**Abstract:**

In our society, we have people with disabilities. The technology is developing day by day but no significant developments are undertaken for the betterment of these people. Communications between deaf-mute and a normal person has always been a challenging task. It is very difficult for mute people to convey their message to normal people. Since normal people are not trained on hand sign language. In emergency times conveying their message is very difficult. The human hand has remained a popular choice to convey information in situations where other forms like speech cannot be used. Voice Conversion System with Hand Gesture Recognition and translation will be very useful to have a proper conversation between a normal person and an impaired person in any language.The project aims to develop a system that converts the sign language into a human hearing voice in the desired language to convey a message to normal people, as well as convert speech into understandable sign language for the deaf and dumb. We are making use of a convolution neural network to create a model that is trained on different hand gestures. An app is built which uses this model. This app enables deaf and dumb people to convey their information using signs which get converted to human-understandable language and speech is given as outpu

## CHAPTER 1

## INTRODUCTION:

### 1.1 Human Computer Interaction:

Human–computer interaction (HCI) involves the study, planning, design and uses of the interaction between people (users) and computers. It is often regarded as the intersection of computer science, behavioral sciences, design, media studies, and several other fields of study. Human–computer interaction (HCI) involves the study, planning, design and uses of the interaction between people (users) and computers. It is often regarded as the intersection of computer science, behavioral sciences, design, media studies, and several other fields of study. Humans interact with computers in many ways, and the interface between humans and the computers they use is crucial to facilitating this interaction. Desktop applications, internet browsers, handheld computers, and computer kiosks make use of the prevalent graphical user interfaces (GUI) of today. Voice user interfaces (VUI) are used for speech recognition and synthesizing systems, and the emerging multi-modal and gestalt User Interfaces (GUI) allow

humans to engage with embodied character agents in a way that cannot be achieved with other interface paradigms.

The Association for Computing Machinery defines human-computer interaction as "a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them". An important facet of HCI is the securing of user satisfaction (or simply End User Computing Satisfaction). "Because human–computer interaction studies a human and a machine in communication, it draws from supporting knowledge on both the machine and the human side. On the machine side, techniques in computer graphics, operating systems, programming languages, and development environments are relevant. On the human side, communication theory, graphic and industrial design disciplines, linguistics, social sciences, cognitive psychology, social psychology, and human factors such as computer user satisfaction are relevant. And, of course, engineering and design methods are relevant." Due to the multidisciplinary nature of HCI, people with different backgrounds contribute to its success. HCI is also sometimes referred to as human–machine interaction (HMI), man–machine interaction (MMI) or computer–human interaction (CHI). A UCLA 2014 study of sixth graders and their use of screen-devices found a lack of face-to-face contact deprived the youngsters of emotional cues including facial expressions and body language.

Poorly designed human-machine interfaces can lead to many unexpected problems. A classic example of this is the Three Mile Island accident, a nuclear meltdown accident, where investigations concluded that the design of the human–machine interface was at least partially responsible for the disaster. Similarly, accidents in aviation have resulted from manufacturers' decisions to use non-standard flight instrument or throttle quadrant layouts: even though the new designs were proposed to be superior in regards to basic human–machine interaction, pilots had already ingrained the "standard" layout and thus the conceptually good idea actually had undesirable results.

HCI (Human Computer Interaction) aims to improve the interactions between users and computers by making computers more usable and receptive to users' needs. Specifically, HCI has interests in: methodologies and processes for designing interfaces (i.e., given a task and a class of users, design the best possible interface within given constraints, optimizing for a desired

property such as learn ability or efficiency of use) methods for implementing interfaces (e.g. software toolkits and libraries) techniques for evaluating and comparing interfaces developing new interfaces and interaction techniques developing descriptive and predictive models and theories of interaction A long term goal of HCI is to design systems that minimize the barrier between the human's mental model of what they want to accomplish and the computer's support of the user's task. Professional practitioners in HCI are usually designers concerned with the practical application of design methodologies to problems in the world. Their work often revolves around designing graphical user interfaces and web interfaces. Researchers in HCI are interested in developing new design methodologies, experimenting with new devices, prototyping new software systems, exploring new interaction paradigms, and developing models and theories of interaction.

**1.2 Sign Language Recognition:**

Sign Language recognition is a topic in computer science and language technology with the goal of interpreting human Sign Languages via mathematical algorithms. Sign Languages can originate from any bodily motion or state but commonly originate from the face or hand. Current focuses in the field include emotion recognition from the face and Sign Language recognition. Many approaches have been made using cameras and computer vision algorithms to interpret sign language. However, the identification and recognition of posture, gait, proxemics, and human behaviors is also the subject of Sign Language recognition techniques. Sign Language recognition can be seen as a way for computers to begin to understand human body language, thus building a richer bridge between machines and humans than primitive text user interfaces or even GUIs (graphical user interfaces), which still limit the majority of input to keyboard and mouse.

Sign Language recognition enables humans to communicate with the machine (HMI) and interact naturally without any mechanical devices. Using the concept of Sign Language recognition, it is possible to point a finger at the computer screen so that the cursor will move accordingly. This could potentially make conventional input devices such as mouse, keyboards and even touch-screens redundant. Sign Language recognition can be conducted with techniques from computer vision and image processing. The literature includes ongoing work in the

computer vision field on capturing Sign Languages or more general human pose and movements by cameras connected to a computer.

Sign Language recognition and pen computing: This computing not only going to reduce the hardware impact of the system but also it increases the range of usage of physical world object instead of digital object like keyboards, mouses. Using this we can implement and can create a new thesis of creating of new hardware no requirement of monitors too. This idea may lead us to the creation of holographic display. The term Sign Language recognition has been used to refer more narrowly to non-text-input handwriting symbols, such as inking on a graphics tablet, multi-touch Sign Languages, and mouse Sign Language recognition. This is computer interaction through the drawing of symbols with a pointing device cursor.

In computer interfaces, two types of Sign Languages are distinguished:[9] We consider online Sign Languages, which can also be regarded as direct manipulations like scaling and rotating. In contrast, offline Sign Languages are usually processed after the interaction is finished; e. g. a circle is drawn to activate a context menu.

Offline Sign Languages: Those Sign Languages that are processed after the user interaction with the object. An example is the Sign Language to activate a menu.

Online Sign Languages: Direct manipulation Sign Languages. They are used to scale or rotate a tangible object.

The ability to track a person's movements and determine what Sign Languages they may be performing can be achieved through various tools. Although there is a large amount of research done in image/video based Sign Language recognition, there is some variation within the tools and environments used between implementations.

Wired gloves: These can provide input to the computer about the position and rotation of the hands using magnetic or inertial tracking devices. Furthermore, some gloves can detect finger bending with a high degree of accuracy (5-10 degrees), or even provide haptic feedback to the user, which is a simulation of the sense of touch. The first commercially available hand-tracking glove-type device was the Data Glove, a glove-type device which could detect hand position, movement and finger bending. This uses fiber optic cables running down the back of the hand.

Light pulses are created and when the fingers are bent, light leaks through small cracks and the loss is registered, giving an approximation of the hand pose.

Depth-aware cameras: Using specialized cameras such as structured light or time-of-flight cameras, one can generate a depth map of what is being seen through the camera at a short range, and use this data to approximate a 3d representation of what is being seen. These can be effective for detection of Sign Languages due to their short range capabilities.

Stereo cameras: Using two cameras whose relations to one another are known, a 3d representation can be approximated by the output of the cameras. To get the cameras' relations, one can use a positioning reference such as a lexian-stripe or infrared emitters. In combination with direct motion measurement (6D-Vision) Sign Languages can directly be detected.

Controller-based Sign Languages: These controllers act as an extension of the body so that when Sign Languages are performed, some of their motion can be conveniently captured by software. Mouse Sign Languages are one such example, where the motion of the mouse is correlated to a symbol being drawn by a person's hand, as is the Wii Remote or the Myo, which can study changes in acceleration over time to represent Sign Languages. Devices such as the LG Electronics Magic Wand, the Loop and the Scoop use Hillcrest Labs' Free space technology, which uses MEMS accelerometers, gyroscopes and other sensors to translate Sign Languages into cursor movement. The software also compensates for human tremor and inadvertent movement. Audio Cubes are another example. The sensors of these smart light emitting cubes can be used to sense hands and fingers as well as other objects nearby, and can be used to process data. Most applications are in music and sound synthesis, but can be applied to other fields.

Single camera: A standard 2D camera can be used for Sign Language recognition where the resources/environment would not be convenient for other forms of image-based recognition. Earlier it was thought that single camera may not be as effective as stereo or depth aware cameras, but some companies are challenging this theory. Software-based Sign Language recognition technology using a standard 2D camera that can detect robust Sign Languages, hand signs, as well as track hands or fingertip at high accuracy has already been embedded in

Lenovo’s Yoga ultrabooks, Pantech’s Vega LTE smartphones, Hisense’s Smart TV models, among other devices.

Depending on the type of the input data, the approach for interpreting a Sign Language could be done in different ways. However, most of the techniques rely on key pointers represented in a 3D coordinate system. Based on the relative motion of these, the Sign Language can be detected with a high accuracy, depending of the quality of the input and the algorithm’s approach.

In order to interpret movements of the body, one has to classify them according to common properties and the message the movements may express. For example, in sign language each Sign Language represents a word or phrase. The taxonomy that seems very appropriate for Human-Computer Interaction has been proposed by Quek in "Toward a Vision-Based Sign Language Interface" He presents several interactive Sign Language systems in order to capture the whole space of the Sign Languages: 1. Manipulative; 2. Semaphoric; 3. Conversational.

Some literature differentiates 2 different approaches in Sign Language recognition: a 3D model based and an appearance-based. The foremost method makes use of 3D information of key elements of the body parts in order to obtain several important parameters, like palm position or joint angles. On the other hand, Appearance-based systems use images or videos for direct interpretation.

A real hand (left) is interpreted as a collection of vertices and lines in the 3D mesh version (right), and the software uses their relative position and interaction in order to infer the Sign Language.

**1.3Sign Language Recognition algorithms:**

**1.3.1 3D model-based algorithms**

The 3D model approach can use volumetric or skeletal models, or even a combination of the two. Volumetric approaches have been heavily used in computer animation industry and for computer vision purposes. The models are generally created of complicated 3D surfaces, like NURBS or polygon meshes.

The drawback of this method is that is very computational intensive and systems for live analysis is still to be developed. For the moment, a more interesting approach would be to map simple primitive objects to the person's most important body parts (for example cylinders for the arms and neck, sphere for the head) and analyze the way these interact with each other. Furthermore, some abstract structures like super-quadrics and generalized cylinders may be even more suitable for approximating the body parts. The exciting thing about this approach is that the parameters for these objects are quite simple. In order to better model the relation between these, we make use of constraints and hierarchies between our objects.

The skeletal version (right) is effectively modelling the hand (left). This has fewer parameters than the volumetric version and it's easier to compute, making it suitable for real-time Sign Language analysis systems.

### 1.3.2 Skeletal-based algorithms

Instead of using intensive processing of the 3D models and dealing with a lot of parameters, one can just use a simplified version of joint angle parameters along with segment lengths. This is known as a skeletal representation of the body, where a virtual skeleton of the person is computed and parts of the body are mapped to certain segments. The analysis here is done using the position and orientation of these segments and the relation between each one of them (for example the angle between the joints and the relative position or orientation)

Advantages of using skeletal models:

- Algorithms are faster because only key parameters are analyzed.
- Pattern matching against a template database is possible
- Using key points allows the detection program to focus on the significant parts of the body

These binary silhouette (left) or contour (right) images represent typical input for appearance-based algorithms. They are compared with different hand templates and if they match, the correspondent Sign Language is inferred.

### 1.3.3 Appearance-based models

These models don’t use a spatial representation of the body anymore, because they derive the parameters directly from the images or videos using a template database. Some are based on the deformable 2D templates of the human parts of the body, particularly hands. Deformable templates are sets of points on the outline of an object, used as interpolation nodes for the object’s outline approximation. One of the simplest interpolation functions is linear, which performs an average shape from point sets, point variability parameters and external deformators. These template-based models are mostly used for hand-tracking, but could also be of use for simple Sign Language classification.

A second approach in Sign Language detecting using appearance-based models uses image sequences as Sign Language templates. Parameters for this method are either the images themselves, or certain features derived from these. Most of the time, only one (monoscopic) or two (stereoscopic) views are used.

### 1.4 Applications based on the hands-free interface:

#### 1.4.1 Interactive expositions

Nowadays, expositions based on new ways of interaction need contact with the visitors that play an important role in the exhibition contents. Museums and expositions are open to all kind of visitors, therefore, these ‘‘sensing expositions’’ look forward to reaching the maximum number of people. This is the case of ‘‘Galicia dixital’’, an exposition. Visitors go through all the phases of the exposition sensing, touching and receiving multimodal feedback such as audio, video, haptics, interactive images or virtual reality. In one phase there is a slider-puzzle with images of Galicia to be solved. There are four computers connected enabling four users to compete to complete the six puzzles included in the application.

Visitors use a touch-screen to interact with the slider-puzzle, but the characteristics of this application make it possible to interact by means of the hands-free interface in a very easy manner. Consequently, the application has been adapted to it and therefore, disabled people can also play this game and participate in a more active way in the exposition.

#### 1.4.2 Non-verbal communication

By means of human–computer interaction, one ambitious objective is to achieve communication for people with speech disorders using new technologies. Nowadays, there are different augmentative communication systems for people with speech limitations, ranging from unaided communication such as sign languages, to computerized iconic languages with voice output systems such as MinspeakTM. We present BlissSpeaker, an application based on a symbolic graphical-visual system for nonverbal communication named Bliss. The Bliss system can be used as an augmentative system or for completely replacing verbal communication. It is commonly used by people with cerebral palsy, but with the following learning aptitude requirements:

1. Cognitive abilities;

2. Good visual discrimination;

3. Possibility of indicating the desired symbol;

4. Good visual and auditory comprehension.

Some speech therapists use it in their sessions to help themselves with children with speech disorders and to help in the prevention of linguistic and cognitive delays in crucial stages of a child's life. The Blissymbolics language is currently composed of over 2000 graphic symbols that can be combined and re-combined to create new symbols. The number of symbols is adaptable to the capabilities and necessities of the user, for example, BlissSpeaker has 92 symbols that correspond to the first set of Bliss symbols for preschool children. BlissSpeaker is an application that verbally reproduces statements built using Bliss symbols, which allows a more ''natural'' communication between a child using Bliss and a person that does not understand or use these symbols, for example, the children's relatives. The application can work with any language, as long as there is an available compatible SAPI (Speech Application Programming Interface). The system's process is shown. The potential users of BlissSpeaker are children with speech disorders; therefore, its operation is to be very simple and intuitive. Moreover, audio, vision and traditional graphical user interfaces combined together configure a very appealing multimodal interface that can help attract and involve the user in its use. Furthermore, the use of the hands-free interface with BlissSpeaker will help to fulfil the third requirement of Bliss user, which is the possibility of indicating the desired symbol. It will offer

children with upper-body physical disabilities and speech difficulties a way to communicate themselves through an easy interface and their teachers or relatives will understand them better due to the symbols' vocal reproduction. Furthermore, the use of the new interface can make learning of Bliss language more enjoyable and entertaining, and it also promotes the children's coordination, because the interface works with head motion. This system was evaluated in a children's scientific fair. The system was tested by more than 60 disabled and non-disabled children from 6 to 14 years of age. A short explanation on how it works was given. They operated the application with surprising ease and even if they had never seen Bliss symbols before, they created statements that made sense and reproduced them for their class mates. Children enjoyed interacting with the computer through the functionalities that the face-based interface offered. Moreover, upper-body physical disabled children are grateful for the opportunity of accessing a computer.

**1.5 Color Models:**

The aim of the proposed project is to overcome the challenge of skin color detection for natural interface between user and machine. So to detect the skin color under dynamic background the study of various color models was done for pixel based skin detection. Three color spaces has been chosen which are commonly used in computer vision applications.

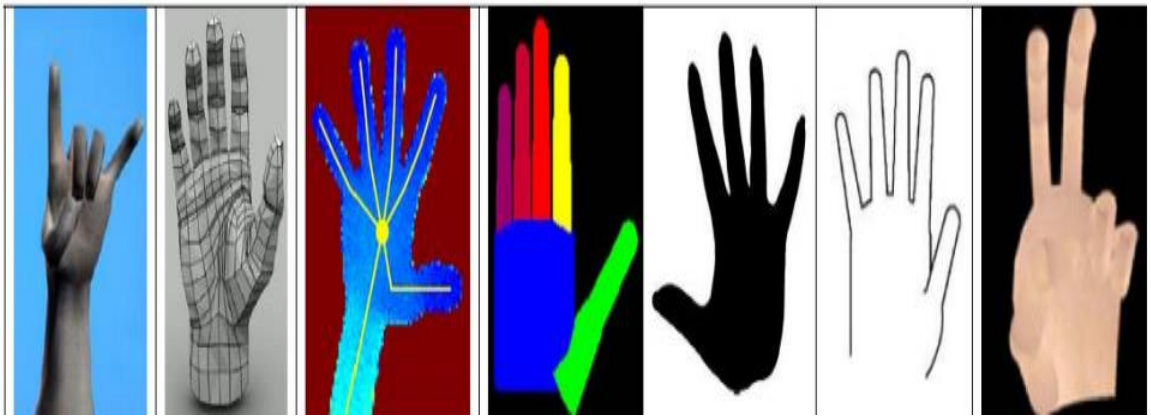
**RGB:** Three primary colors red(R), green(G), and blue(B) are used. The main advantage of this color space is simplicity. However, it is not perceptually uniform. It does not separate luminance and chrominance, and the R, G, and B components are highly correlated.

**HSV (Hue, Saturation, Value):** It express Hue with dominant color (such as red, green, purple and yellow)of an area. Saturation measures the colorfulness of an area in proportion to its brightness. The "intensity", "lightness", or "Values" is related to the color luminance. This model discriminates luminance from chrominance. This is a more intuitive method for describing colors, and because the intensity is independent of the color information this is very useful model for computer vision. This model gives poor result where the brightness is very low. Other similar color spaces are HSI and HSL (HLS).

**CIE –Lab:** It defined by the International Commission on Illumination. It separates a luminance variable L from two perceptually uniform chromaticity variable (a, b).

### 1.6 Hand modeling for Sign Language recognition

Human hand is an articulated object with 27 bones and 5 fingers. Each of these fingers consists of three joints. The four fingers (little, ring, middle and index) are aligned together and connected to the wrist bones in one tie and at a distance there is the thumb. Thumb always stands on the other side of the four fingers for any operation, like capturing, grasping, holding etc. Human hand joints can be classified as flexion, twist, directive or spherical depending up on the type of movement or possible rotation axes. In total human hand has approximately 27 degrees of freedom. As a result, a large number of Sign Languages can be generated. Therefore, for proper recognition of the hand, it should be modeled in a manner understandable as an interface in Human Computer Interaction (HCI). There are two types of Sign Languages, Temporal (dynamic) and Spatial (shape). Temporal models use Hidden Markov Model (HMM), KalmanFilter , Finite State Machines, Neural Network (NN). Hand modeling in spatial domain can be further divided into two categories, 2D (appearance based or view based) model and 3D based model. 2D hand modeling can be represented by deformable templates, shape representation features, motion and coloured markers. Shape representation feature is classified as geometric features (i.e. live feature) and non-geometric feature. Geometric feature deals with location and position of fingertips, location of palm and it can be processed separately. The non – geometric feature includes colour, silhouette and textures, contour, edges, image moments and Eigen vectors. Non-geometric features cannot be seen (blind features) individually and collective processing is required. The deformable templates are flexible in nature and allow changes in shape of the object up to certain limit for little variation in the hand shape. Image motion based model can be obtained with respect to colour cues to track the hand. Coloured markers are also used for tracking the hand and detecting the fingers/ fingertips to model the hand shape. Hand shape can also be represented using 3D modeling. The hand shape in 3D can be volumetric, skeletal and geometric models. Volumetric models are complex in nature and difficult for computation in real-time applications. It uses a lot of parameters to represent the hand shape. Instead other geometric models, such as cylinders, ellipsoids and spheres are considered as alternative for such model for hand shape approximation. Skeletal model represents the hand structure with 3D structure with reduced set of parameters. Geometric models are used for hand animation and real-time applications. Polygon meshes and cardboard models are examples of geometric models.



## CHAPTER 2

## 2. Literature survey:

### 2.1Systematic review of Kinect applications in elderly care and stroke rehabilitation

**Author: David Webster**

As the Kinect is a relatively new piece of hardware, establishing the limitations of the sensor within specific application scenarios is an ongoing process. Nevertheless, provide a list of current limitations of Kinect that noted based on our review of applications in elderly care systems. Current Kinect-based fall risk reduction strategies are derived from gait-based, early intervention methodologies and thus are only indirectly related to true fall prevention which would require some form of feedback prior to a detected potential fall event. Occlusion in fall detection algorithms, while partially accounted for through the methodologies of the various systems discussed, is still a major challenge inherent in Kinect-based fall detection systems. Current strategies focus on a subject who stands, sits, and falls in an ideal location of the Kinect's field of vision, while authentic falls in realistic home environment conditions are more varied, therefore the current results should not be taken as normative. The Kinect sensor must be fixed to a specific location and has a range of capture of roughly ten meters. This limitation dictates that fall events must occur directly in front of the sensor's physical location. While it has been noted that a strategically placed array of Kinect sensors could mitigate this limitation, a system utilizing this methodology has not yet been implemented and evaluated. Without careful consideration of the opinions of a system's proposed user base, concerns regarding ubiquitous always-on video capture systems, such as the Kinect, may inhibit wide-scale system adoption. During the review, it was noted that research related to the reception of alert support systems is at an early phase, likely due to in-home hardware previously being cumbersome and expensive. With the Kinect having the potential to be widely disbursed in in-home setting monitoring systems, this avenue of research has become more viable and relevant. In this section we provide a review of applications of Kinect in stroke rehabilitation grouped under 2 categories: 1) Evaluation of Kinect's Spatial Accuracy, and 2) Kinect based Rehabilitation Methods. These categories follow the trend of the literature to first evaluate the Kinect sensor as a clinically viable tool for rehabilitation. Motor function rehabilitation for stroke patients typically aims to strengthen and retrain muscles to rejuvenate debilitated functions, but inadequate completion of

rehabilitation exercises drastically reduces the potential outcome of overall motor recovery. These exercises are often unpleasant and/or painful leading to patients' tolerance for exercise to decrease as indicated. Then noted that decreased tolerance or motivation often lead to intentional and unintentional 'cheating' or, in the worst case scenario, avoidance of rehabilitation exercises altogether. The Kinect may contain the potential to overcome these barriers to in-home stroke rehabilitation as an engaging and accurate markerless motion capture tool and controller interface; however, a functional foundation of Kinect-based rehabilitation potential needs to be established focusing on the underlying strategies of rehabilitation schemas rather than the placating effects offered by serious games**.** However, some significant technological limitations still present are: a fixed location sensor with a range of capture of only roughly ten meters; a difficulty in fine movement capture; shoulder joint biomechanical accuracy, and fall risk reduction methodologies that only utilize indirect, gait-based preemptive training. The directions for future work are vast and have promise to enhance elderly care; stroke patient motivation to accurately complete rehabilitation exercises; rehabilitation record keeping, and future medical diagnostic and rehabilitation methods. Based on review of the literature, have reported a summary of critical issues and suggestions for future work in this domain.

**2.2 Title: ChaLearn Sign Language Challenge: Design and First Results**

**Author: Isabelle Guyon**

For the unpublished methods, summarize the descriptions provided in the fact sheets. Interestingly, all top ranking methods are based on techniques making no explicit detection and tracking of humans or individual body parts. The winning team (alfnie) used a novel technique called "Motion Signature analyses", inspired by the neural mechanisms underlying information processing in the visual system. This is an unpublished method using a sliding window to perform simultaneously recognition and temporal segmentation, based solely on depth images. The second best ranked participants (team Pennect) did not publish their method yet. From the fact sheets we only know that it is an HMM-style method using HOG/HOF features with a temporal segmentation based on candidate cuts. Only RGB images were used. The methods of the two best ranking participants are quite fast. They claim a linear complexity in image size, number of frames, and number of training examples. The third best ranked team (One Million Monkeys) did not publish either, but they provided a high level description indicating that the system uses a HMM in which a state is created for each frame of the Sign Language exemplars.

The state machine includes skips and self-loops to allow for variation in the speed of the Sign Language execution. The most likely sequence of Sign Languages is determined by a Viterbi search. Comparisons between frames are based on the edges detected in each frame. Edges are associated with several attributes including the X/Y coordinates, their orientation, their sharpness, their depth and location in an area of change. In matching one frame against another, they find the nearest neighbor in the second frame for every edge point in the first frame, and calculate the joint probability of all the nearest neighbors using a simple Gaussian model. The system works exclusively from the depth images. The system is one of the slowest proposed. Its processing speed is linear in number of training examples but quadratic in image size and number of frames per video. The fourth best ranked entrant ManavenderMalgireddy (immortals) published in these proceedings a paper called “Detecting and Localizing Activities/Sign Languages in Video Sequences”. They detect and localize activities from HOG/HOF features in unconstrained real-life video sequences, a more complex problem than that of the challenge. To obtain real-life data, they used video clips from the Human Motion Database (HMDB). The detection and localization paradigm was adapted from the speech recognition community, where a keyword model is used for detecting key phrases in speech. The method learns models for activities-of-interest and creates a network of these models to detect keywords. According to the paper, the approach out-performed all the current state-of-the-art classifiers when tested on publicly available datasets such as KTH and HMDB.The final evaluation performance for the first round of the Sign Language challenge is around 10% error, still far from human performance, which is below 2% error. However, the progress made during round 1, starting at a baseline performance of 60% error indicates that the objective of attaining or surpassing human performance could possibly be reached in the second round. There is room for improvement particularly because the top two ranking participants used only one modality (the first one depth only and the second one RGB only) and because many Sign Languages are recognizable only when details of hand posture are used, yet none of the methods disclosed made use of such information. The most efficient techniques so far have used sequences of features processed by graphical models of the HMM/CRF family, similar to techniques used in speech recognition. No use of skeleton extraction of body part detection was made. Rather, orientation and ad hoc features were extracted. It is possible that progress will also be made in feature extraction by making better use of the development data for transfer learning.

**2.3 Title:Sign Language Recognition: A Survey**

**Author:SushmitaMitra**

Generally, there exist many-to-one mappings from conceptsto Sign Languages and vice versa. Hence, Sign Languages are ambiguous and incompletely specified. For example, to indicate the concept “stop,” one can use Sign Languages such as a raised hand with palm facing forward, or, an exaggerated waving of both hands over the head. Similar to speech and handwriting, Sign Languages vary between individuals, and even for the same individual between different instances. There have been varied approaches to handle Sign Language recognition, ranging from mathematical models based on hidden Markov chains to tools or approaches based on soft computing. In addition to the theoretical aspects, any practical implementation of Sign Language recognition typically requires the use of different imaging and tracking devices or gadgets. These include instrumented gloves, body suits, and marker based optical tracking. Traditional 2-D keyboard-, pen-, and mouse-oriented graphical user interfaces are often not suitable for working in virtual environments. Rather, devices that sense body (e.g., hand, head) position and orientation, direction of gaze, speech and sound, facial expression, galvanic skin response, and other aspects of human behavior or state can be used to model communication between a human and the environment. Sign Languages can be static (the user assumes a certain pose or configuration)or dynamic (with prestroke, stroke, and poststrokephases). Some Sign Languages also have both static and dynamic elements, as in sign languages. Again, the automatic recognition of natural continuous Sign Languages requires their temporal segmentation. Often one needs to specify the start and end points of a Sign Language in terms of the frames of movement, both in time and in space. Sometimes a Sign Language is also affected by the context of preceding as well as following Sign Languages. Moreover, Sign Languages are often language and culture-specific. They can broadly be of the following types: 1) hand and arm Sign Languages: recognition of hand poses, sign languages, and entertainment applications (allowing children to play and interact in virtual environments); 2) head and face Sign Languages: some examples are: a) nodding or shaking of head; b) direction of eye gaze; c) raising the eyebrows; d) opening the mouth to speak; e) winking, f) flaring the nostrils; and g) looks of surprise, happiness, disgust, fear, anger, sadness, contempt, etc.; 3) body Sign Languages: involvement of full body motion, as in: a) tracking movements of two people interacting outdoors; b) analyzing movements of a dancer for generating matching music and

graphics; and c) recognizing human gaits for medical rehabilitation and athletic training. Typically, the meaning of a Sign Language can be dependent on the following:spatial information: where it occurs;pathic information: the path it takes; symbolic information: the sign it makes; affective information: its emotional quality. Facial expressions involve extracting sensitive features (related to emotional state) from facial landmarks such as regions surrounding the mouth, nose, and eyes of a normalized image. Often dynamic image frames of these regions are tracked to generate suitable features. The location, intensity, and dynamics of the facial actions are important for recognizing an expression. Moreover, the intensity measurement of spontaneous facial expressions is often more difficult than that of posed facial expressions. More subtle cues such as hand tension, overall muscle tension, locations of self-contact, and pupil dilation are sometimes used. In order to determine all these aspects, the human body position, configuration (angles and rotations), and movement (velocities) need to be sensed. This can be done either by using sensing devices attached to the user. Those may be magnetic field trackers, instrumented (data) gloves, and body suits, or by using cameras and computer vision techniques. Each sensing technology varies along several dimensions, including accuracy, resolution,latency, range of motion, user comfort, and cost. Glove-based gestural interfaces typically require the user to wear a cumbersome device and carry a load of cables connecting the device to a computer. This hinders the ease and naturalness of the user's interaction with the computer. Vision-based techniques, while overcoming this, need to contend with other problems related to occlusion of parts of the user's body. While tracking devices can detect fast and subtle movements of the fingers when the user's hand is moving, a vision-based system will at best get a general sense of the type of finger motion. Again, vision-based devices can handle properties such as texture and color for analyzing a Sign Language, while tracking devices cannot. Vision-based techniques can also vary among themselves in: 1) the number of cameras used; 2) their speed and latency; 3) the structure of environment (restrictions such as lighting or speed of movement); 4) any user requirements (whether user must wear anything special); 5) the low-level features used (edges, regions, silhouettes, moments, histograms); 6) whether 2-D or 3-D representation is used; and 7) whether time is represented. There is, however, an inherent loss in information whenever a 3-D image is projected to a 2-D plane. Again, elaborate 3-D models involve prohibitive high dimensional parameter spaces. A tracker also needs to handle changing shapes and sizes of the Sign Language-generating object (that varies between individuals), other

moving objects in the background, and noise.Sign Language recognition is an ideal example of multidisciplinary research. There are different tools for Sign Language recognition, basedon the approaches ranging from statistical modeling, computervision and pattern recognition, image processing, connectionist systems, etc. Most of the problems have been addressed based on statistical modeling, such as PCA, HMMs, Kalman filtering, more advanced particle filtering and condensation algorithms. FSM has been effectively employed in modeling human Sign Languages. Computer vision and pattern recognition techniques, involving feature extraction, object detection, clustering, and classification, have been successfully used for many Sign Languagerecognition systems. Image-processing techniques such as analysis and detection of shape, texture, color, motion, optical flow, image enhancement, segmentation, and contour modeling, have also been found to be effective. Connectionist approaches, involving multilayer perceptron (MLP), time delay neural network (TDNN), and radial basis function network (RBFN), have been utilized in Sign Language recognition as well. While static Sign Language (pose) recognition can typically be accomplished by template matching, standard pattern recognition, and neural networks, the dynamic Sign Language recognition problem involves the use of techniques such as time-compressing templates, dynamic time warping, HMMs, and TDNN. In the rest of this section, we discuss the principles and background of some of these popular tools used in Sign Language recognition.

**2.4 Title: Results and Analysis of the ChaLearn Sign Language Challenge 2012**

**Author: I.Guyon**

Sign Language recognition is an important sub-problem in many computer vision applications,including image/video indexing, robot navigation, video surveillance, computer interfaces, and gaming.With simple Sign Languages such as hand waving, Sign Language recognition could enable controlling the lights or thermostat in your home or changing TV channels. The same technology may even make it possible to automatically detect more complex human behaviors, to allow surveillance systems to sound an alarm when someone is acting suspiciously, for example, or to send help whenever a bedridden patient shows signs of distress. Sign Language recognition also provides excellent benchmarks for Adaptive and Intelligent Systems (AIS) and computer vision algorithms. The recognition of continuous, natural Sign Languages is very challenging due to the multi-modal nature of the visual cues (e.g., movements of fingers and lips, facial expressions, body pose), as well as technical limitations such as spatial

and temporal resolution and unreliable depth cues. Technical difficulties include tracking reliably hand, head and body parts, and achieving 3D invariance. The competition we organized helped improve the accuracy of Sign Language recognition using Microsoft Kinect motion sensor technology, a low cost 3D depth-sensing camera. Most of the participants employed image enhancement and filtering techniques, in majority denoising or outlier removal and background removal. Some reduced the image resolution for faster processing. Notably, some of the top ranking participants did not do any such low level preprocessing. The majority of the top ranking participants used HOG/HOF features and/or ad-hoc hand crafted features, edge/corner detectors or SIFT/STIP features. The latter use a bag-of-feature strategy, which ignores exact location of features and therefore provides some robustness against translations. The winner of both rounds of the challenge claims that his features are inspired by the human visual system. Very few participants resorted to using body parts or trained features. Most participants used the depth image only, but about one third used both RGB and depth images. Interestingly, the second place winner in round 1 used the RGB image only. About one third of the participants did no dimensionality reduction at all and one third resorted to feature selection. Other popular techniques included linear transforms (such as PCA) and clustering.For temporal segmentation, most participants used candidate cuts based on similarities with the resting position or based on amount of motion. All the top ranking participants used recognition-based segmentation techniques (in which recognition and segmentation are integrated). As Sign Language representation, all highest ranking participants used a variable length sequence of feature vectors (sometimes in combination with other representations). To handle such variable length representations, the highest ranking participants used Hidden Markov Models (HMM), Conditional Random Fields (CRF) or other similar graphical models. This is a state machine including skips and self-loops to allow for variation in the speed of the Sign Language execution. The most likely sequence of Sign Languages is determined by a Viterbi search. Some highly ranked, but not top ranking, participants used a bag-of-word representation or image templates, including motion energy or motion history representations. The corresponding classifiers were usually nearest neighbors (using as metric the Euclidean distance or correlation). One participant used a linear SVM. Many participants made use of the development data to either learn features or Sign Language representations in the spirit of “transfer learning”. Most participants claimed that the algorithmic complexity of their methods was linear in image size, number of frames per

video, and number of training examples. The median execution time on the 20 batches of the final evaluation set was 2.5 hours, which is very reasonable and close to real time performance. However, there were a few outliers and it took up to 50 hours for the slowest code.

**2.5 Title:A Robust Background Subtraction and Shadow Detection**

**Author: ThanaratHorprasert**

The capability of extracting moving objects from a videosequence is a fundamental and crucial problem of many vision systems that include video surveillance, traffic monitoring, human detection and tracking for video teleconferencing or human-machine interface, video editing, among other applications. Typically, the common approach for discriminating moving object from the background scene is background subtraction. The idea is to subtract the current image from a reference image, which is acquired from a static background during a period of time. The subtraction leaves only non-stationary or new objects, which include the objects' entire silhouette region. The technique has been used for years in many vision systems as a preprocessing step for object detection and tracking. The results of the existing algorithms are fairly good; in addition, many of them run in real-time. However, many of these algorithms are susceptible to both global and local illumination changes such as shadows and highlights. These cause the consequent processes, e.g. tracking, recognition, etc., to fail. The accuracy and efficiency of the detection are very crucial to those tasks. This problem is the underlying motivation of this work and wants to develop a robust and efficiently computed background subtraction algorithm that is able to cope with the local illumination change problems, such as shadows and highlights, as well as the global illumination changes. Being able to detect shadows is also very useful to many applications especially in"Shape from Shadow" problems. The method must also address requirements of sensitivity, reliability, robustness, and speed of detection. In this paper, present a novel algorithm for detecting moving objects from a static background scene that contains shading and shadows using color images. In next section, propose a new computational color model (brightness distortion and chromaticity distortion) that helps us to distinguish shading background from the ordinary background or moving foreground objects. Next, propose an algorithm for pixel classification and threshold selection. Experimental results and sample applications are respectively.One of the fundamental abilities of human vision is color constancy. Humans tend to be able to assign a constantcolor to an object even under changing of illumination overtime or space. The perceived color of a point in a scene dependson

many factors including physical properties of thepoint on the surface of the object. Important physical propertiesof the surface in color vision are surface spectral reflectanceproperties, which are invariant to changes of illumination,scene composition or geometry. On Lambertain,or perfect matte surfaces, the perceived color is the productof illumination and surface spectral reflectance.This led to our idea of designing a color model thatseparates these two terms; in other words that separatesthe brightness from the chromaticity component.As the person moves, he both obscuresthe background and casts shadows on the floor and wall.Red pixels depict the shadow, and we can easily see how theshape of the shadow changes as the person moves. Althoughit is difficult to see, there are green pixels, which depict thehighlighted background pixels, appearing along the edge ofthe person's sweater. Figure 5 shows a frame of an outdoorscene containing a person walking across a street. Althoughthere are small motions of background objects, such as thesmall motions of leaves and water surface, the result showsthe robustness and reliability of the algorithm. It shows another indoor sequenceof a person moving in a room; at the middle of thesequence, the global illumination is changed by turning halfof the fluorescence lamps off. The system is still able to detectthe target successfully.

**2.6 Title: Naked image detection based on adaptive and extensible skin color model**

**Author: Jiann-Shu Lee**

In a relatively short period of time, the Internet has becomereadily accessible in most organizations, schools and homes.Meanwhile, however, the problem of pornography through the Internet access in the workplace, at home and in education hasconsiderably escalated. In the workplace, the pornography relatedaccess not only costs companies millions in non-businessInternet activities, but it also has led to shattering business reputationsand harassment cases. Being anonymous and often anarchic,images that would be illegal to sell even in adult bookstorescan be easily transferred to home through the Internet,causing juveniles to see those obscene images intentionally orunintentionally. Therefore, how to effectively block or filter outpornography has been arousing a serious concern in related researchareas.The mostly used approach to blocking smut from the Internetis based on contextual keyword pattern matching technologythat categorizes URLs by means of checking contexts ofweb pages and then traps the websites assorted as the obscene.Although this method can successfully filter out a mass of obscene websites, it is unable to deal with images, leading to its failure to detect those obscene web sites containingnaked images instead of smut texts. Besides the threat comingfrom the web

sites, a lot of the e-mail image attachmentsare naked. Hence, the development of naked image detectiontechnology is urgently desired to prevent juveniles from gettingaccess to pornographic contents from the Internet morethoroughly. As can be seen in these methods, none of them considerthe inference coming from special lighting and color altering.There exist a large number of naked pictures taken under speciallighting. Usually, warm lighting is applied to make skintone look more attractive, while human skin color deviatesfrom the normal case at the same time. If the skin color modelcannot tolerate the deviation, it will tend to miss a lot of nakedpictures. On the contrary, if the skin color model accommodatesthe deviation, an abundance of non-skin objects likewood, desert sand, rock, foods, and the skin or fur of animalswould be detected in the skin detection phase and deterioratesthe system performance. Accordingly, the above mentionedapproaches suffer from the skin color deviation resulting fromspecial lighting, which is often seen in the naked images.Dealing with the special lighting effect in the naked images isa difficult task. If the skin color model tolerates the deviation,lots of non-skin objects would be detected simultaneously.A feasible solution for the problem is to adapt the adoptedskin chroma distribution to the lighting of the input image.Based on this concept, a new naked image detection system isproposed. We develop a learning-based chromatic distributionmatchingscheme that consists of the online samplingmechanism and the one-class-one-net neural network. Basedon this approach, the object's chroma distribution can be onlinedetermined so that the skin color deviation coming from lightingcan be accommodated without sacrificing the accuracy.The roughness feature is further applied to reject confusioncoming from non-skin objects, so the skin area can be moreeffectively detected. Several representative features inducedfrom the naked images are used to verify these skin areas.Subsequently, the face detection process is employed to filterout those false candidates coming from mug shots. The skin tone is formed by the interaction between skin andlight. Therefore, the captured skin color in an image dependson the surrounding light in addition to the intrinsic skin tone.To make naked images look more attractive, photographersusually apply special lighting, thereby altering the chromadistribution of the skin tone. Hence, if the referenced skinchroma distribution is gathered from the normal conditionsbeforehand, the corresponding skin detection performancewill be dramatically degenerated.

**2.7 Title: Principal motion: PCA-basedreconstruction of motion histograms**

**Author: Hugo Jair Escalante**

The principal motion is the implementation of a reconstruction approachto Sign Language recognition based on principal components analysis (PCA). Theunderlying idea is to perform PCA on the frames in each video from thevocabulary, storing the PCA models. Frames in test-videos are projectedinto the PCA space and reconstructed back using each of the PCA models,one for each Sign Language in the vocabulary. Next we measure the reconstructionerror for each of the models and assign a test video the Sign Language that obtainsthe lowest reconstruction error. The rest of this document provides moredetails about the principal motion object. The PCA reconstruction approach to Sign Language recognition is inspired from the one-class classification task, where the reconstruction error via PCA hasbeen used to identify outlier. The method is also inspired in a recent method for spam classification. The underlying hypothesis of the methodis that a test video will be better reconstructed with a PCA model that wasobtained with another video that contains the same Sign Language.

**2.8 Title: A Framework for Sign Language Recognition Basedon Accelerometer and EMG Sensors**

**Author: Xu Zhang, Xiang Chen**

Sign Language recognition provides an intelligent, natural,and convenient way of human computer interaction(HCI). Sign language recognition (SLR) and Sign Language-basedcontrol are two major applications for Sign Language recognition technologies. SLR aims to interpret sign languages automaticallyby a computer in order to help the deaf communicatewith hearing society conveniently. Since sign language is akind of highly structured and largely symbolic human Sign Languageset, SLR also serves as a good basic for the development ofgeneral Sign Language-based HCI. In particular, most efforts on SLR are based on hidden Markov models (HMMs) whichare employed as effective tools for the recognition of signalschanging over time. On the other hand, Sign Language-basedcontrol translates Sign Languages performed by human subjects intocontrolling commands as the input of terminal devices, whichcomplete the interaction approaches by providing acoustic,visual, or other feedback to human subjects. Many previous researchers investigated various systems whichcould be controlled by Sign Languages, such as media players,remote controllers, robots, and virtual objects or environments.According to the sensing technologies used to capture Sign Languages,conventional researches on Sign

Languagerecognition canbe categorized into two main groups: data glove-based andcomputer vision-based techniques The multichannel signals recorded in the process of the handSign Language actions which represent meaningful Sign Languages arecalled active segments. The intelligent processing of Sign Languagerecognition needs to automatically determine the start andend points of active segments from continuous streams of inputsignals. The Sign Language data segmentation procedure is difficult due to movement epenthesis. The EMG signal level representsdirectly the level of muscle activity. As the hand movementswitches from one Sign Language to another, the correspondingmuscles relax for a while, and the amplitude of the EMGsignal is momentarily very low during movement epenthesis.Thus, the use of EMG signal intensity helps to implement datasegmentation in a multi sensor system. In method, only themultichannel EMG signals are used for determining the startand end points of active segments. The segmentation is basedon a moving average algorithm and thresholding. The ACCsignal stream is segmented synchronously with the EMG signalstream. Thus, the use of EMG would help the SLR system to automatically distinguish between valid Sign Language segments and movement epenthesis from continuous streams of input signals. The detection of active segments consists of four steps based on the instantaneous energy of the average signal of the multiple EMG channels.

**2.9 Title: One-shot Learning Sign Language Recognition from RGB-D Data Using Bag of Features**

**Author: Jun Wan**

DBN includes HMMs and Kalman filters as special cases and defined five classes ofSign Languages for HCI and developed a DBN-based model which used local features (contour, moment, height) and global features (velocity, orientation, distance) as observations. Then proposed a DBN-based system to control media player or slide presentation. They used local features (location, velocity) by skin extraction and motion tracking to design the DBN inference. However, both HMM and DBN models assume that observations given the motion class labels are conditional independent. This restriction makes it difficult or impossible to accommodate long-range dependencies among observations or multiple overlapping features of the observations. Therefore, proposed conditional random fields (CRF) which can avoid the independence assumption between observations and allow nonlocal dependencies between state and observations. Then incorporated hidden state variables into the CRF model, namely, hidden

conditional random field (HCRF). They used HCRF to recognize Sign Language recognition and proved that HCRF can get better performance. Later,the latent-dynamic conditional field (LDCRF) model was proposed, which combines the strengths of CRFs and HCRFs by capturing both extrinsic dynamics and intrinsic sub-structure. The detailed comparisons are evaluated. Another important approach is dynamic time warping (DTW) widely used in Sign Language recognition. Early DTW-based methods were applied to isolated Sign Language recognition. Then proposed an enhanced Level-Building DTW method. This method can handle the movement epenthesis problem and simultaneously segment and match signs to continuous sign language sentences. Besides these methods, other approaches are also widely used for Sign Language recognition, such as linguistic sub-units and topology-preserving self-organizing networks. Although the mentioned methods have delivered promising results, most of them assume that the local features (shape, velocity, orientation, position or trajectory) are detected well. However, the prior successes of hand detection and tracking are major challenging problems in complex surroundings. Moreover, as shown, most of the mentioned methods need dozens or hundreds of training samples to achieve high recognition rates. For example, the authors used at least 50 samples for each class to train HMM and got the average recognition rate 96%. Besides, Yamato et al. suggested that the recognition rate will be unstable if the number of samples is small. When there is only one training sample per class, those methods are difficult to satisfy the requirement of high performance application systems. In recent years, BoF-based methods derived from object categories and action recognition have become an important branch for Sign Language recognition. Dardas and Georganasproposed a method for real-timeSign Language recognition based on standard BoF model, but they first needed to detect and track hands and that would be difficult in a clutter background. For example, when the hand and face are overlapped or the background is similar to skin color, hand detection may fail. Shen et al. extracted maximum stableextremalregions (MSER) features from the motion divergence fields which In this paper, we propose a unified framework based on bag of features for one-shot learning Sign Languagerecognition. The proposed method gives superior recognition performance than many existing approaches. A new feature, named 3D EMoSIFT, fuses RGB-D data to detect interest points and constructs 3D gradient and motion space to calculate SIFT descriptors. Compared with existing features such as Cuboid, Harri3D, MoSIFT (Chen and Hauptmann) and 3D MoSIFT, it gets competitive performance. Additionally, 3D EMoSIFT

features are scale and rotation invariant and can capture more compact and richer video representations even though there is only one training sample for each Sign Language class. Thispaper also introduces SOMP to replace VQ in the descriptor coding stage. Then each feature can be represented by some linear combination of a small number of visual code words. Compared with VQ, SOMP leads to a much lower reconstruction error and achieves better performance. Although the proposed method has achieved promising results, there are several avenues which can be explored. At first, most of the existing local spatio-temporal features are extracted from a static background or a simple dynamic background. In our feature research, we will focus on extending 3D EMoSIFT to extract features from complex background, especially for one-shot learning Sign Language recognition. Next, to speed up processing time, we can achieve fast feature extraction on a Graphics Processing Unit (GPU). Also, we will explore the techniques required to optimize the parameters, such as the codebook size and sparsity.

**2.10 Title: Discovering Motion Primitives for Unsupervised Grouping and One-shot Learning of Human Actions, Sign Languages, and Expressions**

**Author: Yang Yang, Imran Saleemi**

Learning using few labeled examples should be an essentialfeature in any practical action recognition system because collection of a large number of examples for each of many diverse categories is an expensive and laborious task. Although humans are adept at learning new object and action categories, the same cannot be said about most existing computer vision methods, even though such capability is of significant importance. A majority of proposed recognition approaches require large amounts of labeled training data, while testing using either a leave-one-out or a train-test split scenario. In this paper,we put forth a discriminative yet flexible representation of Sign Languages and actions that lends itself well to the task of learning from few as possible examples. We further extend the idea of one-shot learning to attempt a perceptual grouping of unlabelled datasets and to obtain subsets of videos that correspond to a meaningful grouping of actions, for instance, recovering the original class-based partitions. This observation forms the basis of the proposed representation with the underlying idea that intermediate features (action primitives) should: (a) span as large as possible but contiguous $x-y-t$ volumes with smoothly varying motion, and (b) should be flexible enough to allow deformations arising from articulation of body parts. A byproduct of these properties is that the intermediate representation will be conducive to human understanding. In other words, a meaningful action primitive is one

which can be illustrated visually, and described textually, e.g., ‘left arm moving upwards’, or ‘right legmoving outwards and upwards’, etc.We argue and showexperimentally, that such a representation is much more discriminative, and makes the tasks of ‘few-shot’ action, Sign Language, or expression recognition, or unsupervised clustering simpler as compared to traditional methods. This paper proposes such a representation based on motion primitives. A summary of our method to obtain the proposed representation follows. (i) when required, camera motion is compensated to obtain residual actor-only motion, (ii) a frame difference based foreground estimation, and ‘centralization’ of the actor to remove translational motion is performed, thus resulting in a stack of rectangular image regions coarsely centered around the human; (iii) computation of optical flow to obtain 4d feature vectors ($x, y, u, v$); (iv) clustering of feature vectors to obtain components of a Gaussian mixture; (v) spatio-temporal linking of Gaussian components resulting in instances of primitive actions; and (vi) merging of primitive action instances to obtain final statistical representation of the primitives. For supervised recognition, given a test video, instances of action primitives are detected in a similar fashion, which are labeled by comparing against the learned primitives. Sequences of observed primitives in training and test videos are represented as strings and matched using simple alignment to classify the test video. We also experimented with representation of primitive sequences as histograms, followed by classifier learning, as well as using temporal sequences of primitive labels to learn state transition models for each class. Compared to the state of the art action representations the contributions of the proposed work are:Completely unsupervised discovery of representative and discriminative action primitives without assuming any knowledge of the number of primitives present, or their interpretation, A novel representation of human action primitives that captures the spatial layout, shape, temporal extent, as well as the motion flow of a primitive, Statistical description of primitives as motion patterns, thus providing a generative model, capable of estimating confidence in observing a specific motion at a specific point in space-time, and even sampling, Highly abstract, discriminative representation of primitives which can be labeled textually as components of an action, thus making the recognition task straightforward. This paper has proposed a method that automatically discovers a flexible and meaningful vocabulary of actions using raw optical flow learns statistical distributions of these primitives, and because of the discriminative nature of the primitives, very competitive results

are obtained using the simplest recognition and classification schemes. Our representation offers benefits like

recognition of unseen composite action, insensitivity to occlusions (partial primitive list), invariance to splittingof primitive during learning, detection of cycle extentsand number, etc.

## CHAPTER 3

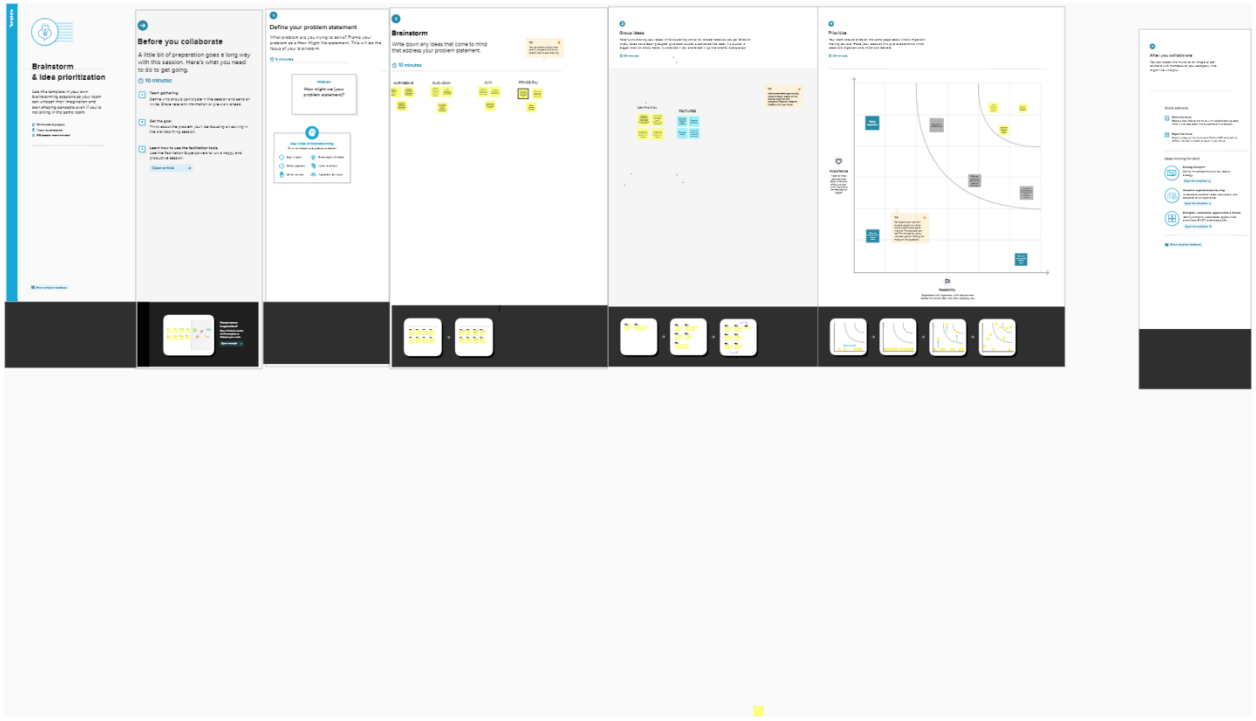
## 3. System Analysis:

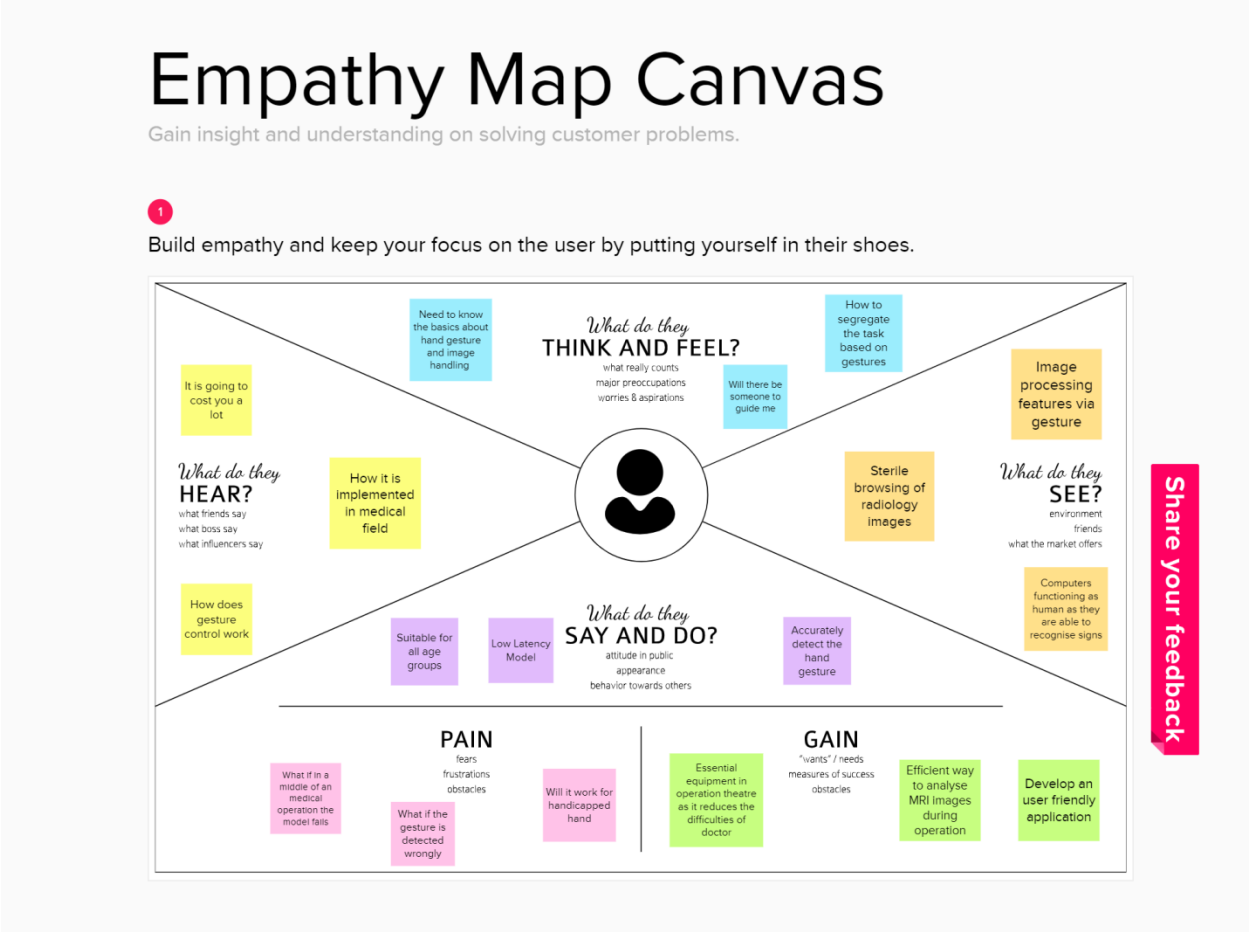
### 3.1 Existing system:

#### 3.1.1 Objective:

Sign Language is a form of non-verbal communication using various body parts, mostly hand and face. Sign Language is the oldest method of communication in human. Primitive men used to communicate the information of food/ prey for hunting, source of water, information about their enemy, request for help etc. within themselves through Sign Languages. Still Sign Languages are used widely for different applications on different domains. This includes human-robot interaction, sign language recognition, interactive games, vision-based augmented reality etc. Another major application of Sign Languages is found in the aviation industry for placing the aircraft in the defined bay after landing, for making the passengers aware about the safety features by the airhostess. For communication by the people at a visible, but not audible distance (surveyors) and by the physically challenged people (mainly the deaf and dumb) Sign Language is the only method. Posture is another term often confused with Sign Language. Posture refers to only a single image corresponding to a single command (such as stop), where as a sequence of postures is called Sign Language (such as move the screen to left or right). Sometimes they are also called static (posture) and dynamic Sign Language (Sign Language). Posture is simple and needs less computational power, but Sign Language (i.e. dynamic) is complex and suitable for real environments. Though sometimes face and other body parts are used along with single hand or double hands, Sign Language is most popular for different applications. With the advancement of human civilization, the difficulty of interpersonal communication, not only in terms of language, but also in terms of communication between common people and hearing impaired people is gradually being abolished. If development of sign language is the first step, then development of hand recognition system using computer vision is the second step. Several works have been carried out worldwide using Artificial Intelligence for different sign languages. The main objective is to perform effective recognition, detection and tracking of hands and a color and depth based 3-D particle filter framework is proposed to solve occlusion

# Ideation & brainstoming






### 3.1.2 Merits:

Surfing the web, typing a letter, playing a video game or storing and retrieving personal or official data are just a few examples of the use of computers or computer-based devices. Due to increase in mass production and constant decrease in price of personal computers, they will even influence our everyday life more in near future. Nevertheless, in order to efficiently utilize the new phenomenon, myriad number of studies has been carried out on computer applications and their requirement of more and more interactions. In existing system presents a novel technique for Sign Language recognition through human–computer interaction based on shape analysis. The main objective of this effort is to explore the utility of a particle filter-based approach to the recognition of the Sign Languages. The goal of static Sign Language recognition is to classify the given Sign Language data represented by some features into some predefined finite number of Sign Language classes. The proposed system presents a recognition algorithm to recognize a set of six specific static Sign Languages, namely: Open, Close, Cut, Paste,

Maximize, and Minimize. The Sign Language image is passed through three stages, preprocessing, feature extraction, and classification. In preprocessing stage some operations are applied to extract the Sign Language from its background and prepare the Sign Language image for the feature extraction stage. In the first method, the hand contour is used as a feature which treats scaling and translation of problems (in some cases). The complex moment algorithm is, however, used to describe the Sign Language and treat the rotation problem in addition to the scaling and translation.

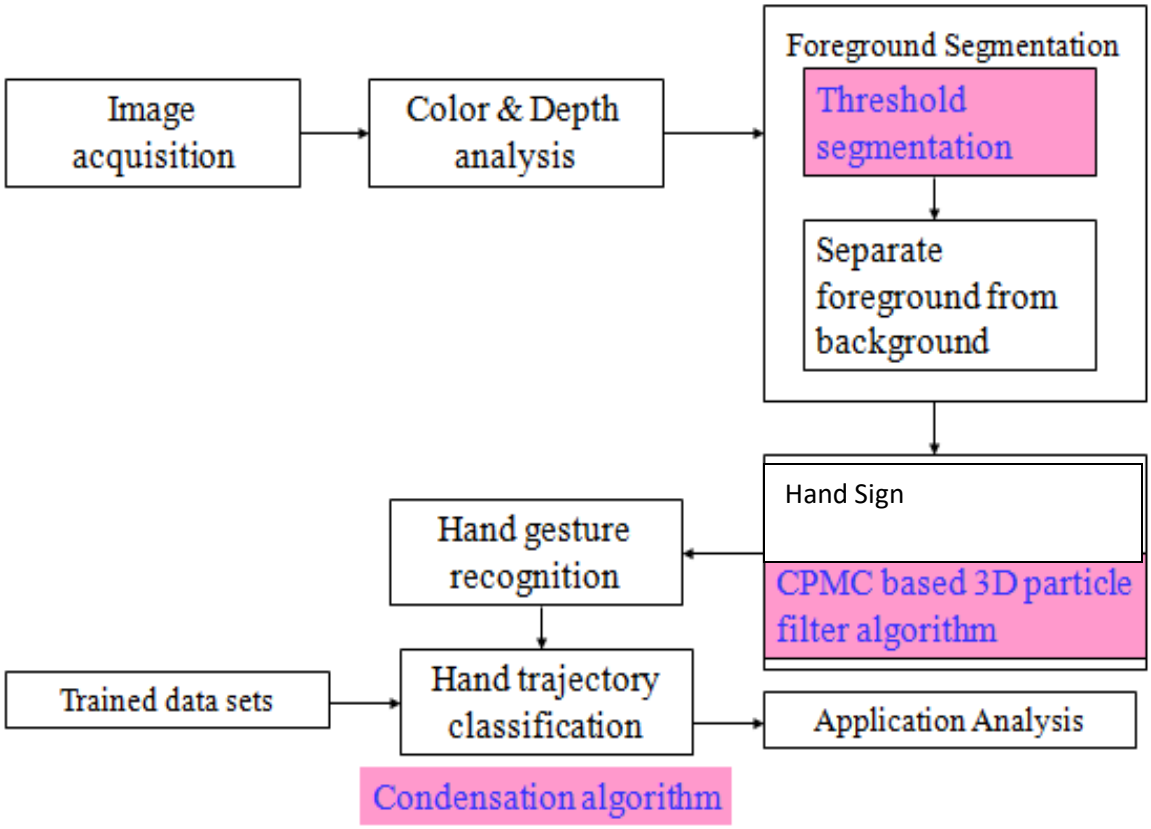
### 3.1.3 Challenges

There are many challenges associated with the accuracy and usefulness of Sign Language recognition software. For image-based Sign Language recognition there are limitations on the equipment used and image noise. Images or video may not be under consistent lighting, or in the same location. Items in the background or distinct features of the users may make recognition more difficult.

The variety of implementations for image-based Sign Language recognition may also cause issue for viability of the technology to general usage. For example, an algorithm calibrated for one camera may not work for a different camera. The amount of background noise also causes tracking and recognition difficulties, especially when occlusions (partial and full) occur. Furthermore, the distance from the camera, and the camera's resolution and quality, also cause variations in recognition accuracy.

In order to capture human Sign Languages by visual sensors, robust computer vision methods are also required, for example for hand tracking and hand posture recognition or for capturing movements of the head, facial expressions or gaze direction.

### 3.1.4 Architecture of Existing System



## CHAPTER 4

## 4. System Design:

### 4.1 Proposed system:

#### 4.1.1 Objective:

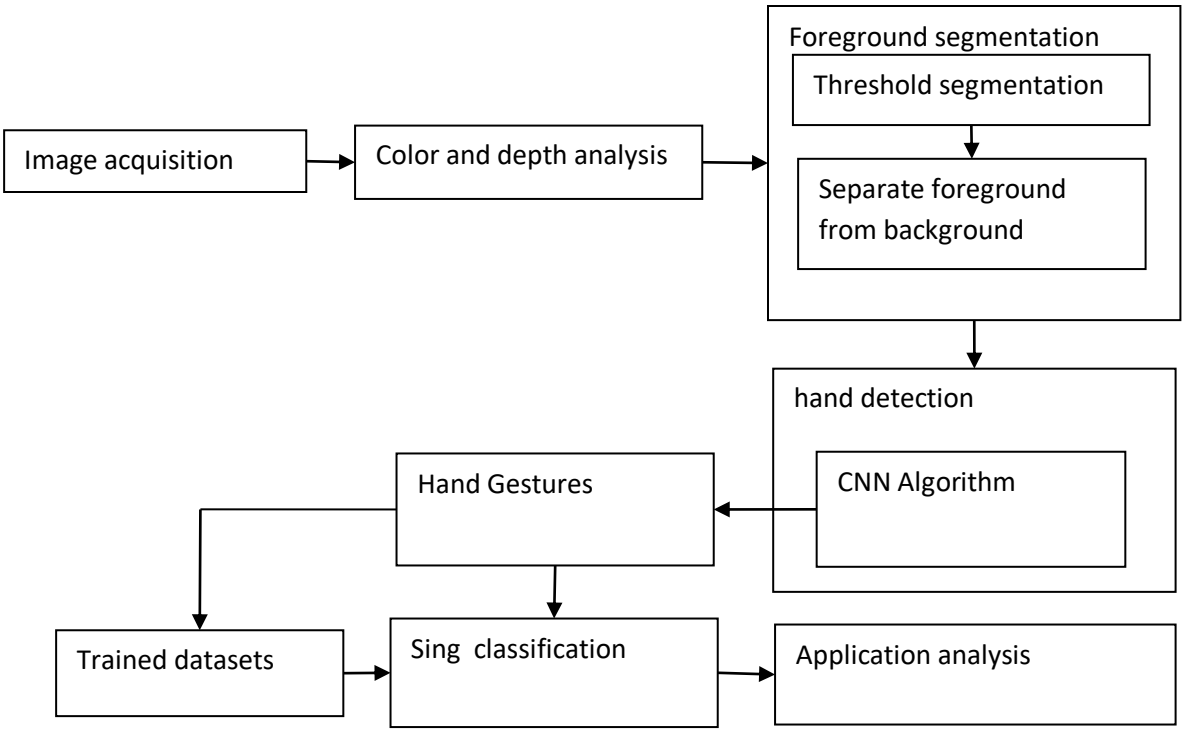
Sign Language was the first mode of communication for the primitive cave men. Later on human civilization has developed the verbal communication very well. But still non-verbal communication has not lost its weight age. Such non – verbal communication are being used not only for the physically challenged people, but also for different applications in diversified areas, such as aviation, surveying, music direction etc. It is the best method to interact with the computer without using other peripheral devices, such as keyboard, mouse. Researchers around the world are actively engaged in development of robust and efficient Sign Language recognition system, more specially, Sign Language recognition system for various applications. The major steps associated with the Sign Language recognition system are; data acquisition, Sign Language modeling, feature extraction and Sign Language recognition. The importance of Sign Language recognition lies in building efficient human–machine interaction. Its applications range from sign language recognition through medical rehabilitation to virtual reality. Given the amount of literature on the problem of Sign Language recognition and the promising recognition rates reported, one would be led to believe that the problem is nearly solved. Sadly this is not so. A main problem hampering most approaches is that they rely on several underlying assumptions that may be suitable in a controlled lab setting but do not generalize to arbitrary settings. Several common assumptions include: assuming high contrast stationary backgrounds and ambient lighting conditions.

**CNN Algorithm:**

Convolution is a phase through which a new product is combined and generated by two functions. We have to think of an image as a matrix of pixels when it comes to pictures. Each pixel has its own value, but it is combined with other pixels, and an image produces a result. To detect certain features in the image, CNN adds filters. The manner in which the convolutional neural network operates entirely depends on the type of filter applied. So, we can provide the

network with as many different features as possible when applying machine learning solutions to image classification. Then upon preparation, it will evaluate their principles. CNNs are composed of three types of layers. When such layers are piled, a CNN architecture has been created. The fundamental functionality of the above example of CNN can be broken down into four main areas

### 4.1.2 System architecture:

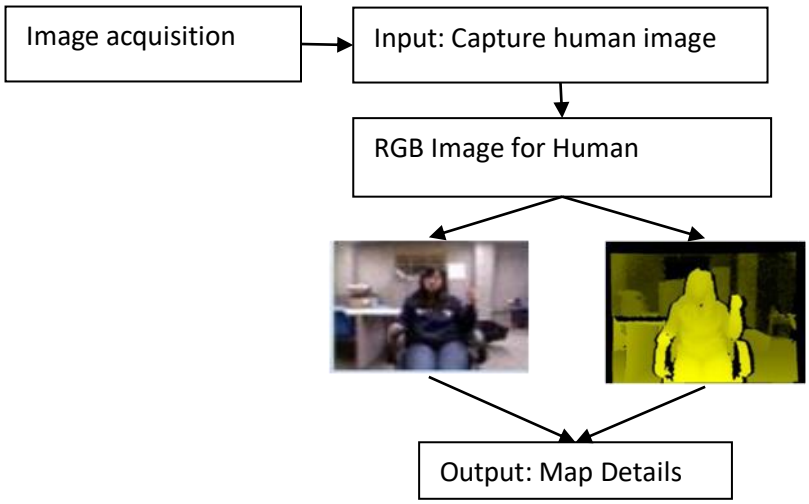


### 4.1.3 Module explanation:

- **Image Acquisition**
- **Foreground segmentation**
- **Face and Hand detection**
- **Hand trajectory classification**
- **Evaluation criteria**

#### 4.1.3.1 Image Acquisition:

For efficient Sign Language recognition, data acquisition should be as much perfect as possible. Suitable input device should be selected for the data acquisition. There are a number of input devices for data acquisition. Some of them are data gloves, marker, hand images (from webcam/ stereo camera/ Kinect 3D sensor) and drawings. Data gloves are the devices for perfect data input with high accuracy and high speed. It can provide accurate data of joint angle, rotation, location etc. for application in different virtual reality environments. At present, wireless data gloves are available commercially so as to remove the hindrance due to the cable. Colored markers attached to the human skin are also used as input technique and hand localization is done by the color localization. Input can also be fed to the system without any external costly hardware, except a low-cost web camera. Bare hand (either single or double) is used to generate the Sign Language and the camera captures the data easily and naturally (without any contact). Sometimes drawing models are used to input commands to the system. The latest addition to this list is Microsft Kinect 3D depth sensor. Kinect is a 3D motion sensing input device widely used for gaming. In this module, we can input image from web camera and also capture hand and face images. And captured both depth and color image.In 3D computer graphics a depth map is an image or image channel that contains information relating to the distance of the surfaces of scene objects from a viewpoint. The term is related to and may be analogous to depth buffer, Z-buffer, Z-buffering and Z-depth.The "Z" in these latter terms relates to a convention that the central axis of view of a camera is in the direction of the camera's Z axis, and not to the absolute Z axis of a scene. And use color map techniques to implement a function that maps (transforms) the colors of one (source) image to the colors of another (target) image. A color mapping may be referred to as the algorithm that results in the mapping function or the algorithm that transforms the image colors.

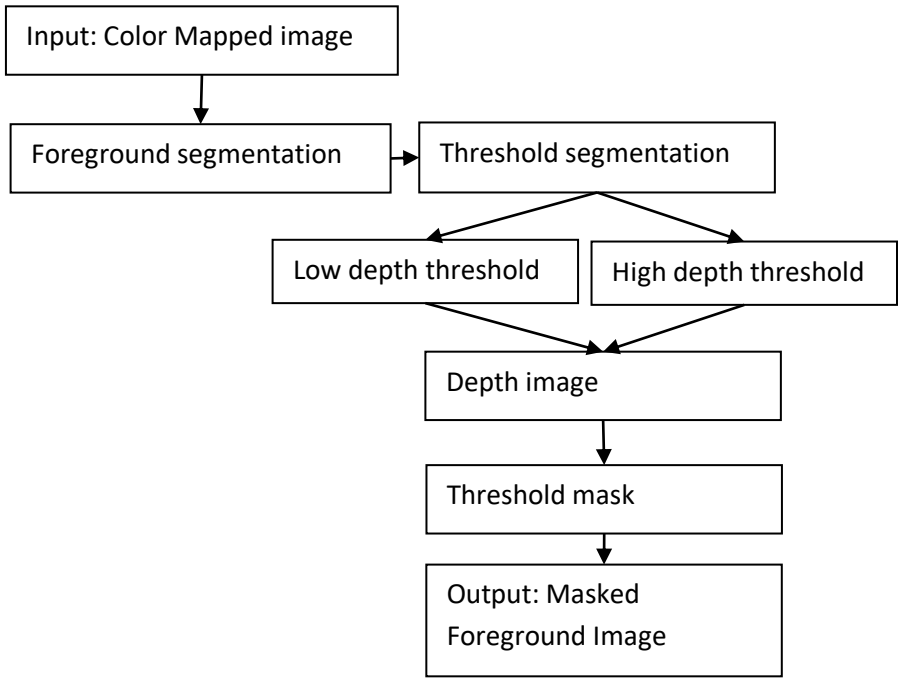



**4.1.3.2 Foreground segmentation:**

Separating foreground objects from natural images and video plays an important role in image and video editing tasks. Despite extensive study in the last two decades, this problem still remains challenging. In particular, extracting a foreground object from the background in a static image involves determining both full and partial pixel coverage, also known as extracting a matte, which is a severely under-constrained problem. Segmenting spatio-temporal video objects from a video sequence is even harder since extracted foregrounds on adjacent frames must be both spatially and temporally coherent. Previous approaches for foreground extraction usually require a large amount of user input and still suffer from inaccurate results and low computational efficiency.

In foreground segmentation section, the background was ruled out from the captured frames and the whole human body was kept as the foreground. In this module, we implement thresholding approach. In computer vision, image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as super pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics. Thresholding is the simplest segmentation method. The pixels are partitioned depending on their intensity value. Global thresholding, using an appropriate threshold T:

$$g(x,y) = \begin{cases} 1 \; iff(x,y) > T \\ 0 \; iff(x,y) \leq T \end{cases}$$




**4.1.3.3 Face and hand detection:**

Face and hand detection was used to initialize the position of the face and hands for the tracking phase. After initialization, both face and hands were tracked through video sequences by the MCMC based HMM method.

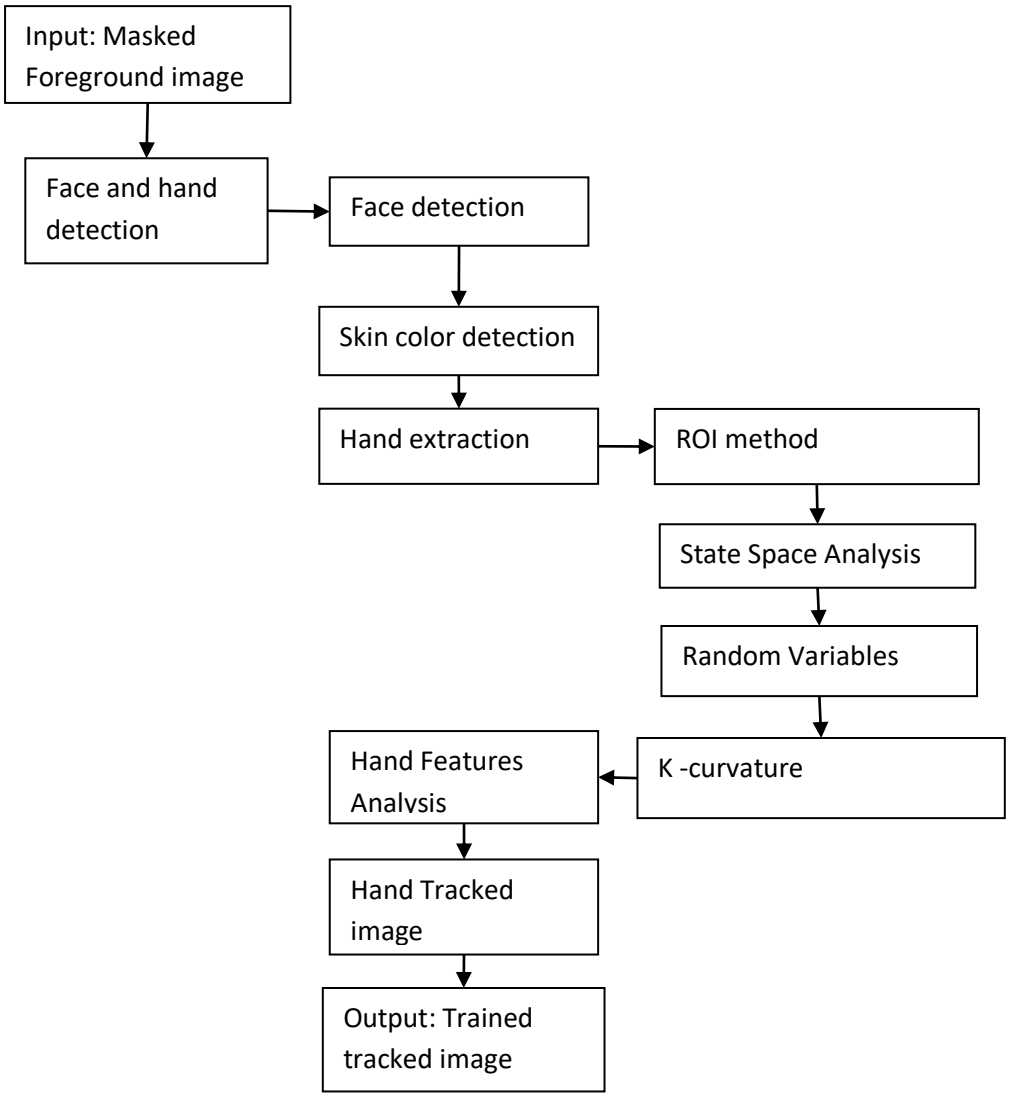
**Monto carlo markovian chain method:**

In mathematics, more specifically in statistics, ROI methods are a class of algorithms for sampling from aprobability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after a number of steps is then used as a sample of the desired distribution. The quality of the sample improves as a function of the number of steps.ROI methods make up a large subclass of MCMC methods. When an ROI method is used for approximating a multi-dimensional integral, an ensemble of "walkers" move around randomly. At each point where a walker steps, the integrand value at that point is counted towards the integral. The walker then may make a number of tentative steps

around the area, looking for a place with a reasonably high contribution to the integral to move into next. Random walk Monte Carlo methods are a kind of random simulation or ROI. However, whereas the random samples of the integrand used in a conventional Monte Carlo integration are statistically independent, those used in k-curvature methods are correlated. A ROI is constructed in such a way as to have the integrand as its equilibrium distribution.

**K-curvature model:**

A time-domain process demonstrates a Markov property if the conditional probability density of the current event, given all present and past events, depends only on the jth most recent event. If the current event depends solely on the most recent past event, then the process is termed a first order Markov process. This is a useful assumption to make, when considering the positions and orientations of the hands of a Sign Languager through time. The K-curvature is a double stochastic process governed by: 1) an underlying Markov chain with a finite number of states and 2) a set of random functions, each associated with one state. In discrete time instants, the process is in one of the states and generates an observation symbol according to the random function corresponding to the current state. Each transition between the states has a pair of probabilities, defined as follows: 1) transition probability, which provides the probability for undergoing the transition; 2) output probability, which defines the conditional probability of emitting an output symbol from a finite alphabet when given a state. The K-curvature is rich in mathematical structures and has been found to efficiently model spatio–temporal information in a natural way. The model is termed “hidden” because all that can be seen is only a sequence of observations. It also involves elegant and efficient algorithms, such as Baum–Welch and Viterbi, for evaluation, learning, and decoding.

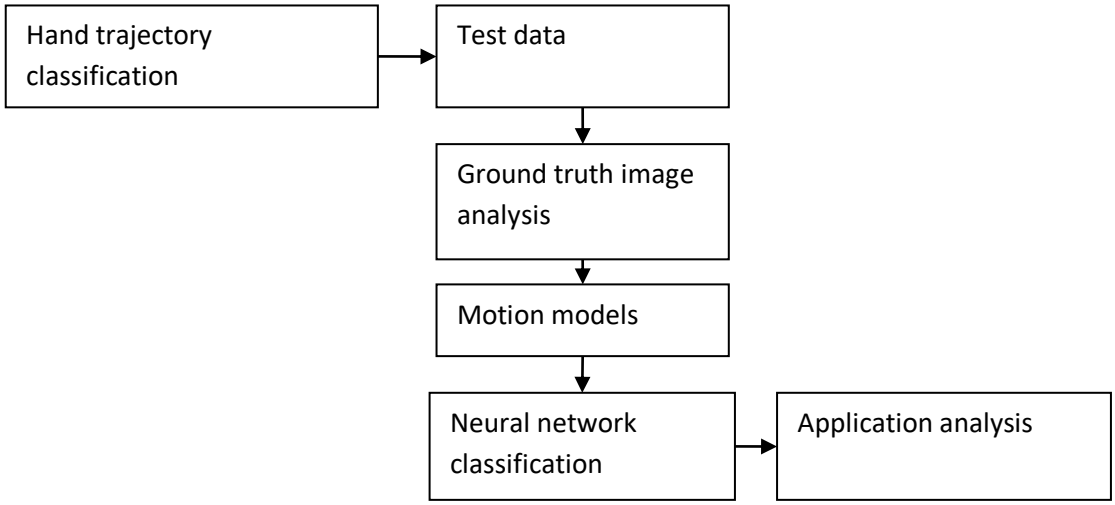



**4.1.3.4 Hand trajectory classification:**

Hand tracking results were segmented as trajectories, compared with motion models, and decoded as commands for robotic control.

Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the network function is determined largely by the connections between elements. We can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements. Commonly neural networks are adjusted, or trained, so that a particular input leads to a specific target output. There, the network is adjusted, based on a comparison of the output and the target,

until the network output matches the target. Typically many such input/target pairs are used, in this supervised learning (training method studied in more detail on following chapter), to train a network.



#### 4.1.3.5 Evaluation criteria:

The proposed system was able to detect finger tips even when it was in front of palm, it reconstruct the 3D image of hand that was visually comparable. This system claimed results 90-95% accurate for open fingers that is quite acceptable while for closed finger it was 10-20% only and closed or bended finger is coming in front of palm, so skin color detection would not make any difference in palm or finger. According to him image quality andoperator was the main reason for low detection and claims about 90% accuracy in the result, if the lighting conditions are good. Then used six different parameters to control the performance of system, if he found much noise there, he could control it using two parameters called as $\alpha$ and $\beta$ respectively. Finally claims about 90.45% accuracy, through hidden finger was not detected in his approach.

## 4.2 System specification:

### 4.2.1 Minimum hardware requirements

- CPU type : Intel Pentium 4
- Clock speed : 3.0 GHz
- Ram size : 512 MB
- Hard disk capacity : 40 GB
- Monitor type : 15 Inch color monitor
- Keyboard type : internet keyboard

### 4.2.2 Minimum software requirements

- Operating System : Windows OS
- Language : Python

**Functional Requirement**

As one of the biggest and most complex tech projects ever for automatic Sign Language recognition and translation, we keep receiving many questions and media inquiries. Ever since gesture recognition was invented, sign language recognition has been one of the enlisted motivations. Even if it is a camera based system or a glove with physical sensors, it is a quite natural idea to use those technologies to enable communication between Deaf and hearing folks.

Calling a system that is able to recognize a set of individual hand signs as a sign language translator is the true evidence of not understanding the challenge. It is venial for a student project – see the remarkable traction of Enable Talk gloves to win the Imagine Cup back in 2012. However, we think that there are many factors that are essential when pursuing a real mission, to provide real value for the Deaf community by respecting Deaf culture.

Here, we highlight the most important three:

Sentence level translation: just like spoken languages, sign languages are individual languages with their own linguistic structures. A system that recognizes separate signs one-by-one could only provide a translation in a situation where SEE (Signed Exact English) is provided. The whole translation and the used words might be completely changed by a sign that comes later in the sentence, like when the signed sentence turns out to be a question. Furthermore, forcing the signer to stop and wait for the result after every sign could result in poor user experience.

Tracking the face: all non-manual markers play a significant role in signing. Face is especially important as it supplies numerous sign locations and also because of facial expressions carrying grammatical meanings which need to be considered. No project that disregard the face can ever result in a real sign language translation.

Engagement of the Deaf Community: There is one true evidence that separates dreamers from real challengers: no good product can be developed without the involvement of the ultimate end users, and this is especially true in this special niche. If there are deaf members in a project, the team will soon realize the two factors above among many others. We believe that the involvement of deaf team members is the most important success factor in all.

Input Set(Hand Gestures): The image acquisition of the person transmitting in the signal language can be obtained by using a camera. It is necessary to initiate the acquisition manually. A camera sensor is necessary to capture the signer's movements. An algorithm for image recognition takes an image as an input and outputs what the image is made up of. In other words, the output is a class mark (for example, "'A Letter' Hand Sign Gesture," B" Letter' Hand Sign Gesture," "'C Letter' Hand Sign Gesture," etc.). In other words, a class mark is the output. The performance is, in other words, a class mark. The algorithm needs to be trained to understand the differences between different classes and to get to know the content of the image.If you want to find 'A letter' in pictures, you need to train an image recognition algorithm with thousands of 'A letter' images and thousands of background images that do not contain 'A letter'. Needless to mention, this algorithm can only understand objects / classes that it has learned. We concentrate on two-class classifiers (binary). Under the hood, several popular object detectors have a binary classifier, Inside a sign language detector is an image classifier that says whether a patch of an image is a hand gesture or a meaning.

Pre Processing To normalize contrast and brightness effects, the input image is pre-processed. An input image or patch of an image is often cropped and resized to a fixed size as part of the preprocessing. This is necessary because on a fixed sized image, the next step, feature extraction, is performed. Filtering is the first preprocessing phase. From the acquired image, unnecessary noise is removed. Background subtraction forms the next major stage. This processing results in a binary image in which white is colored with the pixels that form the hand and all the others are black. This processing includes the classification as part of human skin or not of each pixel of the image

Classification The form of the gesture being recognized is based on the camera location, the distance from the camera of the signer, etc. During real time execution, these methods must maintain a balance between precision and computational complexity. Image recognition is a method of marking and sorting objects by certain groups in the image. We need to educate it by showing thousands of examples and backgrounds before implementing a classification algorithm. Various learning algorithms view things differently, but the general concept is that learning algorithms treat function vectors as points in higher dimensional space, and try to find

planes/surfaces that divide the higher dimensional space in such a way that all examples relating to a certain class are on one side of theplane/surface. For classification purposes, we use the Convolutional Neural Network algorithm as it is the best and most reliable.

**Non Functional Requirements**

The conditions on which system should operate arespecified as non-functional requirements and they are:

1. Real time: the system should provide the recognition ofsigns and their translation to speech in an un noticeable time for its users.

2. Accuracy: signs should not be confused and the systemshould recognize appropriate sign.

3. Environment: the system should provide real time recognition with high accuracy in low light conditions as well.

4. Usability: the system should provide natural interaction to its users. The hearing-impaired person needs to worry nothing else, just for performing signs.

**Scope**

The current scenario in this field is that, lots of research have done and lots of white papers are there but there is no full-fledged Sign Language Interpreter which can translate signs to voice. This project is more inclined towards this particular side. This project will take the input from camera and will convert it to voice and voice again back to sign, as required. Depending upon the dataset and the architecture is used we can expect the accuracy close to 90%. Impaired person when surrounded by the relatives or close people he will not face much problem to communicate for both parties. The problem arises when he wants to communicate with the person who is stranger. Solution provided by us will be helpful in every scenario

## CHAPTER 5

## 5.1 CONCLUSION:

The design of more natural and multimodal forms of interaction with computers or systems is an aim to achieve. Vision-based interfaces can offer appealing solutions to introduce non-intrusive systems with interaction by means of Sign Languages. In order to build reliable and robust perceptual user interfaces based on computer vision, certain practical constraints must be taken in account: the application must be capable of working well in any environment and should make use of low-cost devices. This work has proposed a new mixture of several computer vision techniques for facial and hand features detection and tracking and face Sign Language recognition, some of them have been improved and enhanced to reach more stability and robustness. A hands-free interface able to replace the standard mouse motions and events has been developed using these techniques. Sign Language recognition is finding its application for non-verbal communication between human and computer, general fit person and physically challenged people, 3D gaming, virtual reality etc. With the increase in applications, the Sign Language recognition system demands lots of research in different directions. Finally we implemented effective and robust algorithms to solve false merge and false labeling problems of hand tracking through interaction and occlusion.

## 5.2 FUTURE ENHANCEMENT:

In future we present an idea of Sign Language recognition and Neural Networks approaches. One of the most effective of software computing techniques is Artificial Neural Networks that has many applications on Sign Language recognition problem. Some researches that handle Sign Language recognition problem using different neural networks systems are discussed with detailed showing their advantages and disadvantages. Comparison was made between each of these methods, as seen different Neural Networks systems are used in different stages of recognition systems according to the problem nature, its complexity, and the environment available. The input for all the selected methods was either digitized image camera or using data glove system. Then some preprocessing was made on the input image like normalization, edge detection filter, or thresholding which are necessary for segmenting the Sign Language from the background. Then feature extraction must be made, different methods presented in this paper, geometric features or non geometric features, geometric features that use angles and orientations, palm center, non geometric such as color, silhouette and textures, but they are inadequate in recognition. Neural Networks system can be applied for extracted features from the input image Sign Languages after applying segmentation to extract the shape of the hand.

**References:**


[1] M. R. Ahsan, "EMG signal classification for human computer interaction: A review," Eur. J. Sci. Res., vol. 33, no. 3, pp. 480–501, 2009.

[2] J. A. Jacko, "Human–computer interaction design and development approaches," in Proc. 14th HCI Int. Conf., 2011, pp. 169–180.

[3] I. H. Moon, M. Lee, J. C. Ryu, and M. Mun, "Intelligent robotic wheelchair with EMG-, Sign Language-, and voice-based interface," Intell.Robots Syst., vol. 4, pp. 3453–3458, 2003.

[4] M. Walters, S. Marcos, D. S. Syrdal, and K. Dautenhahn, "An interactive game with a robot: People's perceptions of robot faces and a Sign Language based user interface," in Proc. 6th Int. Conf. Adv. Computer–HumanInteractions, 2013, pp. 123–128.

[5] O. Brdiczka, M. Langet, J. Maisonnasse, and J. L. Crowley, "Detection human behavior models from multimodal observation in a smart home," IEEE Trans. Autom. Sci. Eng., vol. 6, no. 4, pp. 588–597, Oct. 2009.

[6] M. A. Cook and J. M. Polgar, Cook & Hussey's Assistive Technologies: Principles and Practice, 3rd ed. Maryland Heights, MO, USA: MosbyElsevier, 2008, pp. 3–33.

[7] G. R. S. Murthy, and R. S. Jadon, "A review of vision based Sign Language recognition," Int. J. Inform. Technol. Knowl. Manage., vol. 2, no. 2, pp. 405–410, 2009.

[8] D. Debuse, C. Gibb, and C. Chandler, "Effects of hippotherapy on people with cerebral palsy from the users' perspective: A qualitative study," Physiotherapy Theory Practice, vol. 25, no. 3, pp. 174–192, 2009.

[9] J. A. Sterba, B. T. Rogers, A. P. France, and D. A. Vokes, "Horseback riding in children with cerebral palsy: Effect on gross motor function," Develop. Med. Child Neurology, vol. 44, no. 5, pp. 301–308, 2002.

[10] K. L. Kitto, "Development of a low-cost sip and puff mouse," in Proc. 16th Annu Conf. RESNA, 1993, pp. 452–454.

[11] Y. H. Yin, Y. J. Fan, and L. D. Xu, "EMG and EPP-integrated human– machine interface between the paralyzed and rehabilitation exoskeleton," IEEE Trans. Inf. Technol. Biomed., vol. 16, no. 4, pp. 542–549, Jul. 2012.

[12] H. Jiang, J. P. Wachs, and B. S. Duerstock, "Facilitated Sign Language recognition based interfaces for people with upper extremity physical impairments," in Proc. Pattern Recogn., Image Anal., Comput. Vision,Applicat., 2012, pp. 228–235.

[13] J. Wachs, M. K¨olsch, H. Stern, and Y. Edan, “Vision-based Sign Language applications: Challenges and innovations,” Commun. ACM, CoverArticle, vol. 54, no. 2, pp. 60–71, 2011.

[14] Z. Li and R. Jarvis, “A multimodal Sign Language recognition system in ahuman–robot interaction scenario,” in Proc. IEEE Int. Workshop RoboticSensors Environments, 2009, pp. 41–46.

[15] E. A. Suma, B. Lange, A. Rizzo, D. M. Krum, and M. Bolas, “FAAST:The flexible action and articulated skeleton toolkit,” in Proc. IEEEVirtual Reality Conf., Mar. 2011, pp. 247–248.

[16] Leap Motion [Online]. Available: https://www.leapmotion.com/

[17] G. R. Bradski, “Computer vision face tracking as a component of aperceptual user interface,” in Proc. Workshop Applicat. Comput. Vision,1998, pp. 214–219.

[18] M. Isard and A. Black, “Condensation: Conditional density propagationfor visual tracking,” J. Int. J. Comput. Vision, vol. 29, no. 1, pp. 5–28,1998.

[19] S. Maskell, N. Gordon, and T. Clapp, “A tutorial on particle filters foronline nonlinear/non-Gaussian Bayesian tracking,” IEEE Trans. SignalProcess., vol. 50, no. 2, pp. 174–188, Feb. 2002.

[20] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, Color-Based ProbabilisticTracking, vol. 2350. Heidelberg, Germany: Springer, pp. 661–675,2002.

[21] K. Okuma, A. Taleghani, N. Freitas, J. J. Little, and D. G. Lowe, “Aboosted particle filter: Multitarget detection and tracking,” in Proc. Eur.Conf. Comput. Vision, 2004.

[22] M. Kristan, J. Pers, S. Kovacic, and A. Leonardis, “A local-motion-basedprobabilistic model for visual tracking,” Pattern Recogn., vol. 42, no. 9,pp. 2160–2168, 2009.

[23] W. Qu, D. Schonfeld, and M. Mohamed, “Real-time distributed multiobjecttracking using multiple interactive trackers and a magnetic-inertiapotential model,” IEEE Trans. Multimedia, vol. 9, no. 3, pp. 511–519,Apr. 2007.

[24] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, “Articulatedhuman pose estimation and search in (almost) unconstrained still,”ETH Zurich Tech. Rep. 272, Sep. 2010.

[25] Y. Yang, and D. Ramanan, “Articulated pose estimation with flexiblemixture of parts,” in Proc. IEEE Conf. Comput. Vision Pattern Recogn.,Jun. 2011, pp. 1385–1392.

[26] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R.Moore, A. Kipman, and A. Blake, “Real-time human pose recognitionin parts from single depth images,” in Proc. IEEE Conf. Comput. VisionPattern Recogn., Jun. 2011, pp. 1297–1304.

[27] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, “Tracking the articulatedmotion of two strongly interacting hands,” in Proc. IEEE Conf.Comput. Vision Pattern Recogn.,Jun. 2012, pp. 1862–1869.

[28] S. Bilal, R. Akmeliawati, A. A. Shafie, and M. J. E. Salami, “HiddenMarkov model for human to computer interaction: A study on humanSign Language recognition,” Artificial Intell., 2011, pp. 1–22.

[29] B. W. Miners, O. A. Basir, and M. S. Kamel, “UnderstandingSign Languages using approximate graph matching,” IEEE Trans. Syst.,Man, Cybern. A, Syst., Humans, vol. 35, no. 2, pp. 239–248,Mar. 2005.

[30] Z. Xu, C. Xiang, Y. Li, V. Lantz, K. Wang, and J. Yang, “A frameworkfor Sign Language recognition based on accelerometer and EMG sensors,”IEEE Trans. Syst., Man, Cybern. A, Syst. Humans, vol. 41, no. 6, pp.1064–1076, Nov. 2011.

[31] M. J. Black and A. D. Jepson, “A probabilistic framework for matchingtemporal trajectories: CONDENSATION-based recognition of Sign Languageand expressions,” in Proc. Eur. Conf. Comput. Vision, 1998,pp. 909–924.

[32] J. Alon, V. Athitsos, and W. Yuan, “A unified framework forSign Language recognition and spatiotemproal Sign Language segmentation,” IEEETrans. Pattern Anal. Mach. Intell., vol. 31, no. 9, pp. 1685–1699,Sep. 2009.

[33] I. Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner,Results and Analysis of the ChaLearn Sign Language Challenge. 2012.

[34] L. Fei-Fei, R. Fergus, and P. Perona, “One-shot learning of objectcategories,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 4,pp. 594–611, Apr. 2006.

[35] D. Wu, F. Zhu, and L. Shao, “One shot learning Sign Language recognitionfrom RGBD images,” in Proc. IEEE Comput. Soc. Conf. Comput. VisionPattern Recogn. Workshop Sign Language Recogn., Jun. 2012, pp. 7–12.