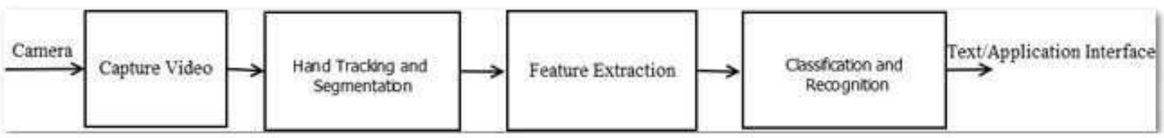

Camera
Capture Video
Hand Tracking and Segmentaton
Feature Extraction
Classification and Recognition
Text/Application Interface

Figure 5: HGR System

Figure 5 shows typical architecture of HGR[12] system. Hand tracking and segmentation are to be done on captured video/ Image Frame and feature extraction is to be done on segmented hand image which is further given to classification and recognition phase. Output is to be printed or executed, depending on the application.

After capturing and separating frames from videos, the elementary and important task is detection and segmentation of hands. There are various approaches and techniques available in literature but the results vary, images to images due to the limitation of vision based approach such as variable lightning condition, variation of skin color, detection of hand in complex background. Pixel and region based segmentation techniques are available. It has been observed that HSV color model gave better result for skin color detection than other models due to the separation property of luminance and chrominance component. Some researcher used additional marker or color gloves for hand segmentation using color threshold, but for natural interface bare hand interaction is always preferred. Supervised as well as Unsupervised Learning Model such as Bayesian classifier can be used for skin color segmentation. Unsupervised learning such as, K-mean clustering is also a good option for skin color segmentation. 2D Tracking algorithm gives the position information of hand such as color tracking, motion tracking, template matching, blob tracking, Multiple cues integrating methods are available. It has been noticed that tracking algorithm such as mean shift, camshaft and viola jones with appropriate color space gave better segmentation result in complex background.

Shape is the important visual feature of the hand. Zhang and Lu gave classification of shape representation and description techniques based on contour and region. In contour based method, shape features are extracted from the shape boundary whereas, in region based method features are extracted from the whole shape.

- Contour-based shape representation and description methods are chain Code, Polygon, B-spline, Perimeter, Compactness, Eccentricity, Shape Signature, Hausdoff Distance, Fourier Descriptor, Wavelet Descriptor, Scale Space, Auto regressive ,Elastic matching.
- Region-based shape representation and description methods are Convex Hull, Media Axis, Area, Euler Number, Eccentricity, Geometric Moments, Zernike Moments, Pseudo-Zernike Moments and Legendre Moments.

In hand recognition problem, shape contour is important than whole region so, contour based methods are mostly used. But for complex sign, sometimes region based methods are more suitable because it contains all the available information. In case of the new signer for performing gesture, there may be chances for angle deviation, shifting of signer space can occur. Hand size of the signer can also vary. So, while choosing feature extraction method, care must be taken that it should be invariant to translation, rotation and scale. SLs contain large set of vocabulary, use of one of the feature extraction techniques is not sufficient. Practically combination of feature vector and motion vector is the better choice to get accuracy. Table 1 shows the earlier reported work on hand gesture recognition on various segmentation and feature extraction techniques.

Table 1: Survey on Different Segmentation and Feature extraction Techniques

Parameter	Techniques	Accuracy
Segmentation and Tracking techniques	YCbCr color space, Kmeans embedded particle filter for two hand tracking	Accuracy: 83% a) worked better than mean shift algorithm, b) tracking fail for rapid movement of hand.
	Tower method for hand tracking	Faster than camshift
	Two hand segmentation with Haar-Like feature and adaptive skin color model	Accuracy: 89% to 98% for four movement
	kalman filtering and a collapsing method	Satisfactory results
	Viola Jones method for tracking	Fast and most accurate learning-based method for object detection
	Color based segmentation using HSV, L*a*b color spaces and camshift method for tracking	Camshift tracking with HSV color model gives better result in complex background, different lighting condition and skin color

Feature extraction techniques	Angle and distance from endpoint	Accuracy: 92.13% No. of gesture used : 10
	Haar wavelet, Code word scheme	Accuracy: 94.89% No. of gesture used : 15
	Location, angle, velocity and motion pattern P2-DHMMS	Accuracy: up to 98% No. of gesture used : 36
	Orientation Histogram, Neural network	Accuracy: up to 90 No. of gesture used : 33
	Co-occurrence Matrix, local and global features	Accuracy: 93.094% No. of gesture used : 30
	Key trajectory point selection, trajectory length selection, location feature extraction, orientation feature extraction, velocity and acceleration	Accuracy: Static-92.81% Dynamic:87.64% No. of gesture used : 26

Apart from above mentioned features and segmentation based algorithms, David Lowe has proposed one scale invariant algorithms which was milestone in pattern matching in image processing. The Scale Invariant Feature Transform [SIFT][5] was first proposed algorithm in history of gesture recognition technology which was robust and effect of 3D orientation, illumination, scaling was lowest. There are many modifications done by different researchers to improve the performance of SIFT[5] and thus we can see currently many modified version of SIFT. The latest improvement in SIFT[5] termed as SURF[11] algorithms which stands for Speeded Up Robust Features. The uniqueness of SIFT[5] was it was invariant to scaling of object but the dynamic movement was not properly covered under SIFT. SURF[11] algorithm takes care of the dynamic movement of object and it is robust in terms of scaling as well as speed of object.

My thesis proposes new approaches of hand gesture recognition which will recognize sign language gestures in a real time environment. A hybrid feature approach, which combines the advantages of SIFT, Principal Component Analysis, Histogram and they are used as a combined feature set to achieve a good recognition rate. To increase the recognition rate and make the recognition system resilient to view-point variations, the concept of principal component analysis introduced. K-Nearest Neighbors (KNN[11]) is used for hybrid classification of single signed letter. In addition, integration of color detection method is under progress to increase the accuracy further. The performance analysis of the proposed approaches is presented along with the experimental results. Comparative study of these methods with other popular techniques shows that the real time efficiency and robustness are better.

Chapter 3

Hand Gesture Recognition System

3.1. HGR Framework Architecture

Hand gesture recognition system consists of the following steps:

(a) Pre-processing and hand segmentation

(b) Hand detection and tracking,

(c) Hand posture recognition

(d) Hand gesture classification

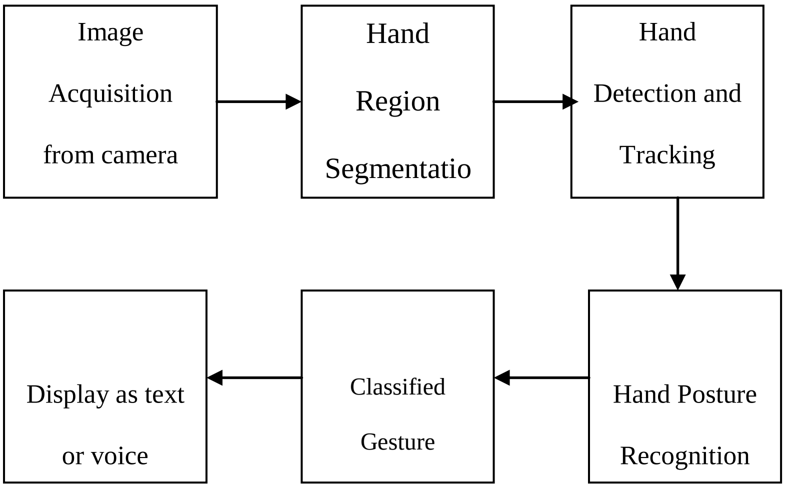


Figure 6: Hand Gesture Recognition System

3.1.1 Hand Segmentation

Skin color segmentation is performed using k-means clustering method or Skin color detection method. RGB color frames.. Skin pixel region is identified from the different color regions using a threshold method in RGB color space where the threshold value is selected experimentally. Repeat the cluster for 3 times to avoid local minima. Figure shows example results of the segmentation algorithm.

3.1.2 Hand Detection using Invariant Feature Descriptors:

After obtaining skin segmented RGB image, it is converted into gray scale. The converted gray scale image is normalized. Invariant features are extracted using Scale Invariant Feature Transform SIFT method. The basic idea is to extract the invariant key point which represents/identifies hand from the segmented image. For this purpose of hand detection, SIFT features are first extracted from a set of reference images and stored in a database. An image frame is matched by individually comparing each feature from the image frame to this previous database and finding candidate matching features based on Euclidean distance of their feature vectors.

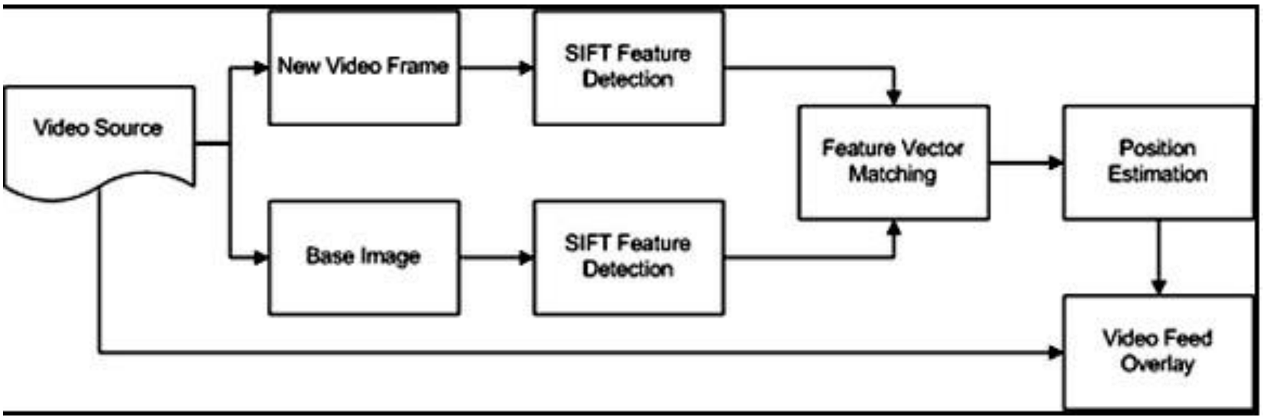


Figure 7: Block diagram for SIFT Algorithm [6]

3.1.3 Scale Invariant Feature Transform [SIFT]:

The method or algorithm adopted is called Scale Invariant Feature Transform (SIFT). When considering images, variance is one major factor that comes when the image appears in a large screen. The image window size might be standard, but the image size within the window may vary in real-time. Basically, there are five types of common invariance that could be found in images, scale invariance, rotation invariance, illumination invariance, perspective invariance and affine transformations. As a basic and first step in building robust gesture recognition system the scale invariance, illumination invariance and rotation invariance is handled in this work. The SIFT algorithm helps in managing this invariance. The method followed is depicted in the figure 2. The feature extraction is done by firs finding the key points. The scale and location variance are eliminated at this stage by sub pixel localization and edge elimination. Sub pixel elimination is done by down sampling the region of interest considered. The edge is identified by the Sobel edge detection method [13] and cropped. The signature images are derived from the image gradients which are sampled over 16*16 array of locations in scale space, then an array of orientation histograms are drawn for the same. The figure 9 shows the image gradients and the keypoint descriptors derived for an image.

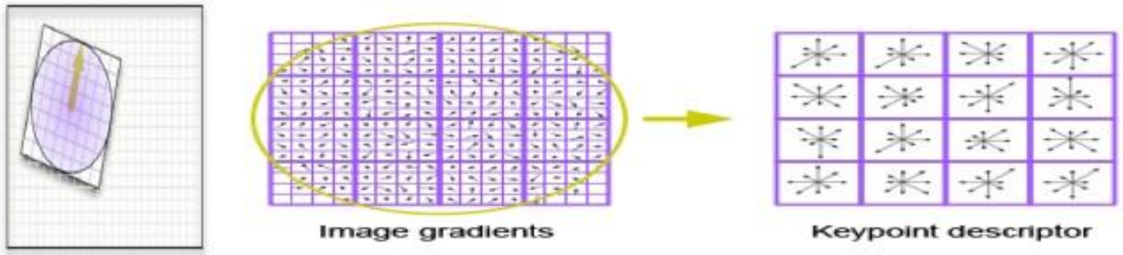


Figure 8: SIFT Keypoint Descriptor [6]

The method adopted to achieve this scale invariance is scale space Difference of Gaussian (DOG method). The figure 4 shows the working of the DOG method. The scale works octave by octave.

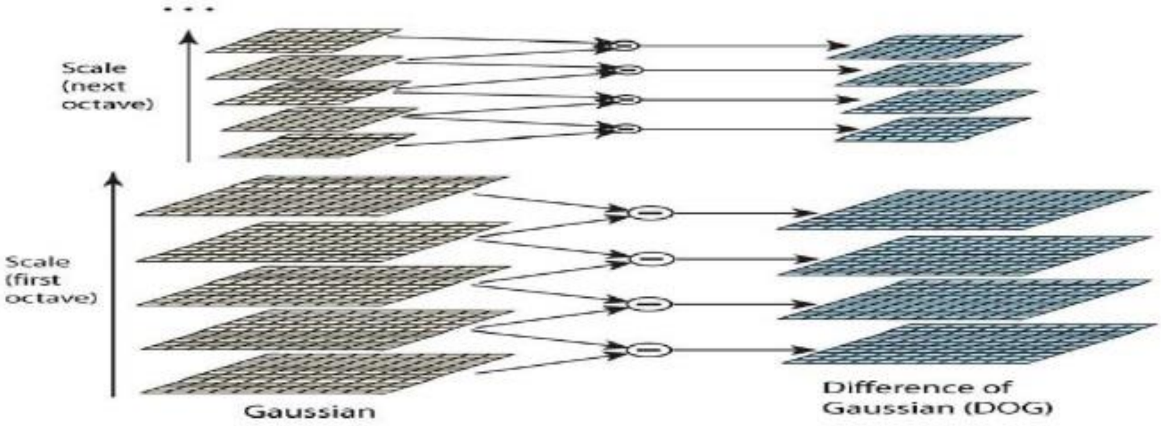


Figure 9: Difference of Gaussian [6]

Every scale is taken in term of an octave and the change in between the next octave is calculated as a Gaussian function. Since the features are derived from the difference of the images, if the feature is repeatedly present in between difference of Gaussians it is scale invariant and it is retained. This paves way as a major key factor for the performance of the system.

The algorithm used is LOWE’s SIFT algorithm [14].SIFT is an invariance algorithm and because of that feature its results are promising for real time as well as formatted images. The scale invariance is the main intention of selecting this algorithm. SIFT as defined by Lowe is a histogram of gradients. The algorithm packages keypoints in each pixel location as [row, col, scale]. Key vectors are sorted and the first 128 values are used as the feature vector for an image. The input image is compared with all its key points with the database image vectors where the nearest neighbor has angle less than the distance ratio, the keypoints are taken as matched. The maximum keypoints matched image is retrieved or recognized as that character.

The algorithm executed is depicted in fig 11. The flow explains the step by step process of the algorithm implemented.

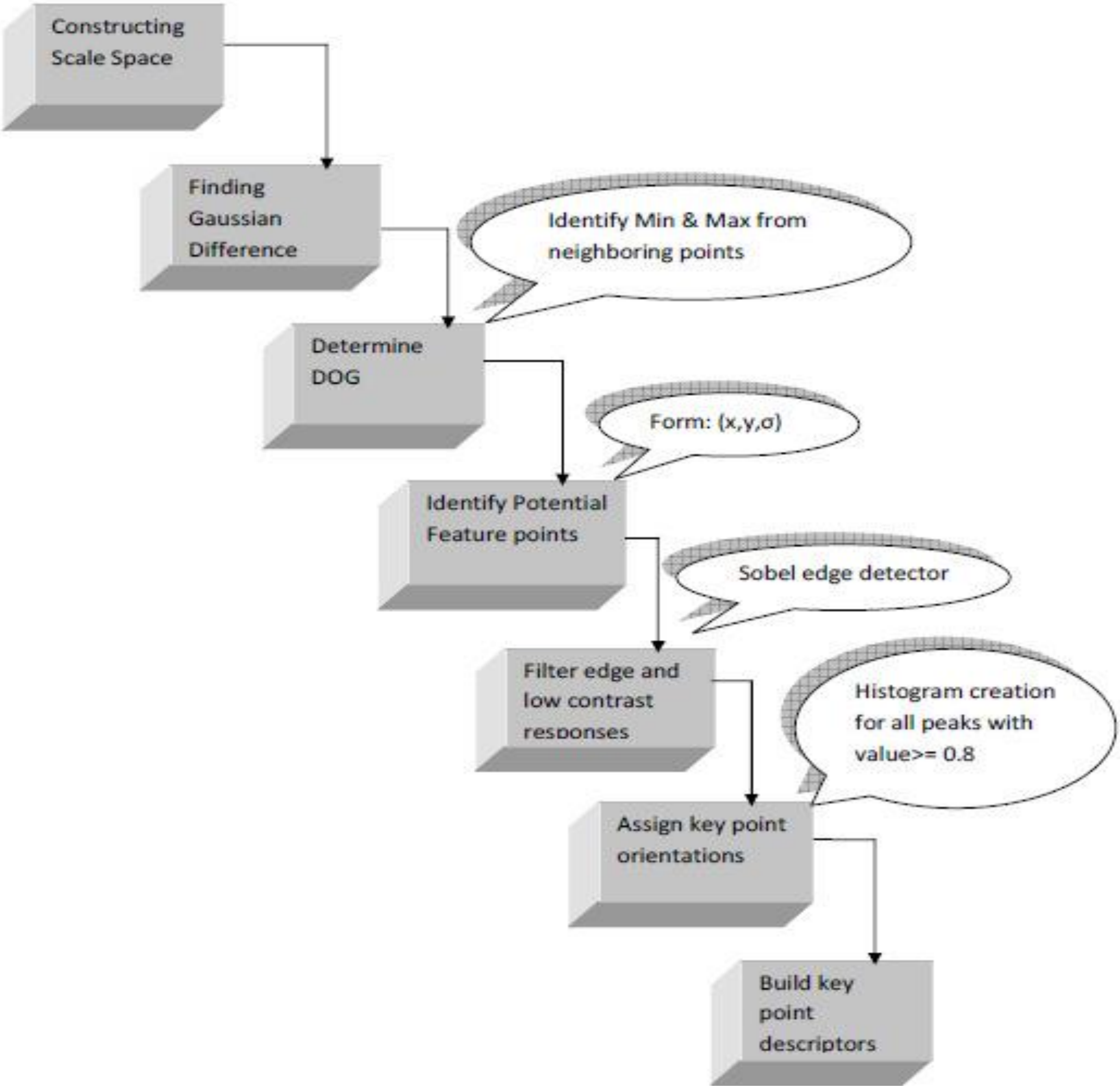

Constructing Scale Space
Finding Gaussian Difference
Identify Min & Max from neighboring points
Determine DOG
Form: (x,y,σ)
Identify Potential Feature points
Sobel edge detector
Filter edge and low contrast responses
Histogram creation for all peaks with value>= 0.8
Assign key point orientations
Build key point descriptors

Figure 10: Flow of SIFT method [6]

3.1.4 Recognition of Letters:

The bounding box of the detected hand in each frame is obtained from the previous section. To recognize the posture of detected hand, a combined feature extraction methodology using Speeded up Robust Features (SURF) and Hu Moment Invariant features is incorporated. Bounding box, BBIm (x, y) is taken as test image. Features are calculated and compared with the database features. Minimum Euclidean distance between the feature vectors recognizes particular hand posture/letter.

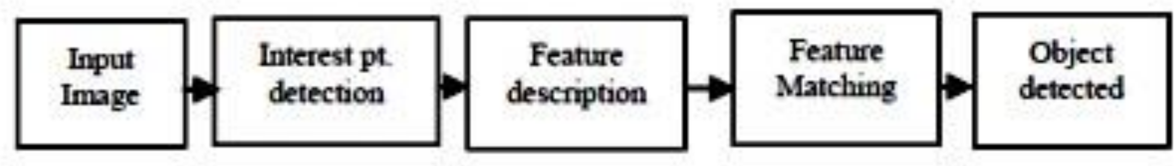


Figure 11: Block diagram for SURF Algorithm [1]

Algorithm consists of four major parts.

» Integral image generation

» Interest point detection

» Descriptor generation

3.1.5 Speeded Up Robust Features:

Given an image BBIm(x, y), integral image ii(x, y) is calculated using,

$$ii(x,y) = \sum_{\substack{x1 \le x \\ y1 \le y}} BBIm(x1, y1) \qquad (1)$$

To find out the interest points from the integral image, Fast Hessian Detector is used. Given a point X = (x, y) in image ii(x, y), the Hessian matrix H(X, σ) in X at scale σ is defined as

$$H(X,\sigma) = \begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix} \qquad \text{———} \bigcirc\!\!\!\!\!\!\;2$$

To localize interest points in the image and over scales, non-maximum suppression in a 3 × 3 × 3 neighborhood is applied. The maxima of the determinant of the Hessian matrix are the interpolated in scale and image space. In order to be invariant to rotation, Haar wavelet responses in x and y direction, within radius 6s around interest point is calculated. For the extraction of the descriptor, the first step consists of constructing a square region centered on the interest point. The region is split up regularly into Smaller 4 × 4 square sub-regions. This keeps important spatial information in. For each sub-region, a few simple features at 5×5 regularly spaced sample Points are computed. dx the Haar wavelet response in horizontal direction and dy the Haar wavelet response in vertical direction (f iltersize2s). The wavelet responses dx and dy are summed up over each sub-region and form a first set of entries to the feature vector. Absolute values of the responses |dx| and |dy| provide polarity information.

Each sub-region has a four dimensional descriptor vector. This results in a descriptor vector for all 4 × 4 sub-regions of length

3.1.6 Hu Moment Invariant Geometric Features:

Two-dimensional moments of detected hand image BBIm(x, y) of size m×m is given as,

$$m_{pq} = \sum_{x=0}^{x=m-1} \sum_{y=0}^{y=m-1} (x)^p.(y)^q BBIm(x,y) \qquad (3)$$

$$\text{p, q} = 0,1,2,3\ldots\ldots$$

The moments BBIm(x, y) translated by an amount (a, b), are defined as

$$\mu_{pq} = \sum_x \sum_y (x+a)^p.(y+b)^q BBIm(x,y) \qquad (4)$$

Thus the central moment's mpq or µpq can be computed. Hu defines seven values, computed by normalizing central moments through order three, that are invariant to object scale, position and orientation. Now a 1 × 64 feature vector from surf and 1×7 feature vector from moment invariant are obtained for all the reference (posture/ letter) images and stored in a database.

3.1.7 Classification using K-Nearest Neighbor:

Combined feature vectors of database images are stored in a database. These feature vectors are classified using KNN[11] classifier. Feature vectors of detected hand image from subsequent frame are compared with the stored feature vectors by means of a Euclidean distance measure in KNN[11].

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to weigh the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A shortcoming of the k-NN algorithm is that it is sensitive to the local structure of the data. The algorithm has nothing to do with and is not to be confused with k-means, another popular machine learning technique.

3.1.8 Classification using Support Vector Machines:

Simultaneously the feature vectors of the dataset are given to SVM[15] classifier for training. The basic principle of SVM[15] is to find an optimal separating hyper plane (OSH) which can separate different classes in a feature space, that is, the distances between these classes should be the furthest. To perform the classification between two classes, a nonlinear SVM [15] classifier is applied by mapping the input data (xi,yi) into a higher dimensional feature space using a non-linear operator φ(x).

The OSH can be computed as a decision surface:

$$f(x) = sign(\sum_i \alpha_i y_i K(x_i, x) + b), \qquad 5$$

Where sign () is the sign function and K (xi, x) = φ(xi) τφ(x) is the predefined kernel function. The coefficients αi and b in can be determined by the quadratic problem. This procedure is carried out for the sequence of detected hand from video frames. For each frame classifier recognizes a single letter as output. The results given by both the classifiers are taken as a combined feature vector for gesture classification

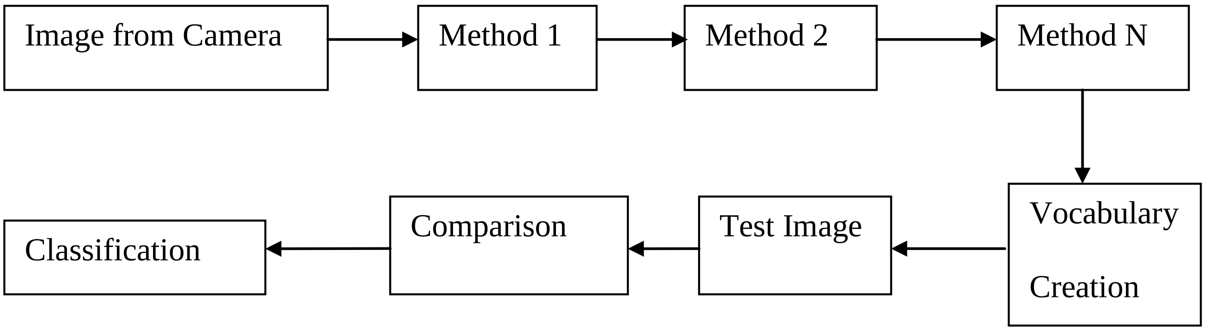
Chapter 4

Cascade of Features for Hand Gesture Recognition

4.1 System Architecture and Implementation:

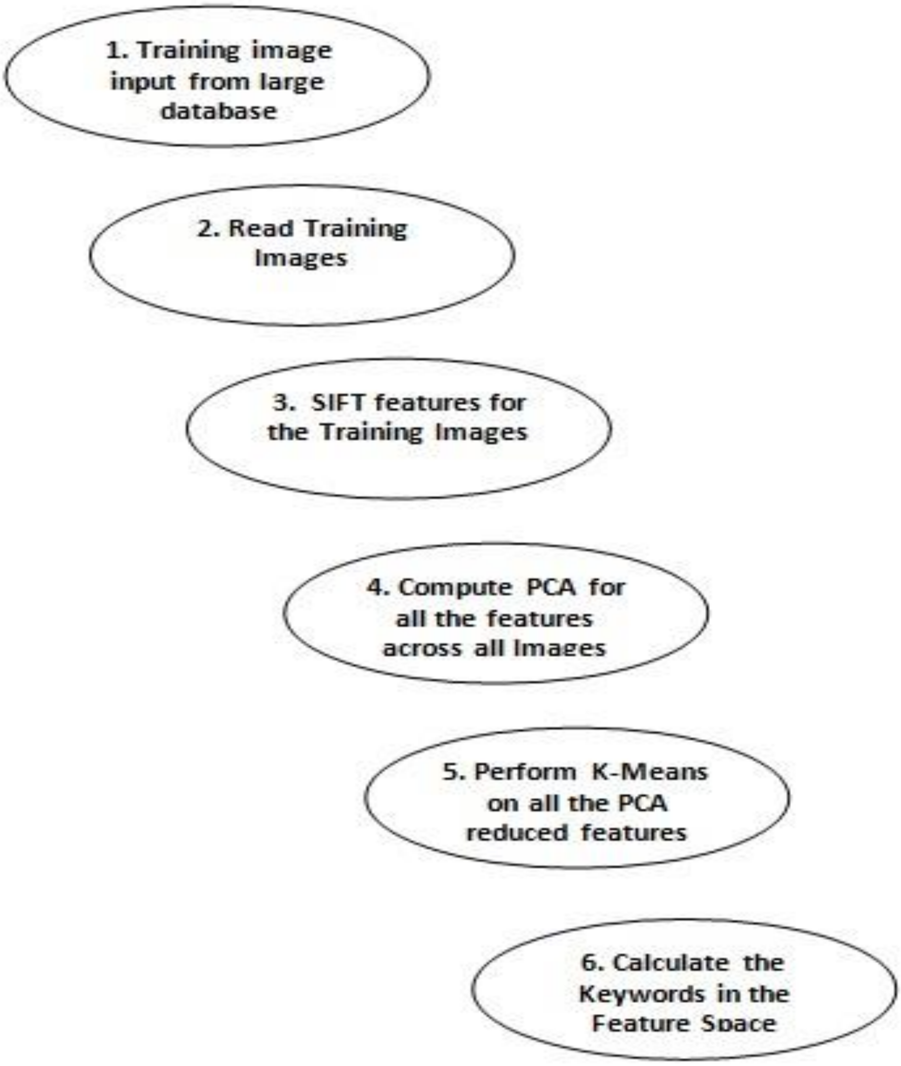
After going through the literature survey and hand gesture recognition system we can clearly see that there are number of methods available. Every method has its own pros and cons. It depends on our variations of input and requirement which method will work better. Starting from Euclidean distance, histogram like basic comparison to Scale invariant feature, Speeded up robust feature comparison.

After studying benefits and limitations of all the method I came to conclusion that existing algorithms can be used in a modified way to achieve good result. I have implemented the cascade of existing method like in circuit theory we use parallel and cascade combination of circuit component to achieve desired value of resistance, capacitance, inductance etc. In the similar way we will use the series of features and method and we will pass our data base images through our system to train it and finally our test data result will be compared with trained data for classification.



There is no limitation on number of methods used in the series as filter to achieve more accuracy but the chosen method should be appropriate and should increase the accuracy of output else these is no benefit of adding one more filter at the cost of increasing the time complexity of method.

The method which we have used in our work is described below:


1. Training image input from large database
2. Read Training Images
3. SIFT features for the Training Images
4. Compute PCA for all the features across all Images
5. Perform K-Means on all the PCA reduced features
6. Calculate the Keywords in the Feature Space

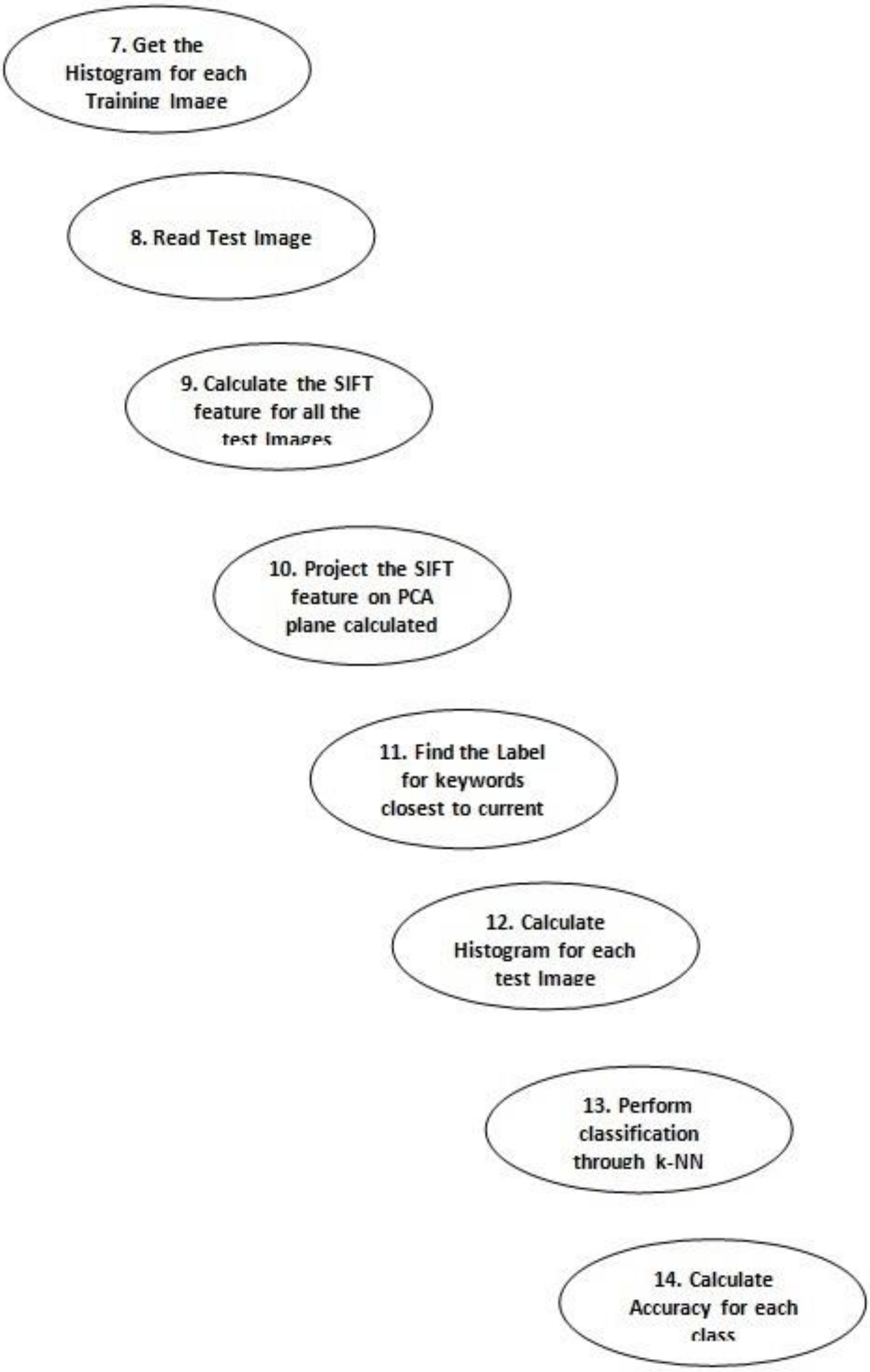

7. Get the Histogram for each Training Image
8. Read Test Image
9. Calculate the SIFT feature for all the test Images
10. Project the SIFT feature on PCA plane calculated
11. Find the Label for keywords closest to current
12. Calculate Histogram for each test Image
13. Perform classification through k-NN
14. Calculate Accuracy for each class

Figure 12: Cascade of features for HGR

1. **Training Image Input from database**: We have used 20 images for each class for training the data. Path for the training images given below :

   HandGestureR\training\Gesture_One\...

   HandGestureR\training\Gesture_Two\...

   HandGestureR\training\Gesture_Three\...

   HandGestureR\training\Gesture_Four\...

   HandGestureR\training\Gesture_Five\...

   We can use any number of images and all the numbers and alphabet class for training of images. To reduce the time taken for execution of cascade of features we have limited the number of images to 20. Images samples are given below:

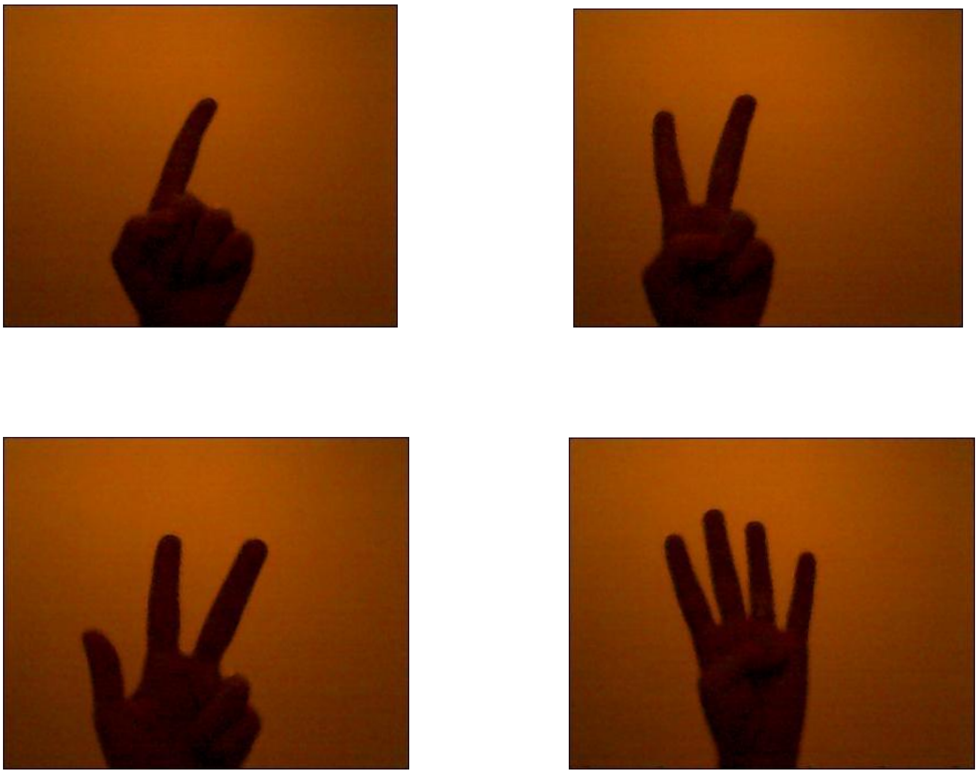


Figure 13: Hand Postures

2. **Read training images**: Using self-defined read_train(Images, name) function which internally calling function imread(file name). In file name parameter I am passing the complete path of the training images. The function imread loads an image from the specified file and returns it. If the image cannot be read (because of missing file, improper permissions, unsupported or invalid format), the function returns an empty matrix (Mat::data==NULL). Currently, the following file formats are supported:
3. **SIFT Features for the training images**: We have used OpenCV library for SIFT features calculation. Using this we will get the key points of our interest for comparison. Using loop instruction calculation of key points of all images have been performed. A SIFT feature is a selected image region (also called keypoint) with an associated descriptor. Keypoints are extracted by the SIFT detector and their descriptors are computed by the SIFT descriptor. It is also common to use independently the SIFT detector (i.e. computing the keypoints without descriptors) or the SIFT descriptor (i.e. computing descriptors of custom keypoints).

   SIFT (Images[i], noArray(), Keypoints[i], Descriptor[i]);

   SIFT detector (Keypoints):

   A SIFT keypoint is a circular image region with an orientation. It is described by a geometric frame of four parameters: the keypoint center coordinates x and y, its scale (the radius of the region), and its orientation (an angle expressed in radians). The SIFT detector uses as keypoints image structures which resemble “blobs”. By searching for blobs at multiple scales and positions, the SIFT detector is invariant (or, more accurately, covariant) to translation, rotations, and re scaling of the image.

The keypoint orientation is also determined from the local image appearance and is covariant to image rotations. Depending on the symmetry of the keypoint appearance, determining the orientation can be ambiguous. In this case, the SIFT detectors returns a list of up to four possible orientations, constructing up to four frames (differing only by their orientation) for each detected image blob.

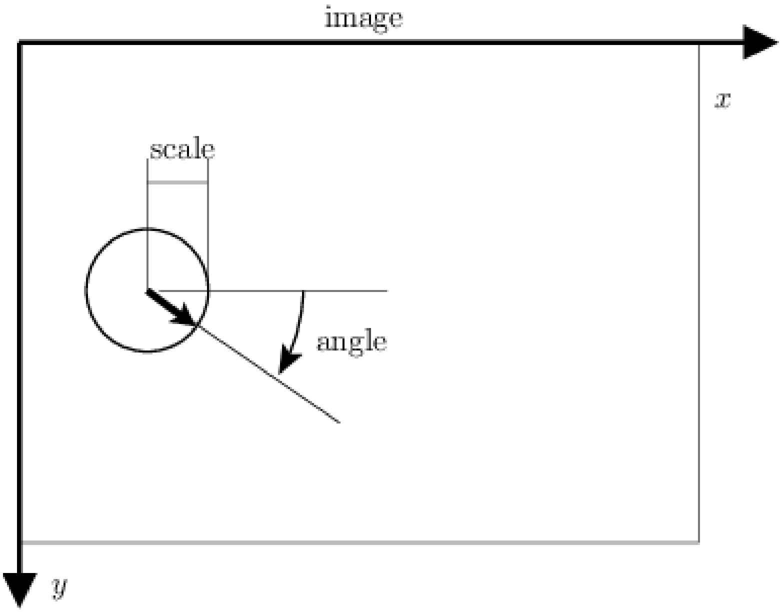


Figure 14: SIFT keypoints are circular image regions with an orientation. [9]

There are several parameters that influence the detection of SIFT keypoints. First, searching keypoints at multiple scales is obtained by constructing a so-called "Gaussian scale space". The scale space is just a collection of images obtained by progressively smoothing the input image, which is analogous to gradually reducing the image resolution. Conventionally, the smoothing level is called scale of the image. The construction of the scale space is influenced by the following parameters, set when creating the SIFT filter object.

Number of octaves:

Increasing the scale by an octave means doubling the size of the smoothing kernel, whose effect is roughly equivalent to halving the image resolution. By default, the scale space spans as many octaves as possible which has the effect of searching keypoints of all possible sizes.

First octave index:

By convention, the octave of index 0 starts with the image full resolution. Specifying an index greater than 0 starts the scale space at a lower resolution (e.g. 1 halves the resolution). Similarly, specifying a negative index starts the scale space at an higher resolution image, and can be useful to extract very small features (since this is obtained by interpolating the input image, it does not make much sense to go past -1).

Number of levels per octave:

Each octave is sampled at this given number of intermediate scales (by default 3). Increasing this number might in principle return more refined keypoints, but in practice can make their selection unstable due to noise (see [1]).

Keypoints are further refined by eliminating those that are likely to be unstable, either because they are selected nearby an image edge, rather than an image blob, or are found on image structures with low contrast. Filtering is controlled by the follow:

Peak threshold: This is the minimum amount of contrast to accept a keypoint. It is set by configuring the SIFT filter.

Edge threshold: This is the edge rejection threshold. It is set by configuring the SIFT filter

SIFT Descriptor:

A SIFT descriptor is a 3-D spatial histogram of the image gradients in characterizing the appearance of a keypoint. The gradient at each pixel is regarded as a sample of a three-dimensional elementary feature vector, formed by the pixel location and the gradient orientation. Samples are weighed by the gradient norm and accumulated in a 3-D histogram h, which (up to normalization and clamping) forms the SIFT descriptor of the region. An additional Gaussian weighting function is applied to give less importance to gradients farther away from the keypoint center. Orientations are quantized into eight bins and the spatial coordinates into four each, as follows:

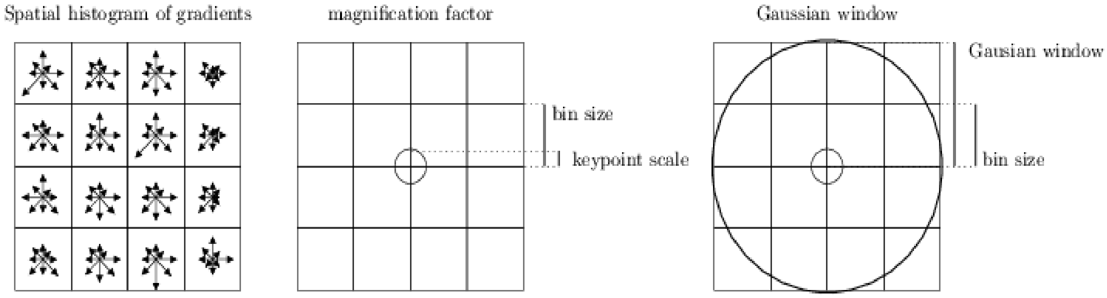


Figure 15: The SIFT descriptor is a spatial histogram of the image gradient [9]

SIFT descriptors are computed by keypoint. They accept as input a keypoint frame, which specifies the descriptor center, its size, and its orientation on the image plane. The following parameters influence the descriptor calculation:

Magnification factor: The descriptor size is determined by multiplying the keypoint scale by this factor.

Gaussian window size: The descriptor support is determined by a Gaussian window, which discounts gradient contributions farther away from the descriptor center.

SIFT descriptor uses the following convention. The y axis points downwards and angles are measured clockwise (to be consistent with the standard image convention). The 3-D histogram (consisting of bins) is stacked as a single 128-dimensional vector, where the fastest varying dimension is the orientation and the slowest the y spatial coordinate. This is illustrated by the following figure.

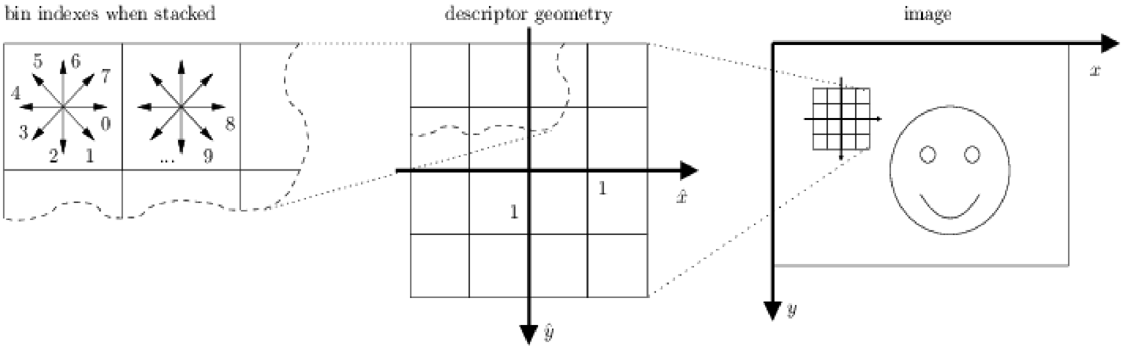


Figure 16: SIFT Descriptor conventions [9]

NOTE:

Keypoints (frames) D. Lowe's SIFT implementation convention is slightly different: The y axis points upwards and the angles are measured counter-clockwise.

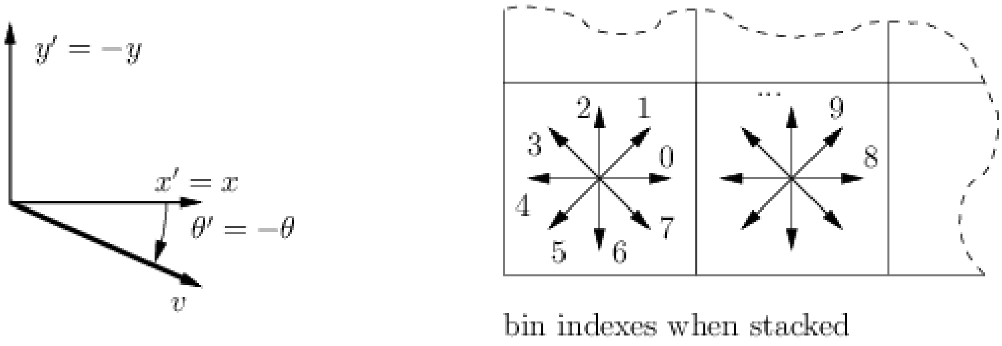


Figure 17: SIFT implementation conventions [9]

4. **Compute PCA for all the features:** Principal Components Analysis (PCA[9]) is technique recommended when there is a large amount of numeric variables and it is desired to find a lower number of artificial variables, or principal components that will be responsible for the higher variance in the observed variables. Then, these principal components can be used as predictor variables in subsequent analyses.

   PCA[9] basically receives an n x m matrix, denoted as M, which represents the actual number of dimensions and the number of feature vectors, respectively. The first step is to obtain a mean vector for each dimension, denoted as mn. Then, mn is substracted from every feature vector in M. Later, we calculate the M x MT covariance matrix. Subsequently, as every covariance matrix is square, in this case n x n, we can calculate the n eigenvalues with their corresponding n-dimensional eigenvectors. Finally, as a higher eigenvalue represents a higher quantity of information, each eigenvector is ordered according to the value of its corresponding eigenvalue, from higher to lower to obtain the kernel PCA[9] matrix, of n x n dimensions denoted as PM. Each row in PM represents an eigenvector. Then, if we have any x x n data matrix, denoted as DM, in which x is the number of n-dimensional vectors, we can reduce their dimensions by projecting them over the first desired features from each vector in PM.

```
pca.project (full_Descriptor, Pcafeature);
```

5. **Perform K-Means on the reduced feature:** k-means clustering is a method of vector quantization; originally from signal processing that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares.

The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the k-means algorithm; it is also referred to as Lloyd's algorithm, particularly in the computer science community.

Given an initial set of k means m1(1),…,mk(1) (see below), the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean.

Update step: Calculate the new means to be the centroids of the observations in the new clusters.

The algorithm has converged when the assignments no longer change. Since both steps optimize the WCSS objective, and there only exists a finite number of such partitioning, the algorithm must converge to an optimum. There is no guarantee that the global optimum is found using this algorithm.

6. **Calculate keyword in the feature space**: In this step we have created one dictionary for each of training image. It’s like creating one code book of feature space based on the center of each cluster.
7. **Histogram of training image**: Construction of histogram from the created codebook of feature space.

   From steps 8 to 12 same operation will get repeated for test image data and continued till Histogram step.

13. Perform classification through Knn Classifier**:** In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression.[1] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

    - In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
    - In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

14. **Calculate Accuracy for each class**: It will output the accuracy of test data with predicted value.

    Implementation of the method described above has been completed and the result is as given below:

4.2 Result Analysis:

We have used different data set of images to calculate the robustness of the algorithm and calculated the accuracy percentage of classification of images:

4.2.1 Result with correct classification

Data Set 1:

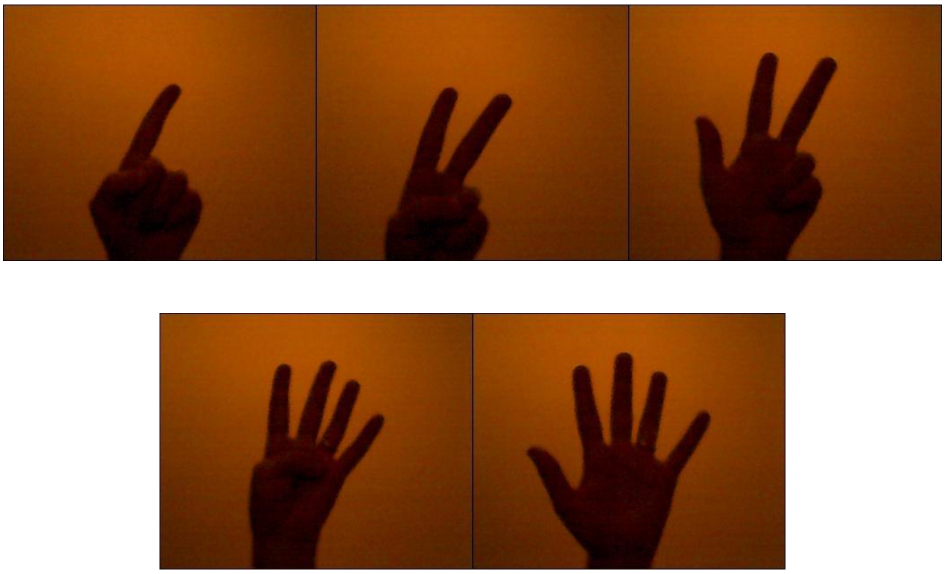
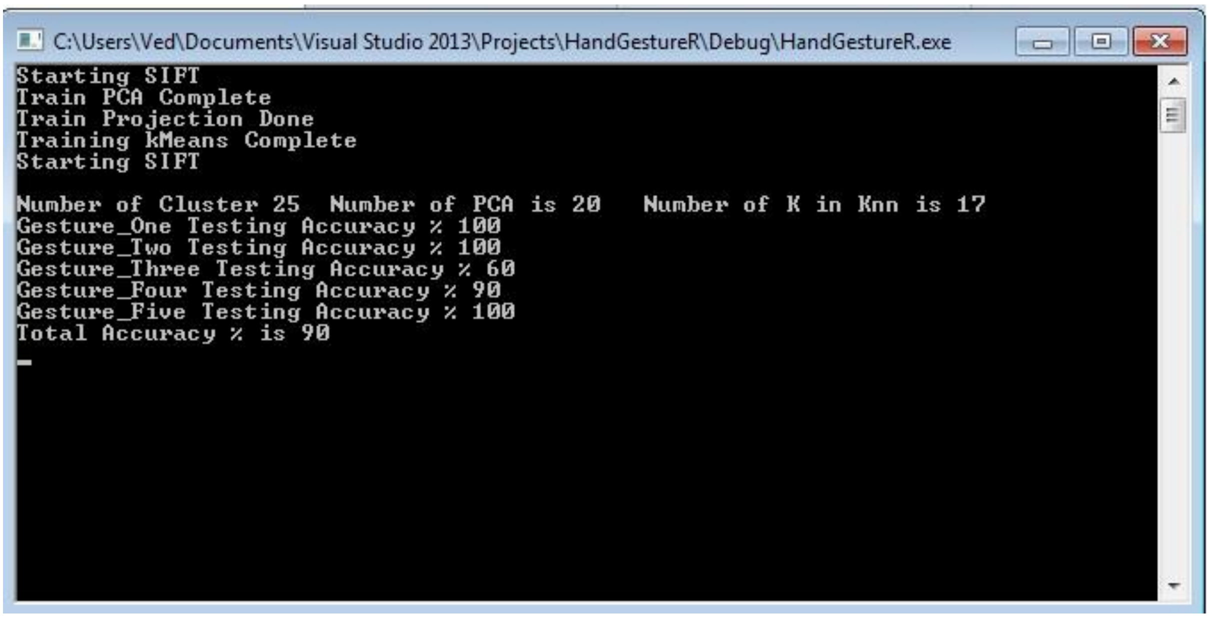


Figure 18: Different hand postures


C:\Users\Ved\Documents\Visual Studio 2013\Projects\HandGestureR\Debug\HandGestureR.exe

```
Starting SIFT
Train PCA Complete
Train Projection Done
Training kMeans Complete
Starting SIFT

Number of Cluster 25  Number of PCA is 20   Number of K in Knn is 17
Gesture_One Testing Accuracy % 100
Gesture_Two Testing Accuracy % 100
Gesture_Three Testing Accuracy % 60
Gesture_Four Testing Accuracy % 90
Gesture_Five Testing Accuracy % 100
Total Accuracy % is 90
_
```

Data Set 2:

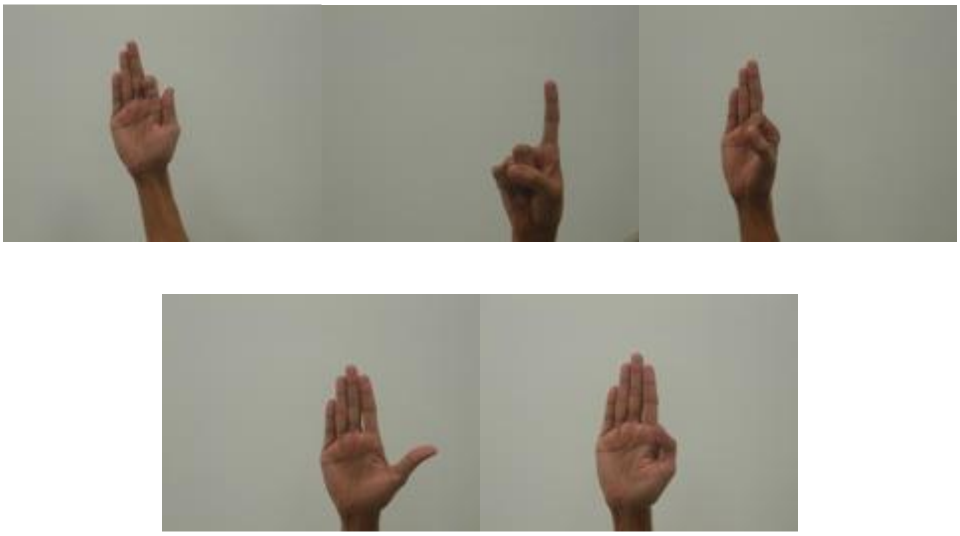
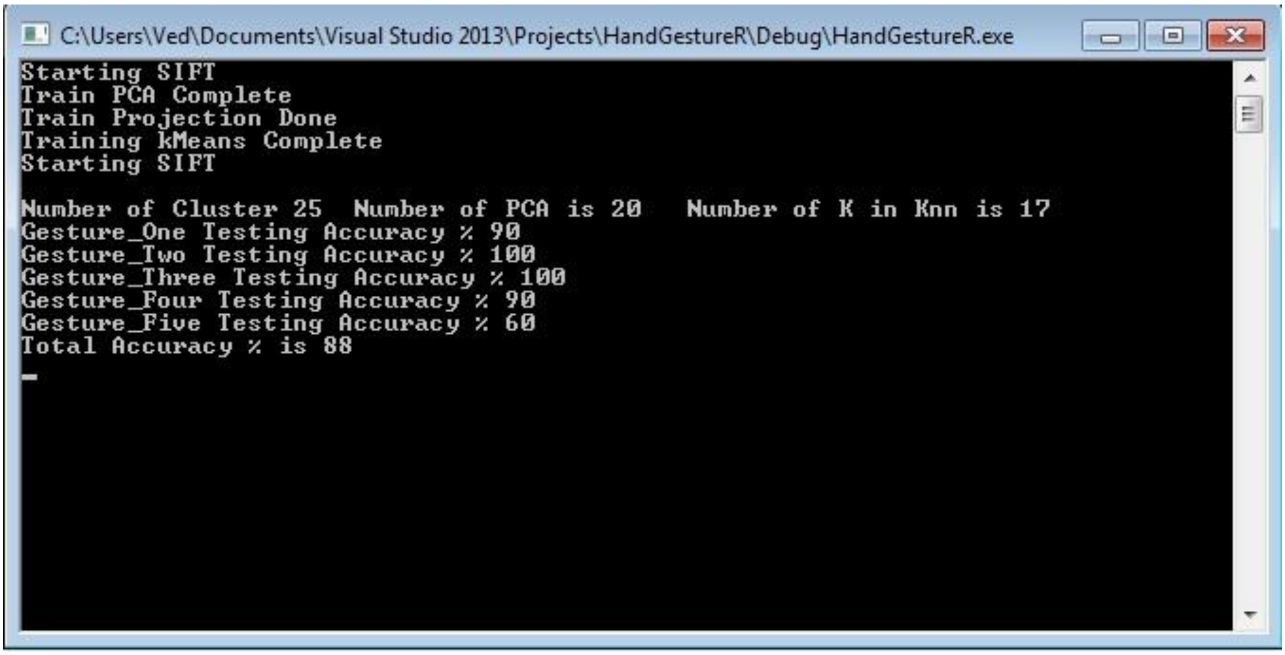


Figure 19: Different hand postures



C:\Users\Ved\Documents\Visual Studio 2013\Projects\HandGestureR\Debug\HandGestureR.exe

```
Starting SIFT
Train PCA Complete
Train Projection Done
Training kMeans Complete
Starting SIFT

Number of Cluster 25  Number of PCA is 20   Number of K in Knn is 17
Gesture_One Testing Accuracy % 90
Gesture_Two Testing Accuracy % 100
Gesture_Three Testing Accuracy % 100
Gesture_Four Testing Accuracy % 90
Gesture_Five Testing Accuracy % 60
Total Accuracy % is 88
_
```

Data Set 3:

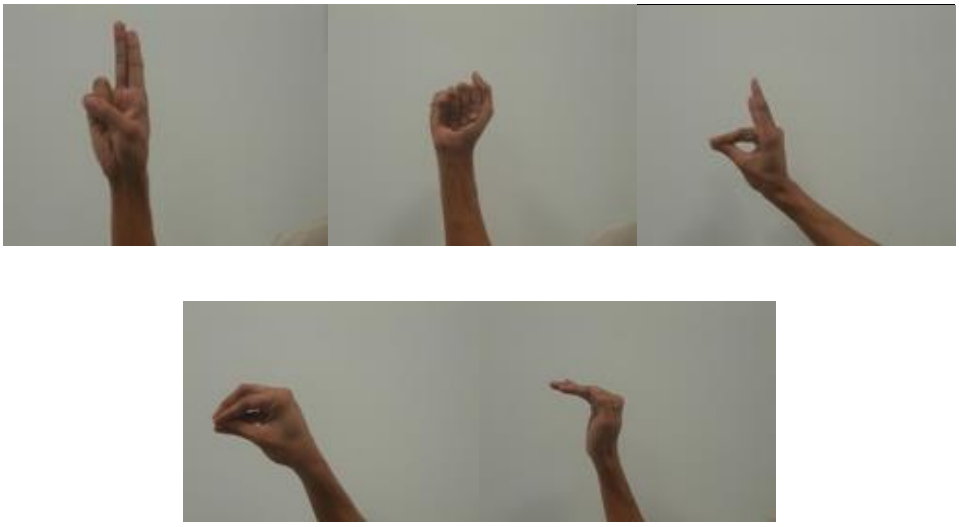
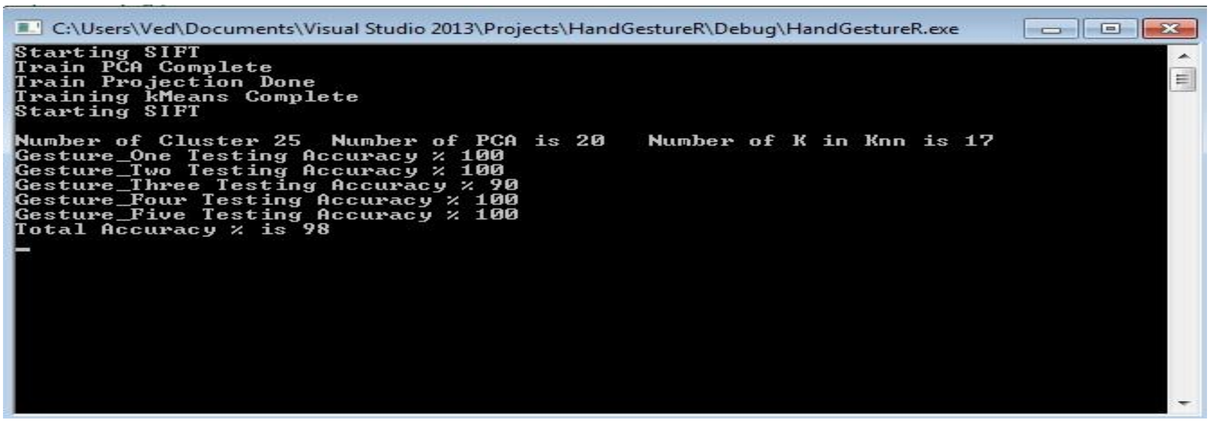


Figure 20: Different hand postures



C:\Users\Ved\Documents\Visual Studio 2013\Projects\HandGestureR\Debug\HandGestureR.exe

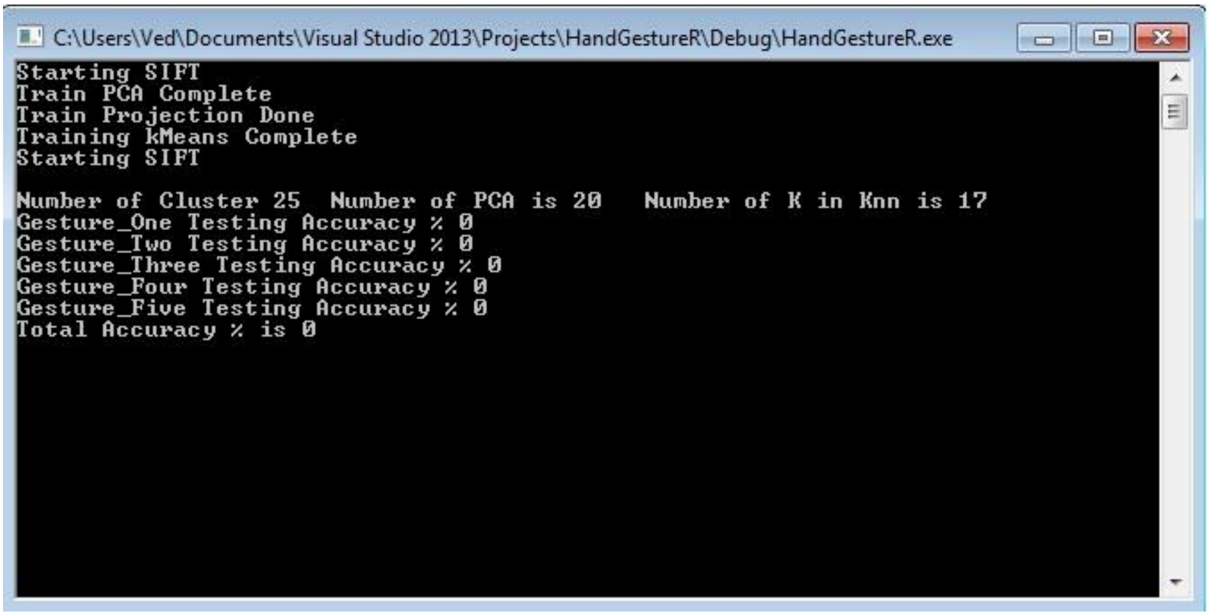
```
Starting SIFT
Train PCA Complete
Train Projection Done
Training kMeans Complete
Starting SIFT

Number of Cluster 25  Number of PCA is 20   Number of K in Knn is 17
Gesture_One Testing Accuracy % 100
Gesture_Two Testing Accuracy % 100
Gesture_Three Testing Accuracy % 90
Gesture_Four Testing Accuracy % 100
Gesture_Five Testing Accuracy % 100
Total Accuracy % is 98
_
```

Table 2: Result with correct classification

Data Set	Gesture 1	Gesture 2	Gesture 3	Gesture 4	Gesture 5	Avg.
1	100	100	60	90	100	90
2	90	100	100	90	60	88
3	100	100	90	100	100	98

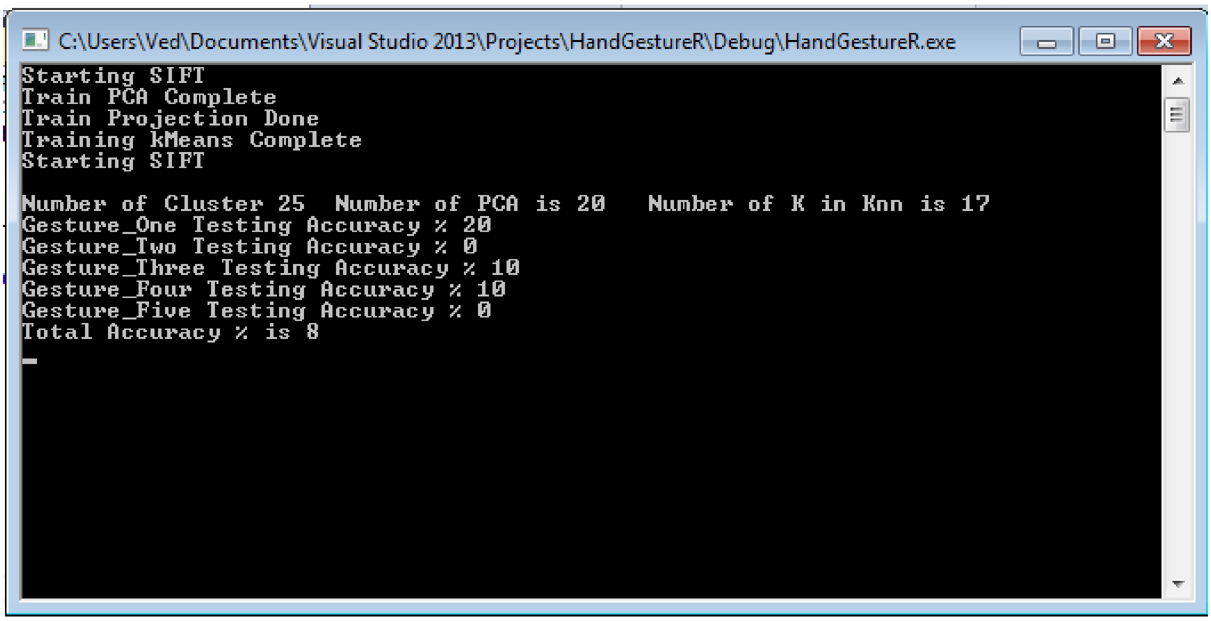
4.2.2 Result with wrong classification

Data Set 1:


C:\Users\Ved\Documents\Visual Studio 2013\Projects\HandGestureR\Debug\HandGestureR.exe

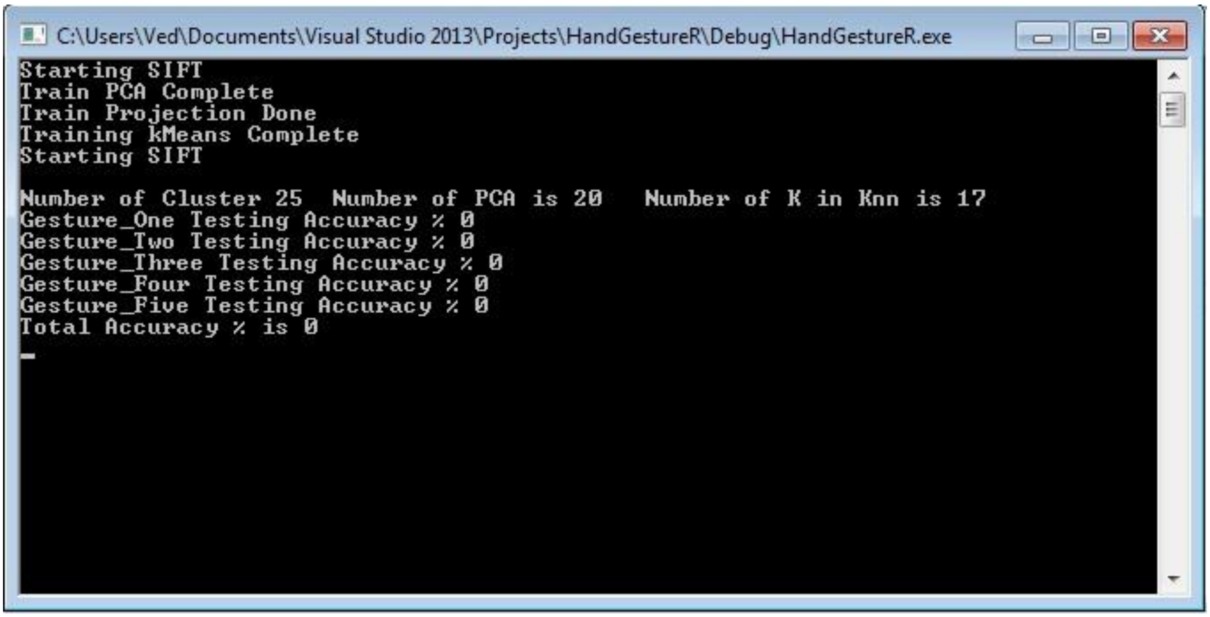
```
Starting SIFT
Train PCA Complete
Train Projection Done
Training kMeans Complete
Starting SIFT

Number of Cluster 25  Number of PCA is 20   Number of K in Knn is 17
Gesture_One Testing Accuracy % 0
Gesture_Two Testing Accuracy % 0
Gesture_Three Testing Accuracy % 0
Gesture_Four Testing Accuracy % 0
Gesture_Five Testing Accuracy % 0
Total Accuracy % is 0
```

Data Set 2:


C:\Users\Ved\Documents\Visual Studio 2013\Projects\HandGestureR\Debug\HandGestureR.exe

```
Starting SIFT
Train PCA Complete
Train Projection Done
Training kMeans Complete
Starting SIFT

Number of Cluster 25  Number of PCA is 20   Number of K in Knn is 17
Gesture_One Testing Accuracy % 20
Gesture_Two Testing Accuracy % 0
Gesture_Three Testing Accuracy % 10
Gesture_Four Testing Accuracy % 10
Gesture_Five Testing Accuracy % 0
Total Accuracy % is 8
```

Data Set 3:



C:\Users\Ved\Documents\Visual Studio 2013\Projects\HandGestureR\Debug\HandGestureR.exe

```
Starting SIFT
Train PCA Complete
Train Projection Done
Training kMeans Complete
Starting SIFT

Number of Cluster 25  Number of PCA is 20   Number of K in Knn is 17
Gesture_One Testing Accuracy % 0
Gesture_Two Testing Accuracy % 0
Gesture_Three Testing Accuracy % 0
Gesture_Four Testing Accuracy % 0
Gesture_Five Testing Accuracy % 0
Total Accuracy % is 0
_
```

Table 3: Result with wrong classification

Data Set	Gesture 1	Gesture 2	Gesture 3	Gesture 4	Gesture 5	Avg.
1	0	0	0	0	0	0
2	20	0	10	10	0	8
3	0	0	0	0	0	0

From the result we can conclude that the cascade of feature comparison method is effective to detect the hand gesture recognition. The accuracy of method in case of correct classification of testing data is more than 90% and in case of wrong classification it is almost 0%. Average accuracy of method for three data set is 92%.

Chapter 5

Future Proposed Improvement: Integration of Color Recognition Technique for HGR

5.1 System Architecture and Implementation:

We have discussed about how we can improve the accuracy of recognition technique using cascading of different method for features comparison. The earlier method discussed was based on the extracted features from images stored in data base or taken from video frame. We have proposed one major improvement in method and working on integration of color recognition technique with the above discussed method. The system will become like:

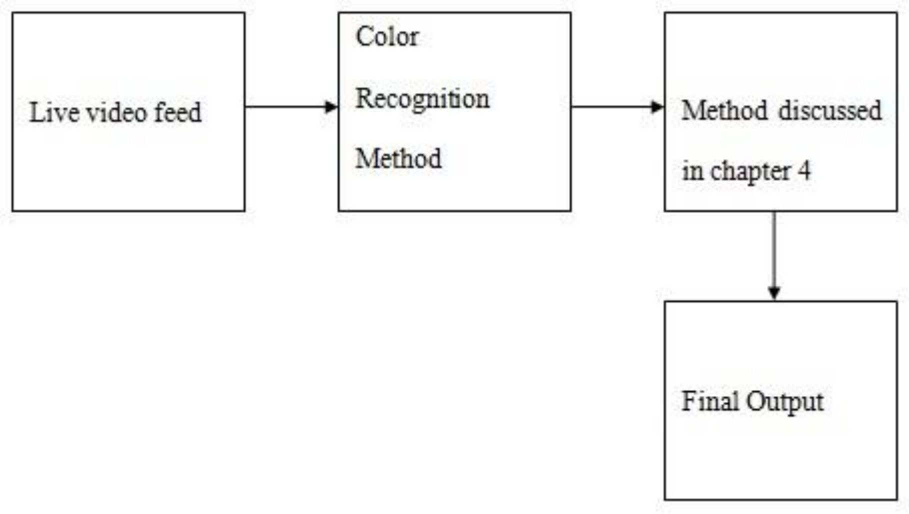


Figure 21: Proposed improvement using Color Recognition

Color Recognition Technique (CRT[2]): It is basic framework for hand gesture recognition. This is based on the color detection of hand skin and then the position of fingers to detect it using convex method. The basic flow of system is as given:

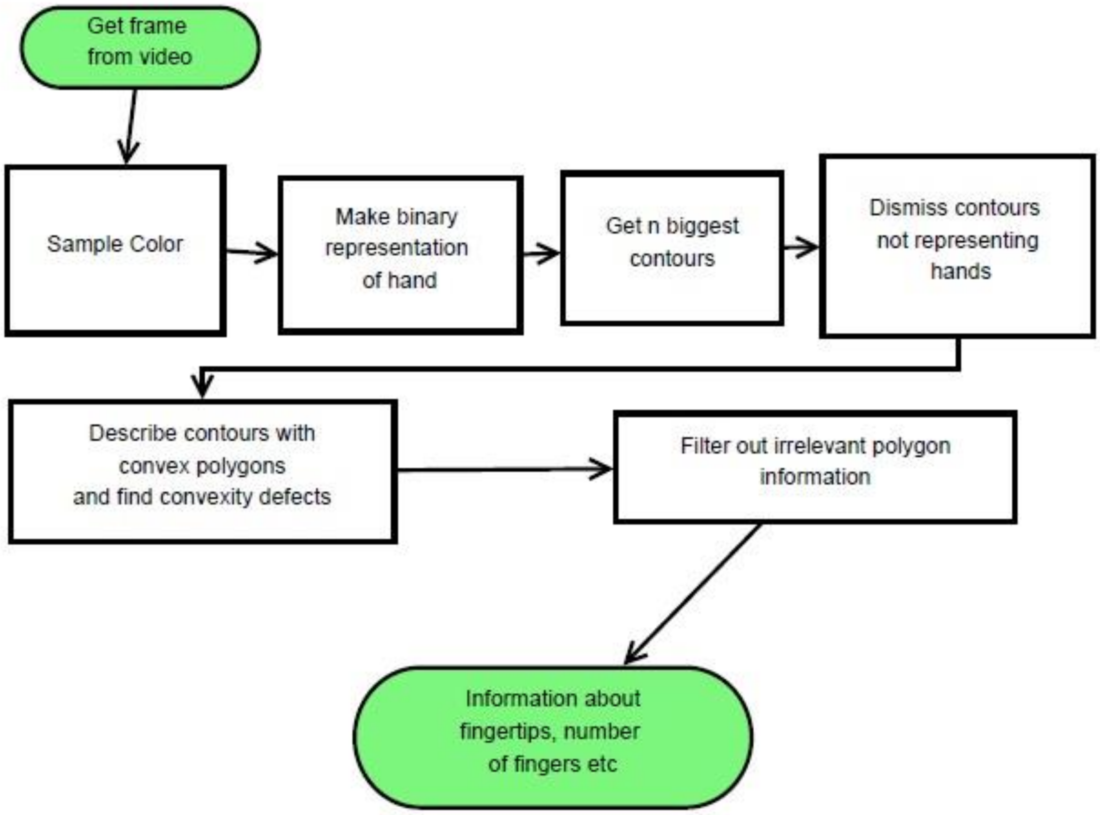

Get frame from video
Sample Color
Make binary representation of hand
Get n biggest contours
Dismiss contours not representing hands
Describe contours with convex polygons and find convexity defects
Filter out irrelevant polygon information
Information about fingertips, number of fingers etc

Figure 22: Basic flow of CRT [4]

It created a color profile of hand based on median color values from different areas of the hand.

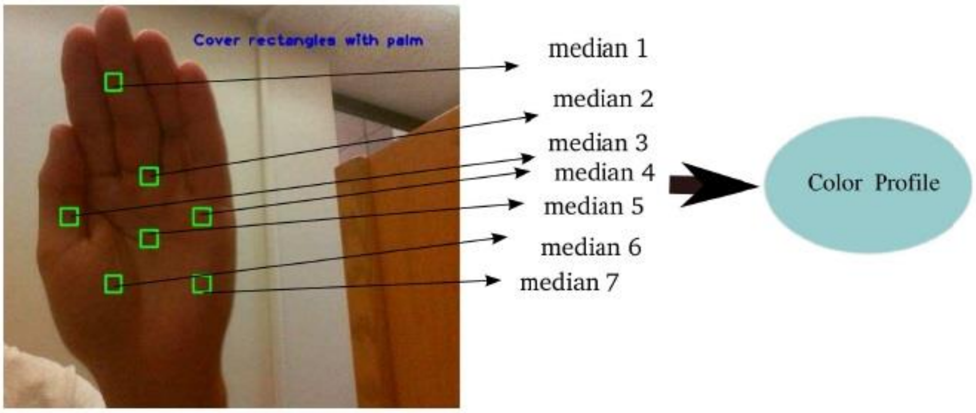


Figure 23: Making a color profile of hand [4]

Extract hand based on color recognition. Compute one binary image based on each sampled median and sum the binary images together. Finally filter the result with the nonlinear median blur filter.

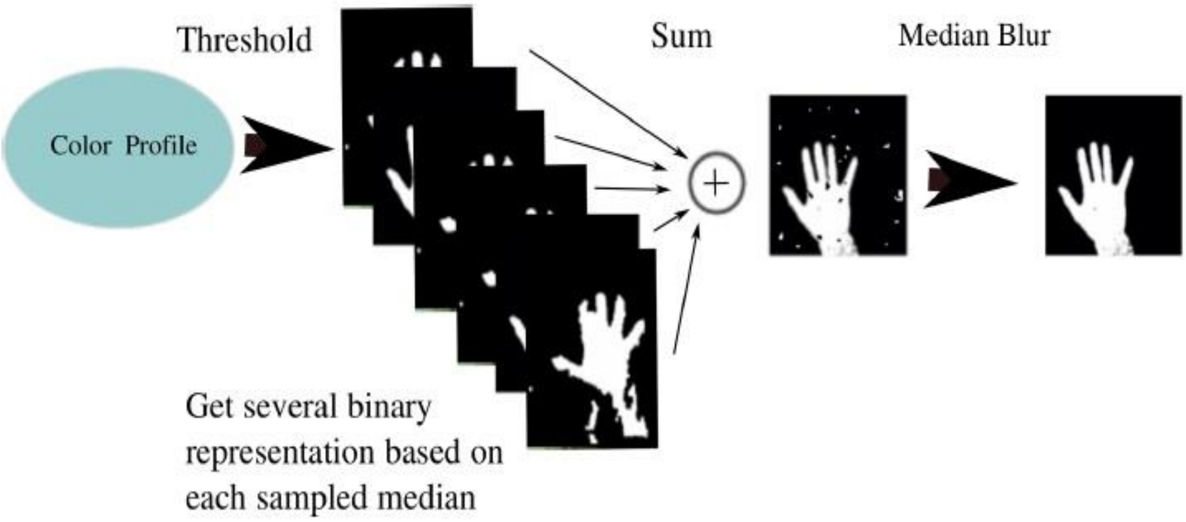

Threshold
Sum
Median Blur
Color Profile
Get several binary representation based on each sampled median

Figure 24: Binary representation [4]

Geometric approach to detect the convex point and convexity defects is shown below:

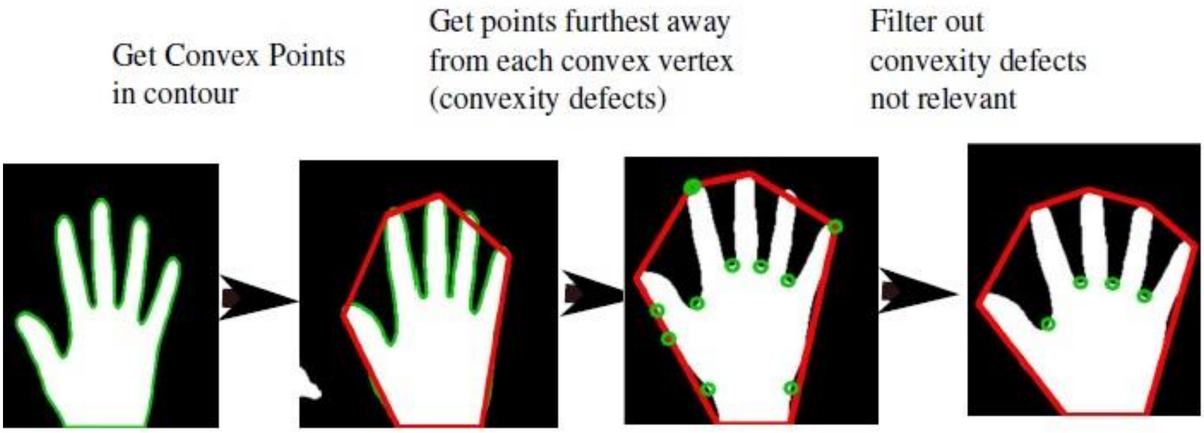

Get Convex Points in contour
Get points furthest away from each convex vertex (convexity defects)
Filter out convexity defects not relevant

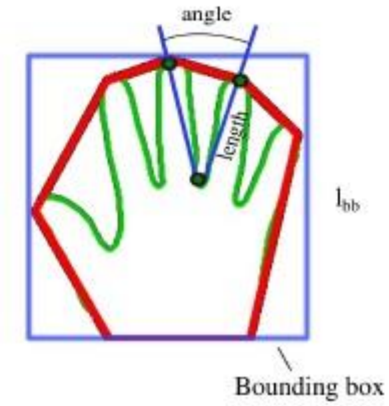
Figure 25: Geometric Approach [4]

The property that determines whether a convexity defect is to be dismissed is the angle between the lines going from the defect to the convex polygon vertices.

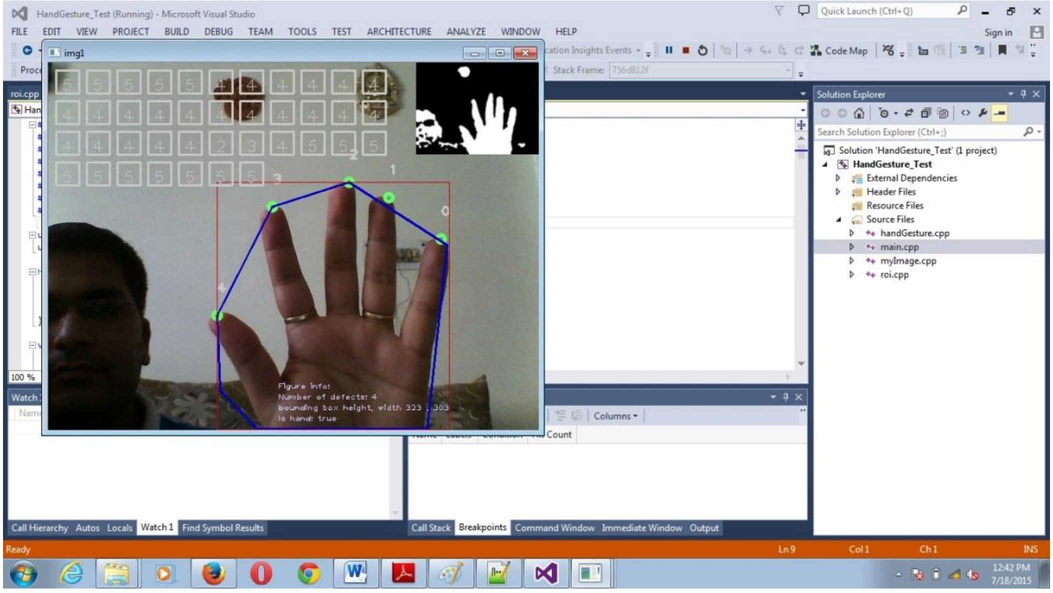
Dismiss convexity defect if:

Length < 0.4lbb

Angle > 80 Degree

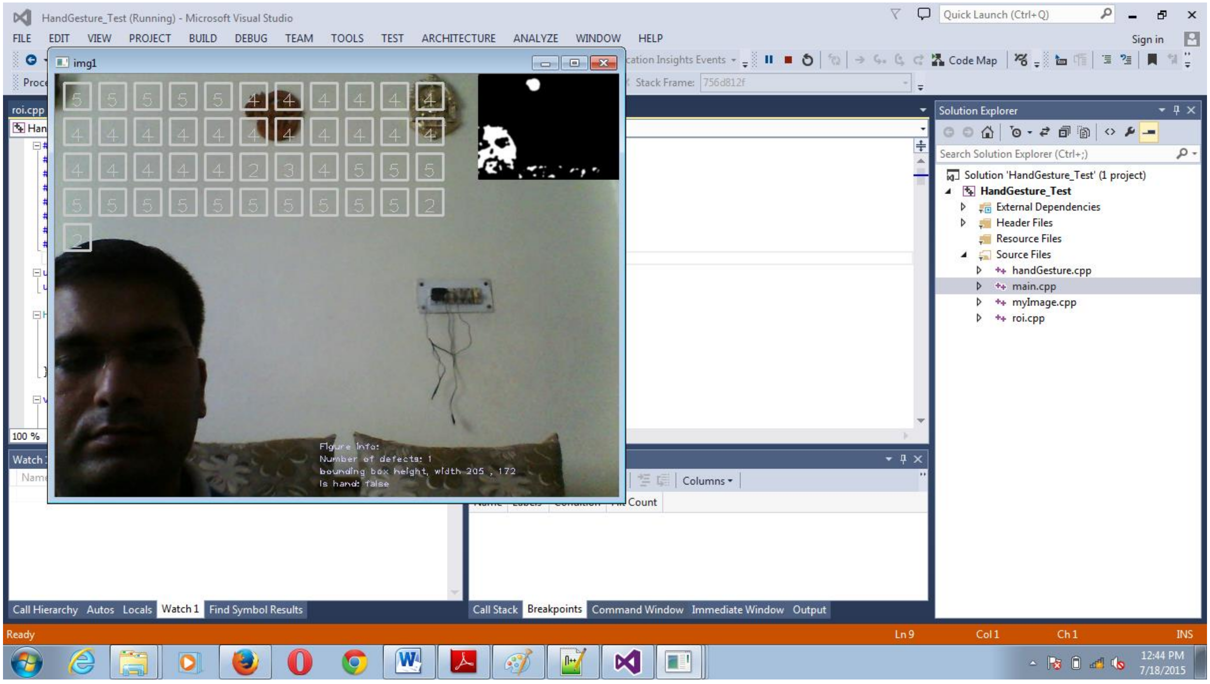


Result of Color Recognition Technique:


HandGesture_Test (Running) - Microsoft Visual Studio
FILE EDIT VIEW PROJECT BUILD DEBUG TEAM TOOLS TEST ARCHITECTURE ANALYZE WINDOW HELP
Quick Launch (Ctrl+Q)
Sign in
img1
Code Map
Stack Frame:
Solution Explorer
Search Solution Explorer (Ctrl+;)
Solution 'HandGesture_Test' (1 project)
HandGesture_Test
External Dependencies
Header Files
Resource Files
Source Files
handGesture.cpp
main.cpp
myImage.cpp
roi.cpp
Columns
Count
Call Hierarchy Autos Locals Watch 1 Find Symbol Results
Call Stack Breakpoints Command Window Immediate Window Output
Ready
Ln 9 Col 1 Ch 1 INS
12:42 PM
7/18/2015

```
Figure info:
Number of defects: 4
bounding box height, width 323 , 303
is hand: true
```

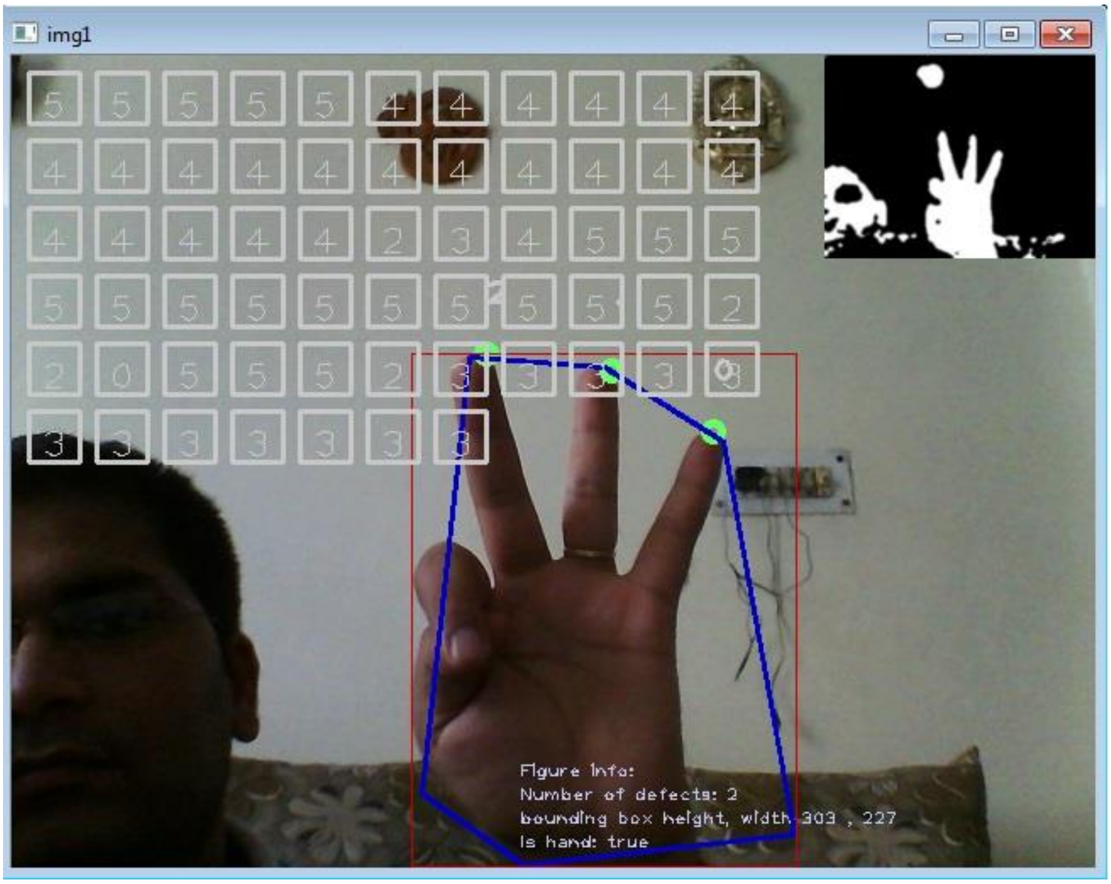
Figure 26: Snapshot of CRT Result



```
Figure Info:
Number of defects: 1
bounding box height, width 205 , 172
Is hand: false
```

HandGesture_Test (Running) - Microsoft Visual Studio
FILE EDIT VIEW PROJECT BUILD DEBUG TEAM TOOLS TEST ARCHITECTURE ANALYZE WINDOW HELP
img1
Solution Explorer
Solution 'HandGesture_Test' (1 project)
HandGesture_Test
External Dependencies
Header Files
Resource Files
Source Files
handGesture.cpp
main.cpp
myImage.cpp
roi.cpp
Call Hierarchy Autos Locals Watch 1 Find Symbol Results
Call Stack Breakpoints Command Window Immediate Window Output
Ready

Figure 27: CRT Stops processing once it finds no hand in video frame



```
Figure Info:
Number of defects: 2
bounding box height, width 303 , 227
Is hand: true
```

Figure 28: Number 3 recognition

5.2 Result Analysis:

Table 4:

CRT Experiment carried out @ Natural Day Light at 10 AM

Exp No.	Total number of frame	True	False
1	55	51	4
2	55	53	2
3	55	51	4
4	55	51	4
5	55	54	1
6	55	49	6
7	55	55	0
8	55	52	3
9	55	52	3
10	55	50	5

Total No. of Exp: 550 Number under observation: 3

True Positive: 518 False Positive: 32

Accuracy % = 94.18

Table 5:

CRT Experiment carried out @ Bad Light Condition at 6:45 PM

Exp No.	Total number of frame	True Positive	False Positive
1	55	31	24
2	55	23	32
3	55	36	19
4	55	13	42
5	55	43	12
6	55	48	7
7	55	43	12
8	55	29	26
9	55	40	15
10	55	18	37

Total No. of Exp: 550 Number under observation: 3

True Positive: 324 False Positive: 226

Accuracy % = 58.9

Above results are under some specific conditions and it may vary depending on the variations of conditions.

We can clearly see the variations of result due to many preconditions. Light conditions play a vital role in CRT. We should have adequate light source to apply this technique in real time system. Limitations of CRT technique given below:

1. Overlapping objects
2. No adaptive learning
3. Noise and light conditions sensitive
4. Camera sensitive

Although this technique is having known limitations but once it will get integrated with the feature comparison methods it can be a great success in terms of real time data processing and increasing the accuracy to achieve the desired output so that we can further modify it for commercialization use.

Chapter 6

Conclusion

It is observed from the experimental results that cascade of features method is robust against multi variations like rotation, scale, lighting and view-point and provides good real time performance. Use of derived features from available feature set along with SIFT & Principal component analysis makes the approach highly robust against multiple variations and shows consistent real time performance with improved processing speed. The tradeoff between accuracy and speed of processing can be maintained by creating dictionary of features in xml file so that the training data can be stored effectively and it can be made adaptive in future. The CRT method is under development and we hope the result will be really remarkable and can be a milestone in hand gesture recognition technique.

References

1. Hand Gesture Recognition for Sign Language: A New Hybrid Approach : J.Rekha1, J.Bhattacharya2 and S.Majumder2
2. Alphabet Recognition of American Sign Language: A Hand Gesture Recognition Approach using SIFT Algorithm: Nachamai. M, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.4, No.1, January 2013
3. Analysis and Implementation of Hand Gesture Recognition System: Poonam Verma, Utkarsh Mor, Puneet Singh, Arpan Khandelwal, International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online) Vol. 3, Issue 1, pp: (119-122), Month: January - March 2015
4. A Review on Feature Extraction for Indian and American Sign Language: Neelam K. Gilorkar, Manisha M. Ingle, International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 314-318
5. Hybrid Method for Hand Gesture Recognition: Aneer P Imunny, International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 5, May 2015
6. REAL TIME HAND GESTURE RECOGNITION USING SIFT: Pallavi Gurjal, Kiran Kunnur, International Journal of Electronics and Electrical Engineering
7. HAND GESTURE RECOGNITION: A LITERATURE REVIEW: Khan Noor Adnan Ibraheem, International Journal of Artificial Intelligence & Applications (IJAIA)
8. Indian Sign Language Recognition System for Deaf People: Arti Thorat, Varsha Satpute, Arati Nehe, Tejashri Atre Yogesh R Ngargoje, International Journal of

Advanced Research in Computer and Communication Engineering Vol. 3, Issue 3, March 2014

9. Modified Sift Algorithm for Appearance Based Recognition of American Sign Language: Jaspreet Kaur ,Navjot Kaur, IJCSET |May 2012| Vol 2, Issue 5,1197-1202
10. Vision Based Hand Gesture Recognition: A Review: G. Simion, V. Gui, and M. Otesteanu, INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING