International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 03, March-2015

A Novel Approach in Hand Gesture Recognition for Sign Language

U. P. K. Jayamini[1], D. K. Withanage[2]
Faculty of Information Technology,
University of Moratuwa, Sri Lanka

***Abstract* — This paper presents an enhanced algorithmic approach for recognizing the symbols of the American Sign Language to reduce the communication gap between the hearing impaired people and the rest of the world. In sign language, each gesture has an assigned meaning (or meanings). This research mainly focuses on depicting the corresponding symbols of the American Sign Language (ASL) mapped to hand gesture. Our approach is based on image pre-processing, image database generation and key point comparison techniques, to detect the corresponding hand gesture through computer vision and machine learning approaches. This could pave way to achieve the goal of implementing an effective system in order to shrink the communication gap between the normal and the hearing impaired people. We've used MATLAB as the programming platform for this research. The Scale Invariant Feature Extraction (SIFT) algorithm, Nearest Neighbor search and new concepts have been used to as an effective and convenient way to process hand gestures and recognize them accurately. This new approach brings more benefits compared to existing approaches and related research concepts.**

Keywords* — *Hearing Impairments, Sign Language, Quality of Match, Gesture Recognition, Image Processing, Nearest Neighbor Search

I. INTRODUCTION

As Hellen Keller once said, "Deafness separates people from people" the deaf become separated by the rest of the world due to their impairments. Approximately one out of 1,000 to 2,000 babies is born deaf [3,4]. The ones who are having hearing impairments are isolated from the rest of the world due to their disability. Sign language that comprises of visually transmitted sign patterns to convey meaning simultaneously combining hand shapes, orientation and movement of hands and facial expressions are used by most of the deaf in their communication. Communication hurdle between the deaf and the hearing community is transpired owing to the ignorance of each other's languages. The emergence of an improved methodology in interpreting sign language into text and voice will be supportive to both hearing and non-hearing community in their interaction to mitigate the communication gap.

Sign language is comprised of structured sets of gestures where each gesture has an assigned meaning (or meanings). ASL (American Sign Language) is the language choice of most of the deaf people [4]. The main purpose of ASL is to allow deaf people communicate with normal people. Interpreting ASL into voice and text with the aid of vision based systems will be beneficial for the deaf to interrelate with others in their day today activities.
The vast majority of hand gesture recognition work use mechanical sensing, most often for direct manipulation of a virtual environment and occasionally for symbolic communication. However, sensing the hand posture mechanically has a range of problems, including reliability, accuracy and electromagnetic interference. Visual sensing has the potential to make gestural interaction more practical, but potentially embodies some of the most difficult problems in machine vision. The hand is a non-rigid object and even becomes worse due to self-occlusion.

This research focuses on a novel concept of recognition system for converting sign language into text/voice for the comprehension of users. Visual biased analysis application is formed to perform Hand Gesture Recognition of American Sign Language (ASL) in that regard. In this research, the nearest neighbor approach is applied by using SIFT descriptors, points' locations, and the technique named as distance calculation is used to find out the ratio of validity to obtain the best match of the gesture. By considering the theory of Quality of Match, the ratio of validity is computed in iterations. Thus the best match can be determined in order to obtain the best output.

The rest of this paper is organized as follows. The section 2 discusses about the requirement of a sign language interpreter to support the hearing impaired community followed by the brief description about the sign language in section3. In section 4 we review the related approaches for sing language recognition. The section 5 presents our approach for hand gesture recognition. A detailed description of the methodology is presented in section 6. The section 7 carries out the evaluation and the discussion of the proposed approach followed by the conclusion in section 8.

II. SIGN LANGAUGE INTERPRETER

For a large portion of the deaf community, English (spoken or written) is not their first language and therefore, they experience the same language issues experienced by any member of a linguistic minority. For many, English is only their second or third language [3]. Due to the issues of written communication (e.g. delay due to the unfamiliarity of language, inability to convey ideas adequately under pressure and potential linguistic issues) the use of written language to

communicate in an environment where effective communication is required, is not recommended.

Lip reading can also be an issue for many Deaf people as many English words use the same lip movements when speaking. It is also not uncommon for individuals who are deaf to have problems with their eyesight [1]. Further most of the time the deaf lacks the capability of writing the spoken languages. Thus when they need to convey their ideas and feelings to the world, they become isolated since the others are lacking comprehending their sign language. When a deaf person signs, the hand gesture recognition system will render the meaning expressed in the signs into the spoken language for the hearing party, which is sometimes referred to as voice interpreting or voicing. The duty of this system for sign language is to build a bridge to link both world of the deaf and the rest of the world by eliminating the language barrier.

III. The American Sign Language

American Sign Language (ASL) is a complete, complex language that employs signs made by moving the hands combined with facial expressions and postures of the body [5]. It is the primary language of many North Americans who are deaf and is one of the several communication options used by people who are deaf or hard-of-hearing. It is a visual language. With signing, the brain processes linguistic information through the eyes. The shape, placement, and movement of the hands, as well as facial expressions and body movements, play important parts in conveying information.

ASL consists of approximately 6000 gestures of common words with finger spelling which are used to communicate proper nouns. Finger spelling that's focused on this research can be performed by one hand and 26 gestures to communicate the 26 letters of the alphabet. Sign language is not a universal language, each country has its own sign language, and regions have dialects, much like the many languages spoken all over the world [2,4]. Like any spoken language, ASL is a language with its own unique rules of grammar and syntax. ASL is a living language that grows and changes over time. In spoken language, voice of the words is produced by using the mouth to make sounds. But for people who are deaf (particularly those who are profoundly deaf), the sound of speech is often not heard, and only a fraction of the speech is visible on the lips. Sign languages are based on the idea that vision is the most useful tool that a deaf person has to communicate and receive information.

IV. Review on Sign Language Recongition Systems

Ways of performing hand gesture recognition can mainly be classified under several approaches. One of them is heavily based on hardware such as glove based analysis, employ sensors (mechanical or optical) attached to a glove that transduces finger flexion into electrical signals to determine the hand posture. Normally, the sensors that are acoustic or magnetic are embedded into the glove. For an instance there is a system able to translate Japanese Sign Language (JSL) and Japanese vise versa. The system operates by recognition of one-handed motions [11,12]. A VPL (Visual Programming Language) Data Glove Model II is used for acquiring hand data. It has two sensors for measuring bending angles of the two joints on each finger, one over the knuckle and the other over the middle joint of the finger. There is also a sensor attached to the back of the glove which measures 3 position data and 3 orientation data relative to the fixed magnetic source [11]. The positional data is calibrated by subtracting the neutral positional data from the raw positional data.

Under the neural network approach which is operated by feeding numerous types of hand gestures images into 'neural network' and network itself is trained by the system. Once the 'neural network' is trained, multiple of hand gesture recognitions of ASL can be performed by this 'neural network' [9,10]. But this system is not safe and robust enough. Moreover, when back similar image from results is used and applied to the system again, it differs from the test result. This occurs since the initial weights and bias from the system is not identical, therefore, each time the system runs the test, result is not the same [9]. This approach is not robust and safe enough since during network retrain, as the results are not guaranteed to be equivalent. Furthermore, there are a lot factors such as number of layers and number of neurons need to be considered.

Another approach is analysis of drawing gesture, which involves the use of special input devices such as stylus. Most of the hand gesture recognition systems currently work using mechanical sensing, most often for direct manipulation of a virtual environment. But this type of sensing has a range of problems such as accuracy, reliability and electromagnetic interference. These two categories involve external hardware devices [11,12]. The third approach is the vision based analysis which is based on the way of perceiving information from the surrounding by human beings. Visual sensing has the potential to make gestural interaction more practical and reliable. This type of method is an intuitive method to perform hand gesture recognition since it doesn't involve external hardware devices, where our hand gestures can be recognized freely. In this case only a camera, webcam, camcorder or any device that can capture images, which can be interfaced with a computer, is required. In this research vision based analysis is focused.

V. Proposed Hand Gesture Recognition System

A system of manual, facial, and other body movements as the means of communication is used within Sign language, especially among deaf people [2]. The deaf is familiar with their sign language and if it's converted into text or voice which is the first language of others, it will be more beneficial and applicable in the real world. In our approach, ASL (American Sign Language) has been chosen as the sign language interpretation. In this hand gesture system, the conversion of sign language into text/voice is focused for the meanings of words depicted only by hand gestures within the broad scope of sign language. Further, finger spelling segment in sign language scope is expressly targeted.

Within our system, SIFT key point extraction is done with regard to the descriptor. For example an image stored in database and query image descriptors perform dot product to find the best matches (nearest neighbor search). Then center point calculation is done to determine the ratio of validity to reach our target of finding the best match. In addition, the ratio of validity, quality of match is also taken into

consideration in order to improve the accuracy by using number of key points. After the identification of data base image from the corresponding ACII value of the sign letter is formed.

A. SIFT algorithm

The SIFT algorithm (Scale Invariant Feature Transform) proposed by Lowe is an approach for extracting distinctive invariant features from images. It has been successfully applied to a variety of computer vision problems based on feature matching including object recognition, pose estimation, image retrieval and many others. Following are the major stages of computation used to generate the set of image features [6,7],

- Scale space-extrema detection
- Key point localization
- Orientation assignment
- Key point descriptor

Image key point descriptor is the fundamental cell in this project. It is the major output parameter that gets out after applying SIFT algorithm. Once such descriptors have been generated for more than one image, one can begin image matching techniques. Therefore basic techniques depend on this key point descriptor extraction. A key point is an image feature which is so distinct that image scaling, noise, or rotation does not, or rather should not, distort the key point itself. A key point descriptor is a 128-dimensional vector that describes a key point. The reason for this high dimension is that each key point descriptor contains a lot of information about the point it describes [6,8]. By using this algorithm locale, image, descriptor and number of key points (to find the quality match) of every image can be output to find the best match with ratio calculation.

B. Image pre-processing

It's frequently the case that images which are to be offered to a computer vision system are already pre-recorded. This is particularly true in virtually all real world images which are used for research studies into computer vision. These input images are therefore pre-sampled at discrete spatial and temporal intervals, may include patterning due to photographic grain, photo-detector noise etc. Converting web cam captured RGB images to gray-scale is done by eliminating the hue and saturation information while retaining the luminance. To reduce image processing time, the number of key points should be reduced. This is achieved by reducing the image resolution and converting training images to the portable gray map (PGM) format. Moreover, lighting compensation, extracting skin, removing noise and finding skin color blocks are applied respectively as image pre-processing techniques.

C. MATLAB as the environment

MATLAB is not only a programming language, but also a programming environment. The Hand Gesture Recognition research project is developed using MATLAB software. Problems involved with vector and matrix formulations are optimally done by using MATLAB since its basic data element is an array which does not require dimensioning [13]. Therefore, in this research, an image descriptor (matrix) is extracted for further calculations as MATLAB environment is beneficial to be used. The dot product calculations of searches are calculated and even array, matrices manipulations are done simply and accurately. Furthermore, MATLAB's enhanced toolboxes (comprehensive collections of MATLAB functions) can be applied throughout the research implementation in an appropriate manner.

VI. PROCEDURE

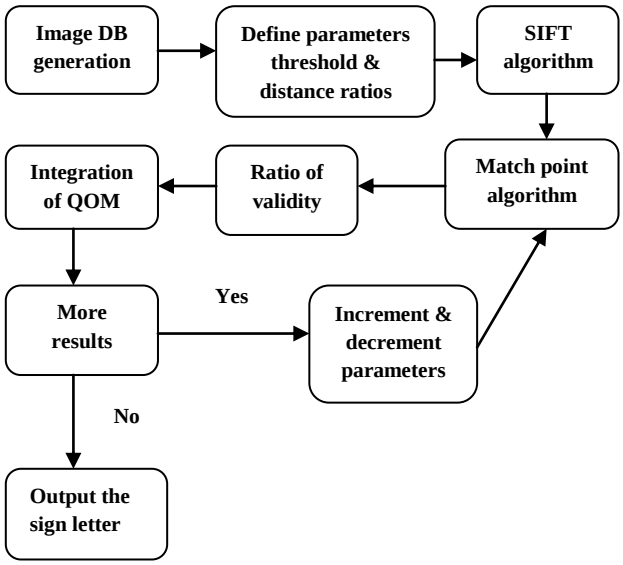

Figure:1 Algorithmic approach for the procedure

Before extracting the descriptors, key point locations, image and number of key points image pre-processing techniques like de-blurring, reading ,converting into gray-scale and skin extraction (to remove unnecessary objects) should be applied to images in database as well as to web cam captured real time images for point comparison. When the image file is input, with the aid of SIFT algorithm, following will be output [1,8],

- Descriptors

A, K-by-128 matrix, where each row gives an invariant descriptor for one of the K key points. The descriptor is a vector of 128 values normalized to unit length.

- Location of key point

K-by-4 matrix, in which each row has the 4 values for a key point location (row, column, scale, orientation).The orientation is in the range [-PI, PI] radians.

- Image

The image array in double precision

- Number of Key points (K)

For performing quality of match, number of key points is needed per each image

The critical parameters like ratio of distance for comparing methods should be defined at the beginning. Even default threshold and their increment values should be assigned. Images (their numbers) are stored in an array. Alphabet letter is identified for each captured image by adding 64 to

array elements to represent the equivalent ASCII letter. Figure 2 depicts the training data set.

A. *Matching approach*

Data base image, image captured from the web cam, output of SIFT algorithm and pre-defined ratio of distance are used here in order to find out the values of the following,

- Matched key point locations of the database image
- Matched key point locations of the web cam captured image
- X position of the database image's center point
- Y position of the database image's center point
- X position of the query image's center point
- Y position of the query image's center point
- Number of matched key points

To find this information modification of nearest neighbor search is used. It is convenient to compute dot products of unit vectors rather than computing Euclidean distances in MATLAB i.e. ratio of angles (acos of dot products of unit vectors) is a close approximation to the ratio of Euclidean distances for small angles. For each descriptor in the first image, selection is made to match the corresponding second image by using this dot product comparison. If nearest neighbor has angle less than ratio of distance (initially this is a constant value, but when next iterations are taken placed, this is changed and treated as a variable value), it will be treated as a matched point.

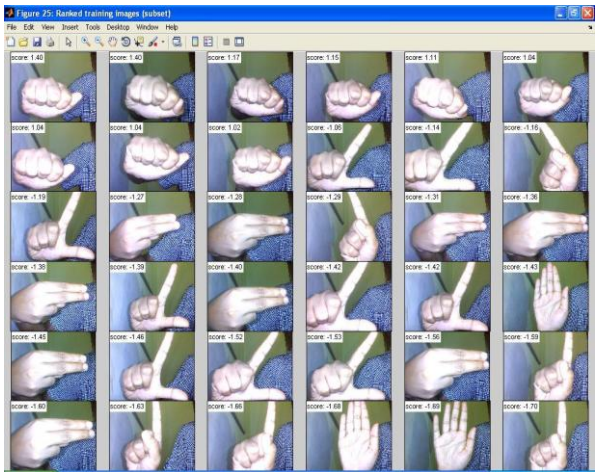

Figure 2: The training dataset

B. *Resulting algorithm*

After the matching algorithm is processed, its output is passed to this newly defined algorithm to calculate the distances of the matched key points to the center of the key points using Euclidean distance computation. For the data base image and query image, those distances are calculated for matched points chosen before [3]. These distances are stored in arrays to compute the sum of distances for further computation.

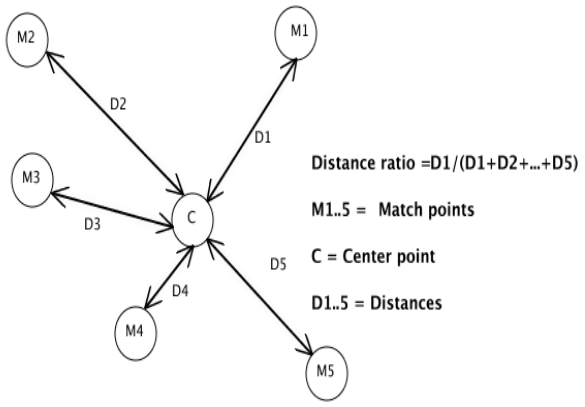

Figure 3: Calculating distance ratios

For each key point the distance ratio is calculated in order to compare the difference with the database images. If the difference is less than the pre-defined threshold, it will be considered as a valid matched key point. Under the process, calculate the validity ratio of the total number of key points simply by dividing the total number of valid matched key points by the matched key points. Then store the ratio inside the results array. Finding the highest value of ratio will be the next step of the algorithm. In this case three best matching images are computed repeatedly while incrementing the ratio of distance and decrementing the threshold value (Figure 5). With the experimental results, it's observed that granting the maximum value of validity ratio alone can create inaccurate results. To avoid that erroneous state, the concept of quality of match is integrated with the validity ratio calculation.

C. *Quality of match*

Let Ks be the number of key points in the source image, Kc be the number of key points in the compared image, and Km be the number of matching key points.

- Formula : $(Km*Kc)/Ks^2*100$

This formula now takes into account the ratio as well as the percentage of matches in the source image.

D. *Process in brief*

- Image pre-processing.
- Database generation to depict alphabetic letters.
- Following the SIFT algorithm and returning descriptor, locale, image and number of key points.
- Pre-computing those parameters for database images and storing in three different .mat files.
- Nearest neighbor search - Computing dot products between unit vectors rather than Euclidean distances
- Matching approach for data base and input image with pre-defined ratio of distance to output matched key point locations of the database image, matched key point locations of the images captured by the web cam.
- Calculating the validity of ratio
- Integration with Quality of Match
- Find the best match that has the highest value

E. Way of initiation

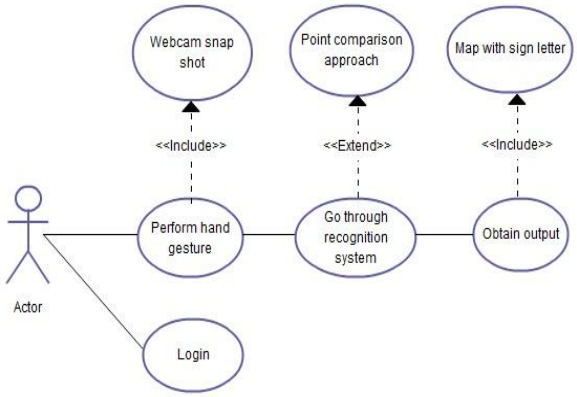

Figure 4: High level use case diagram of the system

With reference to the research implementation training images are stored, for each letter in finger-spelling sign alphabet. Under the SIFT the number of key points, descriptors and locales are pre-computed and saved in four separate .mat files to improve efficiency of the process. When a user logs into the system, he performs the gesture in front of web cam in real time, there a snap shot is taken and processed in the system. Real time generated snap shot images' key point parameters are generated in real time and compared with the stored pre-computed ones and find the best match according to the process discussed previously.

VII. Evaluation and Discussion

Within this approach, image points are compared and it recognizes ASL input (query) images by comparing it with the database images and outputs the equivalent ASCII representation of it. Large numbers of features can be extracted from typical images with efficient algorithms. In addition, like in other approaches SIFT key points are used since the features are highly distinctive, which allows a single feature to be correctly matched with high probability against a large database of features, providing a basis for object and scene recognition. According to the methodology for image matching and recognition, SIFT features are first extracted from a set of reference images and stored in a database.

A new image is matched by comparing each feature from the new image to this stored database and finding candidate matching features based on Euclidean distance and ratios of distances of their feature vectors (Figure 6). The concept behind this approach is matching key points independently to the database of key points extracted from training images. In existing solutions the best candidate match for each key point is found only by identifying its nearest neighbor in the database of key points from training images (only by using quality of match). The nearest neighbor is defined as the key point with minimum Euclidean distance for the invariant descriptor vector.

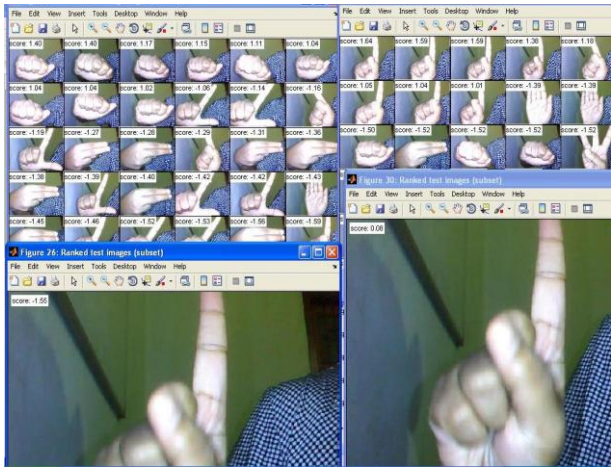

Figure 5: Real time hand gesture recognition

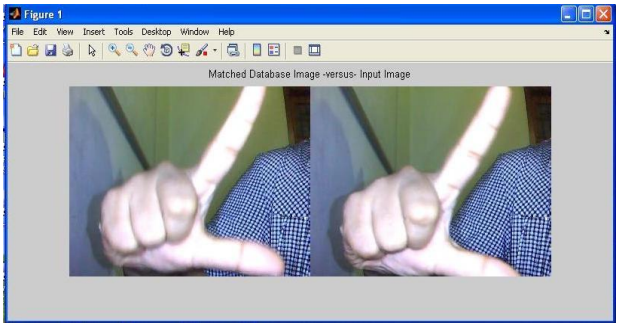

Figure 6: The corresponding database image for a given hand gesture

In this system, the ratio of validity is clarified in order to obtain the best matched image in database. The algorithm used here is newly developed with the aid of some prevailing concepts and the main target is to identify the best match through the key points of an image. In most of the prevailing solutions, vision sensors are used to generate more sharp images. But here only a 4mp web cam is used and appropriate technologies are applied without the expensive use of hardware. For each key point, the key point descriptor is created by sampling the magnitudes and orientations of the image gradients in the N*N neighboring region around the point. This research implementation is done without using any sophisticated tools or hardware and anyone can own this with trust since it's very simple and accurate.

VIII. Conclusion

Hand gestures provide a natural intuitive communication modality for human computer interaction. Efficient human computer interfaces (HCIs) have to be developed to allow computers to visually recognize real time hand gestures. However, vision-based hand tracking and gesture recognition is a challenging problem due to the complexity of hand gestures, which are rich in diversities due to high degrees of freedom (DOF) involved by the human hand. In order to successfully fulfill their role, the hand gesture HCIs have to meet the requirements in terms of real-time performance, recognition accuracy, and robustness against transformations and cluttered background [14]. To meet these requirements, many gesture recognition systems obtain the help of colored markers or data gloves to make

the task convenient. In this paper, we focus on bare hand gesture recognition without the help of any markers and gloves. SIFT features, proposed by Lowe, are features (key points) extracted from images to help in reliable matching between different views of the same object, image classification, and object recognition.

To keep this system cost minimum and to give everyone the opportunity to own and use this application easily, only a computer equipped with a web cam is required as hardware. Since this system is very limited in hardware, the programming part done with the use of MATLAB becomes more prominent. MATLAB is chosen since it is perfect for speeding up development process in which it allows user to work faster and concentrate on the results rather on the design of the programming. According to the discussed approach the query image is compared with the data base images in real time. For the image captured from the web cam, the feature key points are detected and further computation is done for the points in real time. Then they are compared with the pre-computed values for the data base images. With this system, approximately six sign letters can be recognized by using minimum hardware and higher efficiency gained by algorithmic approaches.

REFERENCES

1. William T. Freeman, M. R. (1994, 12). Orientation Histograms for Hand Gesture. p. 8.
2. *How Many People Are Born Deaf?* (n.d.). Retrieved 12 2011, from eHow health: http://www.ehow.com/facts_5199696_many-people-born-deaf_.html.
3. Carter-Johnson, N. (n.d.). *Kwintessential*. Retrieved from Sign language interpreters: http://www.kwintessential.co.uk/BSL/sign-language-interpreter.html.
4. Perlmutter, David M. "The Language of the Deaf." New York Review of Books, March 28,1991, pp. 65-72
5. Sign Language. (n.d.). Retrieved from The free dictionary: http://www.thefreedictionary.com
6. S. Se, D. Lowe, and J. Little. Global localization using distinctive visual features. In Proceedings of the International Conference on Intelligent Robots and Systems, pages 226– 231, 2002.
7. Y. Ke, R. Sukthankar: PCA-SIFT: A more distinctive representation for local image descriptors. In: Proc. CVPR. Volume 2. (2004) 506–513
8. Bakken, T. (2007, 08 16). An evaluation of the SIFT algorithm for. *An evaluation of the SIFT* , p. 39
9. Thian, A. T. (2008, 09 21). *Artificial Intelligence Project based on Neural Network Concept*. Retrieved from Hand Gesture Recognition Using Neural Network: http://angtzuthian.webs.com/
10. Symeonidis, K. (2000, 10 23). Hand Gesture Recognition Using Neural Networks. p. 68
11. Ganzeboom, M. (2009, 08 12). How hand gestures are recognized using a dataglove. p. 8
12. Cyber Glove Systems. Cyber Glove Systems website. Retrieved April 8, 2009 from http://www.cyberglovesystems.com/, 2009
13. Mathworks. (n.d.). *1*. Retrieved from MATLAB: http://www.mathworks.in/products/matlab/
14. Burande, C. (26-28 Feb. 2010). Advanced recognition techniques for human computer interaction., (pp. 480 - 483).