SIGN LANGUAGE RECOGNITION

Tarun Singh

ABSTRACT

Sign language is used by deaf and hard hearing people to exchange information between their own community and with other people. Computer recognition of sign language deals from sign gesture acquisition and continues till text/speech generation. Sign gestures can be classified as static and dynamic. However static gesture recognition is simpler than dynamic gesture recognition but both recognition systems are important to the human community. The sign language recognition steps are described in this survey. The data acquisition, data preprocessing and transformation, feature extraction, classification and results obtained are examined. Some future directions for research in this area also suggested.

Keywords: sign language recognition, hand tracking, hand gesture recognition, gesture analysis, face recognition.

1. INTRODUCTION

Sign language (SL) [1] is a visual-gestural language used by deaf and hard-hearing people for communication purposes. Three dimensional spaces and the hand movements are used (and other parts of the body) to convey meanings. It has its own vocabulary and syntax which is purely different from spoken languages/written language. Spoken languages use the oratory faculties to produce sounds mapped against specific words and grammatical combinations to convey meaningful information. Then the oratory elements are received by the auditory faculties and processed accordingly. Sign language uses the visual faculties which is different from spoken language. Spoken language makes use of rules to produce comprehensive messages; similarly sign language is also governed by a complex grammar. A sign language recognition system consists of an easy, efficient and accurate mechanism to transform sign language into text or speech. The computerized digital image processing and a wide variety of classification methods used to recognize the alphabet flow and interpret sign language words and phrases. Sign language information can be conveyed using gestures of hands, position of head and body parts. Four essential components in a gesture recognition system are: gesture modeling, gesture analysis, gesture recognition and gesture-based application systems [2].

1.1. Indian sign language: history

Professionals in India believe in an acute shortage of special schools for deaf people. A very few schools use sign language as a medium of instruction. There is also a lack of proper and effective audio visual support in oral education in these schools. This results in inadequate communication and language skills in the majority of deaf children, impacting on poor literacy skills in the deaf community. The reality is that deaf schools mainly do not use ISL and nearly 5% of deaf people [3] attend deaf schools. The use of ISL is restricted only to vocational programs and short term courses. ISL was partly influenced by British Sign Language in the finger spelling system and some other signs, but most are unrelated to European sign system.

There was no formal ISL until 1978. Banerjee [4] compared the signs used in some schools for the deaf located in West Bengal and Assam. His conclusion was that the gestures used in each school were not the same. He believed that signing started in India in the 18th century but its use was strongly discouraged. Madan Vasishta [5] sent a questionnaire to the heads of more than hundred schools for the deaf in India in 1975. Almost all the respondents agreed that there was no ISL. But they also acknowledged that deaf children used some kind of gestures. A similar survey was conducted 20 years later, using a set of questionnaires sent to deaf schools. Some of the responses showed the same misconceptions about sign language that signing is "based on spoken language", or "based on English", or "difficult to provide a sign for every spoken word". Some statements showed that a more positive attitude towards manual communication, and here respondents talked about sign language, rather than gestures. Increasing awareness about the nature of sign languages was verified later on.

Observing the advantages of works on sign language recognition of different countries in aiding the deaf people for communication in public places and accessing/communicating with latest gadgets like Telephone, Computers, etc, linguistic studies on Indian Sign Language started in 1978 in India. These works resulted in ISL and discovered that it is a language on its own with specific syntax, grammar, phonology and morphology. Static numerals of ISL isolated gestures are shown in Figure-1.

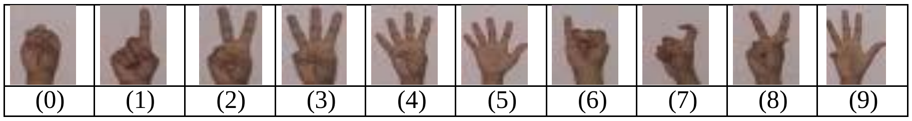

Figure-1. ISL static gestures for isolated numerals.

While significant progress has already been made in computer recognition of sign languages of other countries but a very limited work has been done in ISL computerization [6].

The works carried by various researchers worldwide are summarized in this paper. The domain is isolated sign language but continuous sign language recognition is also discussed due to similarity with isolated sign language recognition. We also select research papers where no special image acquiring devices are required. The reason is that in common places no special image acquiring devices are available at all the times, and all deaf/mute/hard hearing persons might be unable to wear due to their economic conditions and in most cases it is cumbersome to carry and wear. Also we select few research papers in which special wearable devices are used as inputs due to their better performance for comparison purposes.

The organization of the paper is as follows. We are summarizing the research papers from various authors according to following characteristics:

a) Sign language used.
b) The domain of the sign language used.
c) Data acquisition methods employed.
d) Data transformation techniques.
e) Feature extraction methods.
f) Classification techniques.
g) Results obtained.
h) Conclusion.

2. SIGN LANGUAGES USED

It is reported that about 5% of world population consists of deaf mute and hard hearing people. They used some kind of hand, head, and body gesture to exchange their feelings/ideas. So almost all nation have its own Sign Language. The sign language development is different for each country or sub-continent.

Table-1. Major sign languages of the world.

S. No.	Country/ sub-continent	Sign Language	Abbn.	No. of papers included
1	United Kingdom	British Sign Language	BSL	NIL
2	United States of America	American Sign Language	ASL	13
3	Commonwealth of Australia	Australian Sign Language	Auslan	NIL
4	Japan	Japanese Sign Language	JSL	NIL/1
5	People's Republic of China	Chinese Sign Language	CSL	5
	Taiwan	Taiwanese Sign Language	TSL	1
6	Middle-East	Arabic Sign Language	ArSL	2
	Islamic Republic of Iran and other Gulf countries	Persian Sign Language	PSL	2
7	Republic of India	Indian Sign Language	ISL	NIL/2
8	Socialist Republic of Vietnam	Vietnam Sign Language	VSL	1
9	Ukraine	Ukrainian Sign Language	UKL	1
10	Democratic Socialist Republic of Sri Lanka	Sri Lankan Sign Language	SLTSL	1
11	Federative Republic of Brazil	Brazilian Sign Language (Lingua Brasileira de Sinais)	Libras	1
12	Republic of Poland (Rzeczpospolita Polska)	Polish Sign Language (Polski Jezyk Migowy)	PJM	1
13	The Netherlands (Nederland)	Nederlandse Gebarentaal or Sign Language of the Netherlands	NGT/ SLN	1

The Table-1 represents the sign languages of influencing countries/sub-continent. The Table-1 indicates the most dominating research is going on ASL, next comes CSL and others follows. The reason is that a large number of standard database for ASL gesture are available publicly. The developing countries are currently focuses on the research in this field. Although two research papers from India are reported in this survey but the work was performed on ASL. We also include two survey papers on ISL.

3. THE DOMAIN OF THE SL USED

SL is an independent language which is entirely different from spoken/written language. It has its own set of alphabet, numeral, word/phrases/sentences and so on. The basic difference is that it has limited vocabulary compared to written/spoken language. Also in most of the developing countries and under developed countries it is in the initial phase. The development of the sign language in these countries will take years to become an independent language. But the computer recognition for sign language for these countries is started and significant works are reported in literature.

A Sign Language has a set of alphabets and is the same to the written/spoken language of the country it belongs to. If we consider the case of ASL or BSL it is nothing but the alphabet set A to Z. Similarly the numerals 0 to 9 are communicated by any sign language [2, 7, 8 and 9]. Secondly the words/phrases of any sign language belong to a particular domain. Examples are "Why? ","Award ","What for?", "How much? [10]; a coin, cigarette, flower; reluctantly, row, take, immediately, understand, hate, left, seven, moon, eight, walk, conscience [11] and other set used like Friend, To Eat, Neighbor, To sleep, Guest, To Drink, Gift, To wake up, Enemy, To listen, Peace upon you, To stop talking, Welcome, To smell, Thank you, To help, Come in, Yesterday, Shame, To go, House, To come and I/me [12]. The main aim is that when a researcher wants to produce a system of recognition of sign language he/she used a set of words/phrases in a particular domain like banking, railways, public telephone booths or something that focuses very general conversations in public places. Thirdly combinations of sign gestures for simple sentences/phrases are used in recognition of sign languages.

The databases used by various researchers are classified according to:

- Availability of standard database
- Creating own database

3.1. Availability of standard database

The standard databases used by various researchers are available in various libraries. The following library data in Table-2 are included in this survey.

Table-2. Standard database for sign language/face.

Library database	**Sign language**
Lifeprint Fingurespell Library	ASL
CAS-PEAL	CSL
Extended Yale B Yale B frontal (Subset of extended Yale B)	Face Database
Weizmann face database	Face Database
American Sign Language Linguistic Research Project with transcription using SignStream	ASL
ASL Lexicon Video Dataset	ASL
eNTERFACE	ASL
PETS 2002	Similar to PSL
RWTH-BOSTON-104 Database	ASL

3.1.1. Life print Fingurespell Library

American Sign Language University provides [9, 13] online sign language instruction from the year 1997. The program is as an effort to support parents and relatives of deaf-mute children living in rural areas where access to sign language programs is limited. The technical details about how the database is acquired are not available in the literature. However it has a rich library having all types of dataset ranging from static alphabets to simple and complex phrases including medical terms and up to advanced phrases.

3.1.2. CAS-PEAL database

CAS-PEAL face database [14] was developed by Joint Research and Development Laboratory (JDL) for Advanced Computer and Communication Technologies of Chinese Academy of Sciences (CAS), under the support of the Chinese National Hi-Tech Program and the ISVISION Tech. Co. Ltd. The construction of the CAS-PEAL face database was aimed for providing the researchers a large-scale Chinese face database for studying, developing, and evaluating their algorithms. The CAS-PEAL large-scale face images with different sources of variations, like Pose, Expression, Accessories, and Lighting (PEAL) were used to advance the state-of-the-art face recognition technologies.

The database contains 99, 594 images from 1040 individuals (595 males and 445 females). For each subject of the database, nine cameras with equal spaced in a horizontal semicircular layer were setup to capture images across different poses in one shot. The person who was used to perform sign gestures also asked to look up and down to capture 18 images in another two shots. The developers also considered five kinds of expressions, six kinds accessories (three goggles, and three caps), and fifteen lighting directions, also with varying backgrounds, distance from cameras, and aging.

A specially designed photographic room was designed in the JDL Lab CAS. The room size was about 4m×5m×3½m. Some special apparatus were configured in the room including multiple digital cameras, all kinds of lamps, accessories in order to capture faces with different poses, expression, accessories, and lighting.

The camera system consists of nine digital cameras and a specially designed computer. All the nine cameras were placed in a horizontal semicircular layer with radius and height being 0.8m and 1.1m respectively. The cameras used in the experiments were web-eye PC631 with 370, 000 pixels CCD. They were all pointed to the center of the semicircular layer and labeled as 0 to 8 from the subject signer’s right to left.

All of the nine cameras were connected and controlled by the same computer through USB interface. The computer had been specially designed to support nine USB ports. A software package was designed to control the cameras and capture images simultaneously in one shot. In each shot, the software package obtained nine images of the subject across different poses within no more than two seconds and stores these images in the hard disk using uniform naming conventions.

A lighting system was designed using multiple lamps and lanterns. To simulate the ambient lighting, two photographic sunlamps of high power covered with ground glass were used to imitate the indoor lighting environment.

Fluorescent lamps were roughly arranged as ‘lighting sources’ to form the varying lighting conditions. The lamps were arranged in a spherical coordinate, whose origin is the center of the circle, which matched with the semicircular layer. Fifteen fluorescent lamps were used at the ‘lamp’ positions, which are uniformly located at specific five azimuths (-90^o,-45^o, 0^o, +45^o, +90^o) and three elevations (-45^o, 0^o, +45^o). By turning on/off each lamp, different lighting conditions are replicated. To decrease the labor cost and time, a multiplex switch circuit was used to control the on/off of these lamps.

Several types of glasses and hats were used as accessories to further increase the diversity of the database. The glasses consisted of dark frame glasses, thin and white frame glasses, glasses without frame. The hats also had brims of different size and shape.

Face images were captured with a blue cloth as the default background. But in practical applications, many cameras were working under the auto-white balance mode, which changed the face appearance.

3.1.3. Yale B frontal

To capture the images in this database [15], the geodesic lighting rig was used with 64 computer controlled xenon strobes whose positions in spherical coordinates to capture the images. The illumination with the rig was modified at frame rate and images were captured under variable illumination and pose. Images of ten persons were acquired under sixty four different lighting conditions in nine poses (a frontal pose, five poses at 12^o, and three poses at 24^o from the camera axis). The sixty four images of a face in a particular pose were acquired in two seconds.

The original size of the images was 640 × 480 pixels. In experiments [16], all images were manually cropped to include only the face with as little hair and background.

3.1.4. Extended Yale B

In the database [15, 17], forty five images were captured in different lighting directions in average. The authors randomly generated configurations of five lighting directions among these forty five different lighting directions, and the corresponding five images were taken to form the basis vector of a subspace. For each randomly generated lighting configuration, there were five images for training and forty images for testing. They randomly generated 16, 000 different configurations of five lighting positions, and this number corresponds to roughly 1.5 percent of the total number of configurations C (45, 5) = 1; 221; 759.

By randomly picking lighting conditions to form subspaces, the expected error rate should be an order of magnitude larger than the configuration. Most impressive fact that there were only three lighting configurations (out of the total 16, 000 tested configurations) that performed better than the configuration. These three configurations all shared the same basic pattern with the configuration in the spatial distribution of their lighting directions, is a frontal lighting direction coupled with four lateral directions.

3.1.5. Weizmann face database

In acquiring face images [17] about 1000 random coordinates were selected in a single acquired image. Sixteen different, overlapping 5×5 patches around each coordinate were then used to produce a subspace by taking their four principal components. These were stored in the subspace database. Additionally all sixteen patches were stored for the patch database. A novel test image was subdivided into a grid of non-overlapping 5×5 patches. For each patch the database was searched for a similar point patch using the exact sequential and point in ANN based search. The selected database patch was then used as an approximation to the original input patch. Similarly, both exact sequential and ANN searches were used to select a matching subspace in the subspace database for each patch. The point on the selected subspace, closest to the query patch, was then taken as its approximation.

3.1.6. Sign stream

Sign Stream [18, 19], a multimedia database tool distributed on a nonprofit basis to educators, students, and researchers and it provides a single computing environment within which to view, annotate, and analyze digital video and/or audio data. It also provides direct on-screen access to video and/or audio files and facilitates detailed and accurate annotation, making it valuable for linguistic research on signed languages and the gestural component of spoken languages. The database is useful in different domains involving annotation and analysis of digital video data.

Each facility was equipped with multiple synchronized digital cameras to capture different views of the subject data. The facilities were able to capture four simultaneous digital video streams at up to 85 frames per second, while storing the video to disk for editing and annotation. A substantial corpus of ASL video data from native signers was collected and is now publicly available for research purpose. The data acquisition consist the followings:

a. Four PCs, each with a 500-MHz Pentium III processor, 256 MB RAM; 64 GB of hard drive storage, and Bit Flow Road Runner video capture cards.
b. Four Kodak ES-310 digital video cameras. Each camera was connected to any one PC.

c. A video sync generator which was used to synchronize the cameras. Videos were captured in 30, 60, or 85 frames per second.
d. An Ethernet switch allows the four PCs to communicate with each other efficiently.
e. IO Industries 'Video Savant software' installed on all PCs in order to synchronize video capture across the four cameras.
f. Various illumination sources, dark (black) backgrounds, chairs for subjects, and so on.

Out of the four PCs, one was designated as a server and other three PCs were act as clients. To capture a video sequence the appropriate program was executed on server PC and corresponding client programs run on client PCs. Instructions were given to the server about how many frames were to be captured and the starting time of the recording. The captured frames ware then stored in the hard drives in real time mode. With 64 GB of hard drive storage available, continuously video could be recorded for 60 min, at 60 frames per second, in all four machines simultaneously, at an image resolution of 648×484 (width × height).

Video sequences had been collected with four video cameras configured in two different ways:

a) All cameras were focused on a single ASL subject signer. Two cameras make a stereo pair, facing toward the signer and covering the upper half of the signer's body. One camera faces toward the signer and zooms in on the head of the signer. One camera was placed on the side of the viewer and covers the upper half of the subject signer's body.
b) The cameras were focused on two ASL subject signers engaged in conversation, facing each other. In this setup, the cameras stand low on tripods placed in the middle (between the two subject signers, but not obstructing their conversation).

One pair of cameras was focused so as to give a close-up facial view of each subject signer. The other pair of cameras was facing one toward each subject signer, and covering the upper half of the subject signer's body. The video data are now available in both uncompressed and compressed formats. Significant portions of the collected data were also being linguistically annotated using Sign Stream, and these data and the associated Sign Stream annotations publicly available through Internet (Refer to http://www.bu.edu/asllrp/ncslgr.html).

3.1.7. eNTERFACE

A single web camera with 640×480 resolution and 25 frames per second rate was used for the recordings of signs in ASL. The camera was placed in front of the subject signer. Eight subject signers performed five repetitions of each sign and the video data were collected [20] in the database. The database was divided into training and test sets. For training 532 examples are used for training (28 examples per sign) and 228 examples for reporting the test results (12 examples per sign). The subjects in training and test sets were different except for one subject whose examples are divided between training and test sets. Equal numbers of sign classes were used training and testing sets. The authors applied a stratified seven-fold cross validation (CV) on the training sets where a validation set was needed. Sign language features were extracted from both manual signs (hand motion, hand shape, hand position with respect to face) and non-manual signs (head motion). The center of mass (CoM) of each hand was tracked and filtered by a Kalman Filter for hand motion analysis. Appearance-based shape features were calculated on the binary hand images for hand shape features. It includes the parameters of an ellipse fitted to the binary hand and statistics from a rectangular mask placed on top of the binary hand. The system detects rigid head motions such as head rotations and head nods for head motion analysis [21]. The orientation and velocity information of the head and the quantity of motion were also used as head motion features. More details can be found in [20] for further study.

3.1.8. ASL lexicon video dataset

The authors introduced a new large-scale dataset, the ASL Lexicon Video Dataset [11, 18, 22, 23 and 24], containing video sequences of a large number of distinct sign classes of ASL. This dataset is publicly available now and expanding rapidly. The authors believe that this dataset will be an important resource for researchers in sign language recognition and human activity analysis, by providing a large amount of data that can be used for training and testing, and by providing a public benchmark dataset on which different methods can be evaluated. The dataset is currently a part of a computer vision system that allows users to look up the meaning of a sign automatically. The authors suggested that the dataset can be used for testing a wide variety of computer vision, machine learning, and database indexing algorithms. The signer performs the sign in front of a camera (or, possibly, in a multi-camera set up), and the computer retrieves and displays the most similar signs in the lexicon dataset in this system.

The video sequences were captured simultaneously from four different cameras, providing four views namely a side view, two frontal views, and a view zoomed in on the face of the signer. The upper body occupies a relatively large part of the visible scene in both the side views and two frontal views. A frontal view of the face occupies a large part of the image in the face views. Video was captured at 60 frames per second, non-interlaced, at a resolution of 640×480 pixels per frame; for the side view, first frontal view, and face view. For the second frontal view, video was captured at 30 frames per second, non-interlaced, at a resolution of 1600×1200 pixels per frame. This high-resolution frontal view facilitated the application of existing hand pose estimation and hand tracking systems on the dataset, by displaying the hand in significantly more detail than in the 640×480 views.

The authors applied a motion energy method, using a test set of 206 video sequences belonging to 108 distinct glosses (used as class labels) and a training set of 999 video sequences belonging to 992 distinct glosses. For almost all sign classes the authors had only one training example. The worst possible ranking result for the correct class of any test sign was 992. The test sequences were signed by two signers, and the training sequences were signed by another signer, who did not sign in any of the test sequences. Thus, the experiments are user-independent.

3.1.9. PETS 2002 dataset

The database [25, 26, and 27] consists of 1,000 color images of 1282 pixels of 12 hand postures performed by 19 persons against simple and complex backgrounds with varying amount of skin color. The images of three subjects signing against uniform light and dark backgrounds formed the training set, giving six training images per posture, the remaining images formed the test set were used for experimental purposes. For the images in the training set, they constructed graphs of 15 nodes. All 15 nodes were manually placed at anatomically significant points. The number of training images is quite small, but because preliminary experiments were very encouraging and, due to the amount of manual work involved in creating the model graphs, they chose not to add more images to the training set. Gabor jets of the three different feature types were extracted at the node positions of every training image.

3.1.10 The RWTH-BOSTON-104 database

The National Center for sign language and Gesture Resources of the Boston University published this database of ASL sentences [26, 27, 28 and 29]. This Database consists of 201 annotated video streams of ASL sentences. The images were captured simultaneously by using four standard stationary cameras where three of them were black/white and one was a color camera. Two of the black/white cameras were placed towards the signer's face, form a stereo pair and another camera was installed on the side of the signer. The color camera was placed between the stereo camera pair and was zoomed to capture only the face of the signer. The videos published on the Internet were at 30 frames per second, the size of the videos is 366×312 pixels, and the size of the frames without any additional recording information was 312×242 pixels.

To use the RWTH-BOSTON-104 database for ASL sentence recognition, the authors separated the video streams into a training and test set. The training set consists of 161 sign language sentences and the test set includes the 40 remaining sign language sentences. Further the training set was splited again into a smaller training set with 131 sequences and a development set with 30 sequences, in order to tune the parameters of the system used. The database is freely available at http://www-i6.informatik.rwth-aachen.de/~dreuw/ database.php.

3.2. Creating own database

Most of the researchers create their own database for sign language recognition. This database can be also classified into digits, alphabets and phrases (simple or complex). The Table-3 describes the characteristics of the dataset created by various researchers.

4. DATA ACQUISITION METHODS EMPLOYED

In creating standard database a set of digital cameras/video cameras with different positions/places before the object are used by different researchers. Also they employed different lighting illuminations, background selection and other equipments like hat/cap, dresses and spectacles to acquire data in the form of static gestures (photographs) or dynamic gestures (videos).

The same kinds of procedures are also followed by other researchers for acquiring their own dataset. Some researchers use specially designed input devices to capture gestures dataset. Although specially designed devises like the CyberGlove® [30] may be expensive but it is self sufficient to acquire desired data and no other supporting input devices are required. In the following section we will explain each of them. In general digital cameras are used by various researchers to acquire static gestures (signs).

4.1. Digital still camera

Two digital cameras [7, 31] with their optical axis parallel to the Z axis were used for data acquisition. The distance between the lenses of the cameras, called the baseline was set to 10 centimeters. The world coordinate system {X, Y, Z} was chosen parallel to the camera coordinate system (x, y, z). The origin of the coordinate system was located exactly between the two cameras.

Image for each sign was taken [2] by a camera. Portable Document Format (PDF) was used for images preprocessing due to the system will be able to deal with images that have a uniform background. Images of signs were resized to 80×64. The authors used 'bicubic' method to reduce aliasing. The default filter size used was 11×11.

In [8, 32] images of 30 letters from the Chinese manual alphabet were collected from a camera. A total of 195 images were captured for each letter, therefore 5850 images in all.

In the designated experiment [33], the authors used 26 hand gestures to express 26 letters. The hand gesture images were captured in five different views. So, five cameras were mounted at approximately 2/3 body height with looking directions that was parallel to the ground plane. There were 130 hand gestures from all the views were cropped by the minimum exterior rectangle of hand region and then resized to 80×80 pixels. All images were transformed to gray scale images and binary images, respectively.

A digital camera was used to image acquisition and a colored glove was used [34] to allow image processing using the color system (Hue, Saturation, Intensity) HSI. A total of 900 colored images were used to represent the 30 different hand gestures. Segmentation process divides the image into six layers representing the five fingertips and the wrist.

The required images were acquired using a digital camera by the authors [35]. Backgrounds of all images were kept black for uniformity. Five persons (male and females) with mean age 25 years were recruited for the study. They were familiar with the experimental procedure before the experimental data collection. By varying the hand orientation and its distance from camera, 30 images for each sign were captured.

A camera was used [36] to collect the gesture data. Because the main focus was on the adaptation module, the vocabulary and the feature extraction were simplified. The vocabulary consisting of 10 gestures, each of which is a number with 3 connected digits. The digit was signed by Chinese spelling way. The authors extracted eight features from these gestures: the area, the circumference, the length of two axes of the ellipse to fit the gesture region and their derivatives. Experimental data set consists of 1200 samples over 10 gestures and 5 subjects. Among the 5 subjects, 4 of them are selected as the training subjects, and each of them signs 5 times of each gesture.

In this work [37], 32 signs of PSL alphabet were captured to train and test the proposed system. The selected static signs were collected from one hand. The required images were obtained using a digital camera. Background of all images was black for uniformity, and the experimental data were collected by varying the hand orientation and its distance from camera. 640 images for the selected signs were prepared. Among these 416 images were utilized as the training set, and the remaining 224 images were employed as the test set.

4.2 Video camera

For the identification of the image areas with hand and face color several points belonging to the corresponding areas in the image was obtained from web camera. The proposed hand shape comparison algorithm [10] was tested on a set of 240 images taken from 12 signs. Locating the user's face and hands was solved in two stages, first, determination of the image areas with skin color and, secondly, segmentation of the obtained area for faces and hand recognition. Face segment was considered to the larger part of the image, and two smaller ones were hands. The user should wear a long sleeved garment which was differing in color from his/her skin.

The video sequence of the subject signer was obtained by using a camera. In this work, the authors [38] suggested that the camera faces towards the signer in order to capture the front view of the hand gestures. The initiation of the acquisition was carried out manually. A camera sensor was required in order to capture the image frames from the video sequence of the signer.

In the experiments [11] the subject signers were asked to wear dark clothes with long sleeves with white gloves and stand before dark curtains under normal lighting illumination. 15 different gestures from TSL were captured with the help of a video camera. Each input gestures consists of a sequence of 30 image frames captured using a single hand moving in different directions with constant or time-varying hand shape. Gestures had similar or different moving trajectories. The shape patterns and trajectory patterns to find the resemblance between the input model and the stored model were used in the recognition phase. Four different hand motion directions were available in the database. Accordingly the direction of hand's motion was used to classify a sign in a group of motion directions. Motion history image are used for finding motion direction. After this coarse classification, key frames in a sign were selected by means of Fourier descriptor [39]. Tow frames with differences corresponding to max FD coefficients were selected as key frames. Each sign had two key frames. The features were extracted from each key frame with GCD and stored in a vector. So for each sign there was a vector. Then sign language recognition was conducted according to the GCD features of hand shapes in key frames and Euclidean distance classifier.

The database [12] consists of 23 Arabic gestured words/phrases collected from 3 different signers. The list of words was the most commonly used in the communication between the deaf and the hearing society. Each signer was asked to repeat each gesture for 50 times over 3 different sessions, so a total of 150 repetitions of the 23 gestures results a grand total of 3450 video segments. The signers were videotaped using a digital camcorder without imposing any restriction on clothing or image background for user independent recognition applications. The three signer participants (one male and two females) were quite diverse in terms of size and height.

The authors [40] decided to select vision based system to continue this project as they felt data gloves were very expensive and inconvenient to the signer. Sri Lankan Tamil finger spellers' average speed of the finger spelling per minute is forty five (45) letters per minute. That implies a signer will be able to sign only 0.75 signs per second. A standard web camera is capable to capture fifteen frames per second and processing of fifteen frames in a second is computationally very expensive. So, the authors proposed a method in video capturing module to capture only three frames per second, which will allow system to speed up the recognition process.

The vision-based sign language recognition system in Intelligent Building **(**IB) [41] was able to capture images of the sign language user by a video camera. An integrated algorithm (AdaBoost) was incorporated in the system for the face and hands detection to address the problem of real-time recognition system. The system extracts features of sign language gestures, facial expressions and lip movements separately after pretreatment. These features were matched with sign language database and facial expressions database. Simultaneously, lip movements are extracted through image edge detection and matched with mouth shape database. After processing of semantics disambiguation, all of the recognition results are integrated for translating the sign language into speech.

Video sequences in [42] were captured from a CCD camera for the proposed system. Since hand images are two-fold, the 2-DHMM, an extension to the standard HMM, offers a great potential for analyzing and recognizing gesture patterns were used. Due to fully connected 2-DHMMs lead to an algorithm of exponential complexity, the connectivity of the network has been reduced in several ways, two among which are Markov random field and its variants and pseudo 2-DHMMs. The latter model, called P2-DHMMs, is a very simple and efficient 2-D model that retains all of the useful HMMs features. This paper focused on the real-time construction of hand gesture P2-DHMMs. The proposed P2-DHMMs use observation vectors that are composed of two-dimensional Discrete Cosine Transform 2-D DCT) coefficients

The gesture recognition system uses both the temporal and characteristics of the gesture for recognition. The system was also robust to background clutter, did not use special glove to be worn and runs in real time.

In the experiments which were performed in a user-independent manner [23], the database contains 933 examples of signs, corresponding to 921 unique sign classes. The persons performing signs in the query videos did not appeared in the database videos. All test images were obtained from video sequences of a native ASL signer either performing individual hand shapes in isolation or signing in ASL. The test images were obtained from original frames by extracting the sub-window corresponding to the hand region, and performing the same normalization that had performed for database images, so that the image were resized into 256×256 pixels, and the minimum enclosing circle of the hand region is centered at pixel (128, 128), and has radius 120.

In the experiments 20 different hand shapes were included. Those 20 hand shapes are all commonly used in ASL. For each hand shape, the authors synthetically generated a total of 4, 032 database images that correspond to different 3D orientations of the hand. The 3D orientation depends on the viewpoint and on the image plane rotation. A sample of 84 different viewpoints from the viewing sphere was collected, so the viewpoints were approximately spaced 22.5 degrees apart. Also a sample of 48 image plane rotations was used, so that rotations were spaced 7.5 degrees apart. A total of 80,640 images were collected for the experiment. Each image was normalized to be of size 256×256 pixels, and the hand region in the image was normalized so that the minimum enclosing circle of the hand region is centered at pixel (128, 128), and has radius 120. All database images were generated using computer graphics using Poser 5 software.

For the purpose of a typical gesture recognition system, the proposed [14] system has the following prominent characteristics:

(i) Hand gestures were recognized without resorting to any special marks, limited or uniform background, or particular illumination.
(ii) Only one un-calibrated video camera was utilized.
(iii) The user was allowed to perform sign language letters, within the view field of the camera.
(iv) The proposed system observes the user and give feedback in real-time.

4.3. Specially designed input devices

Specially designed data acquisition devices are also used by some of the researchers in order to acquire input signs. These are the list of input devices:

- CyberGlove®
- Sensor Glove
- Polhemus FASTRAK

4.3.1. CyberGlove®

The CyberGlove® [30, 43] consists of two bend sensors for fingers, four abduction sensors, and some additional sensors to measure thumb crossover, palm arch, wrist flexion and wrist abduction. The 18 sensors are based on a linear, resistive bend sensing technology, which was used in experiments to transform hand and finger configuration into real-time joint-angle data that were converted and digitized to 8 bits. These data from the 18 sensors were captured at a rate of 112 samples per second, and provide 18-D feature vectors to describe the handshape. In the experiments, the authors attached one receiver to the chest area of a signer to serve as a reference, and attached another to the wrist of the dominant signing hand, to obtain hand tracking data at 60 Hz. The hand and glove data were simultaneously acquired at a synchronized rate of 60 Hz.

4.3.2. Sensor glove

For experiments [44], the authors used the MEMS sensors ADXL202 accelerometers (www.analog.com). The ADXL202 is of low cost, low power, and complete two-axis accelerometers on a single IC chip with a measurement range of ±2g. The ADXL202 was used to measure both dynamic acceleration (e.g., vibration) and static acceleration (e.g., gravity).

The surface micromachining technology is used to fabricate the accelerometer. It is composed of a small mass suspended by springs. Capacitive sensors distributed along two orthogonal axes provide a measurement proportional to the displacement of the mass with respect to its rest position. The sensor is able to measure absolute angular position due to the mass is displaced from the center, either because of acceleration or due to an inclination with respect to the gravitational vector. The outputs produced by the Sensor Glove are digital signals whose duty cycles (ratio of pulse width to period) are proportional to the acceleration in each of the two sensitive axes. The output period can be adjusted from 0.5 to 10 ms via a single resistor R_{SET}. If a voltage output is required, a voltage output proportional to acceleration can be achievable from the X_{FILT} and Y_{FILT} pins, or may be reconstructed by filtering the duty cycle outputs. The bandwidth of the ADXL202 can be set from 0.01 Hz to 5 kHz via capacitors C_X and $C_{Y,}$ if required. The typical noise floor is 500 µg/ (Hz) ^{½} which allow signals below 5 mg to be resolved for bandwidths below 60 Hz. The sensing device used for experiments consists of six ADXL202 accelerometers attached on a glove, five on the fingers, and one on the back of the palm. The Y axis of the sensor on each finger points toward the fingertip, which provides a measure of joint flexion. The Y axis of the sensor located on the back of the palm is able to measure the flexing angle of the palm. The X axis of the sensor on the back of the palm can be used to extract information of hand roll, and the X axis of the sensor on each finger can provide information of individual finger abduction.

Data can be collected by measuring the duty cycle of a train of pulses of 1 kHz. The duty cycle is 50%, when a sensor is in its horizontal position. When it is tilted from +90º to -90º, the duty cycle varies from 37.5% (0.375 ms) to 62.5% (0.625 ms), respectively. The duty cycle is measured using a BASIC Stamp microcontroller in the device used. The Parallax BASIC Stamp module is a small, low cost general-purpose I/O computer that is programmed in a simple form of BASIC (refer www.parallax.com for details). The pulse width modulated output of the ADXL202 can be read directly of the BASIC Stamp module, so no ADC is required. Twelve pulse widths are read sequentially by the microcontroller, beginning with the X axis followed by the Y axis, thumb first. The data are then sent through the serial port to a PC for the purpose of analyses.

4.3.3. Polhemus FASTRAK

It [43, 45] provides real time 6 Degree-of-freedom (6DOF) tracking with virtually no latency. It may be used for head, hand and instrument tracking for biomedical analysis, graphics and cursor control, digitizing and pointing, steriotaxic localization, telerobotics and other applications. It captures data with high accuracy with maximum reliability and is used for electromagnetic motion tracking system.

It tracks the position (X, Y, and Z coordinates) and orientation (azimuth, elevation, and roll) of a small sensor as it moves through space. The tracking system's near zero latency makes it very ideal for virtual reality interfacing, simulators and other real time response applications. It also converts the acquired data that can be used in popular computer graphics programs. By just pressing buttons data are captured in a simpler manner. It provides exceptional stability in power grid fluctuations. The system setup is very easy as it can be connected to a PCs USB/RS-232 port.

5. DATA TRANSFORMATION

There are several reference points [7] which can be used for image analysis. In sign language recognition where the motion of the hand and its location in consecutive frames is a key feature in the classification of different signs, a fixed reference point must be chosen. The hand's contour was chosen to obtain information on the shape of the hand and also used the hand's center of gravity (COG) as the reference point which alleviated the bias and applied as other reference points. After defining the reference point, the distance between all the different points of a contour respect to the COG of the hand were estimated. The location of the tip of the hand was easily extracted by extracting the local maximum of the distance vector. To reduce the noise introduced by the quantization of the image and the contour extracting methods, a moving average filter to smooth the distance vector was used in the experiments.

The RGB color space [8, 37] (Red, Green and Blue) was converted to gray scale image and then to a binary image. Binary images are images whose pixels have only two possible intensity values. They are normally displayed as black and white. Numerically, the two values are often 0 for black, and either 1 or 255 for white. Binary images can be produced by thresholding (0.25 in case of [37]) a grayscale or color image, in order to separate an object in the image from the background. The color of the object (usually white) is referred to as the foreground color. The rest (usually black) is referred to as the background color. However, depending on the image which is to be thresholded, this polarity might be inverted in which case the object is displayed with zero and the background is with a non-zero value. This conversion resulted in sharp and clear details for the image.

These actions proposed [10] were (1) the downscaling of initial video, (2) skin color area detection using neural network classifier and the inverse Phong reflection model, (3) 10×10 rectangle pixel averaging, which rejects minor objects with color of the user's skin, (4) skin area clustering and label assignment (the larger cluster was a face, two smaller clusters were hands) and (5) hand motion trajectory refinement by means of the closest neighbor method.

Images were captured [31] from a parallel binocular system the hand was extracted by a color-based segmentation process and its contour was used in order to model the hand's geometry. After the contour was extracted from the image, the extracted contour was represented by Elliptical Fourier coefficients and a limited set of harmonies.

The conversion of video sequences into frame format (any size of frame format) was the first step in [13]. Backgrounds in the used video sequences were uniform and non uniform. In the proposed hand gesture recognition system the first step was video object plane (VOP) generation. Inter-frame change detection algorithm was used for extracting the VOP using contour mapping.

Various factors including lighting illumination, background, camera parameters, and viewpoint or camera location were used to address the scene complexity in the research [38]. These scene conditions affect images of the same object dramatically. The first step of preprocessing block was filtering. A moving average or median filter was used to remove the unwanted noise from the image scenes. Background subtraction forms the next major step in the preprocessing block. Running Gaussian average method [46] is used in order to obtain the background subtraction as it is very fast and consumes low memory when compared to other similar methods.

The hand gesture image sequence was analyzed for key frame selection after global motion analysis. As the hand shapes between two consecutive view models were very similar to each other, the authors [46] select some key frames for the stored model generation and the input model generation. The closed boundary of segmented hand shape was described by a Fourier Descriptor (FD) vector with the first 25 coefficients. Due to the properties of rotation, translation, dilation invariant the database space of the stored models was reduced.

The video sequences of a given gesture were segmented in the RGB color space prior to feature extraction [12]. This step had the advantage of colored gloves worn by the signers. Samples of pixel vectors representatives of the glove's color were used to estimate the mean and covariance matrix of the color which was segmented. So the segmentation process was automated with no user intervention. The measure of pixel similarities was used by the Mahalanobis distance. A pixel vector that falls within the locus of points that describe the 3D ellipsoid was classified as a glove pixel. The threshold used to define the locus of points was set to the maximum standard deviation of the three color components. Once the images were segmented, a 5×5 median filter was used to counter affect any imperfections as a result of the segmentation process.

In the proposed work [35] color images were first resized to 250×250 pixels and then, the RGB (Red, Green and Blue) images were converted to gray scale images. Users were not required to use any gloves or visual markings; instead the system uses only the images of the bare hand taken by a digital camera.

In color object tracking method [47] the video frames were converted into color HSV (Hue-Saturation-Value) space. Then the pixels with the tracked color were identified and marked and the resultant images were converted to a binary (Gray Scale image). In image preprocessing, all the images were cropped and their eye-points were manually aligned.

Then all the image vectors were normalized to unity [17].

The system [42] identifies image regions corresponding to human skin by binarizing the input image with a proper threshold value. Then small regions from the binarized image were removed by applying a morphological operator and select the regions to obtain an image as candidate of hand.

At the first step [25, 48] in the image processing phase a hand region extraction was performed. The experiments have been done in front of a simple background and in constant lightning conditions. Three well-known models, namely: normalized RGB, Gaussian distribution model of a skin color and morphological image processing have been used for this purpose.

6. FEATURE EXTRACTION

Refer to Table-4 for details.

7. CLASSIFICATION

Various classification techniques which are used by researchers to recognize sign language gestures are summarized in the Table-5.

8. RESULTS

The results obtained by various research papers are summarized in Table-6. The Table-6 (b) shows the results obtained from standard datasets that are available for research work, which we described in section 3.1. Similarly the result from creators own datasets are summarized in Table-6(c). The result includes the parameters like input Sign Language, Dataset size, Training set, Testing set, standard dataset/ creators of own dataset, classification methods and finally recognition rate[49, 50].

The table indicates that neural network and HMM variations [51] are widely used by the researchers due to their popularity in terms of recognition percentage.

9. CONCLUSIONS

After thorough analysis, the following are conclusions for future research in sign language recognition:

- Current systems are mainly focused on static signs/ manual signs/ alphabets/ numerals.
- Standard dataset not available for all countries/sub continents / languages.
- A need for large vocabulary database is the demand for current scenario.
- Focus should be on continuous or dynamic signs and nonverbal type of communication.
- Sign language recognition systems should adopt data acquisition in any situation (not restricted to laboratory data).
- Systems should be able to distinguish face, hand (right/left) and other parts of body simultaneously.
- Systems should perform recognition task in a convenient and faster manner.

Table-3. Database formed by several researchers for sign language recognition.

SL	Description	Example set	Ref.
ASL	ALS alphabets, single digit numbers used in ASL and a sample of words using bare hands.	e.g. A, B, D… e.g. 3, 5, 7 e.g. love, meet, more.	[2]
CSL	The Chinese manual alphabet, 30 hand gestures, each of them instead of a Pinyin letter.	A-Z, ZH, CH, SH, NG	[8, 32]
Not Mentioned	The sign language of 25 sentences consists of 183 words as experimental data.	No example set mentioned	[43]
Not Mentioned	The vocabulary of the database consists of 262 signs representing words from ten word types such as nouns, verbs, adjectives etc	No example set mentioned	[53]
VSL	The Latin-based Vietnamese alphabet consists of 23 based letters.	A, B, C, D, Đ, E, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y.	[44]
ASL	26 hand gestures to express 26 letters.	A, B, C, …, Z	[33]
UKL	85 gestures of the UKL Sample images of sign handshapes.	Why? Award, What for? , How much?, etc.	[10]
ASL	Static alphabet signs and second database consist of dynamic alphabet sequences of American Sign Language.	A, B, C, …, Z e.g. A name JOHN (J-O-H-N)	[13]
ASL	The training set consists of all 26 alphabets.	A, B, C, …, Z	[38]
TSL	15 different gestures from TSL.	A coin, cigarette, flower, reluctantly, row, take, immediately, understand, hate, left, seven, moon, eight, walk, conscience.	[11]
ArSL	A database of 23 Arabic gestured words/phrases.	Friend, To Eat, Neighbor, To sleep, Guest, To Drink, Gift, To wake up, Enemy, To listen, Peace upon you, To stop talking, Welcome, To smell, Thank you, To help, Come in, Yesterday, Shame, To go, House, To come and I/me.	[12]
ArSL	900 colored images used so as to represent the 30 different hand gestures.	alif , tad, ba, zad, ta, ayn, tha, gayn, jim, fa, ha, qaf, kha, kaf, dal, lam, thal, mim, ra, nun, za, he, sin, waw, shin, la, sad, ya, dhad, t.	[34]
ASL	26 CHARACTERS	A, B, C, …, Z	[54]
SLTSL	Sri Lankan Tamil finger spelling symbols.	Twelve vowels, eighteen consonants, four Grantha consonants and one special charter 'therefore’.	[40]
PSL	Eight signs from PSL.	Good, bowl, permission, child, date, stop, sentence, electricity	[35]
ASL	ASL gestures in the Real time Sign Language Gesture Recognition System (SemiosiS)	Good Night, Good Afternoon , Good Morning, Wife, Daughter, Sister Increase, , God, Jesus, Birthday, Elementary, Yesterday.	[47]
Libras	Brazilian Signs (words).	Tree, Pineapple, Yesterday, Day before Yesterday, etc	[22]
PSL	32 selected PSL alphabets.	B, OO, …, Y, H.	[37]
ASL	30 different images of 36 gestures.	A-Z, 0-9	[42]
PJM	48 hand postures from PJM.	22 postures correspond to characters: A, B, C,… ,Z. 5 postures correspond to cardinal numbers:1, 2, 4, 5, 100. 7 postures correspond to International Sign Language postures: Bm, Cm, . . . , Xm 10 postures are modifications of PSL: Aw, Bk, Bz,., Yk 4 postures correspond to modifications of number postures: 1z, 4z, 5s, 5z.	[25]
CSL	30 letters in CSL alphabet	A, B, C, …, Z. ZH, CH, SH, NG.	[14]
ASL	Based on American Sign Language (ASL) and is expanded with words, tenses, suffixes and prefixes to give a clear and complete visual presentation of English.	Scene 1 (both fingers are bent): A, C, D, E, G, I, J, L, M, N, O, Q, S, T, X, Y, Z, SMALL-C, I-L-HAND, BENT-V, L-1 HAND; Scene 2 (the ring finger is bent while the middle finger is straight): H, K, P, R, U, V; Scene 3 (both fingers are straight): B, F, W, BENT, FLAT.	[30]

Table-4. Feature extraction methods employed by various researchers.

(a) Feature extraction using digital still camera		
Ref.	**Method**	**Description**
[7]	Contour	The distance vector was used to extract some control points in order to calculate the motion parameters.
[2]	Hidden Markov Model	Hough transformation [52] with excellent image processing and neural networks was employed.
[8, 32]	Hu moments, Gabor wavelets, Scale Invariant feature Transform	The feature extraction process includes: 16-bin color histogram; 7Hu moments; 48 dimensional Gabor wavelets and a number of interest points and their SIFT features to characterize both of the global visual features and the local visual features of images.
[31]	Elliptical Fourier representation	The use of Elliptical Fourier representation of a hand could be very useful in complex deformable situations. The proposed method has been tested on both synthetic data and real data.
[34]	Thirty features used	According to the color layers expanded from the segmented color regions. Thirty features are extracted and grouped in one vector that represents a single gesture. Each vector consists of thirty entries; fifteen entries to express the angles between the fingertips themselves and between them and the wrist, and another fifteen entries that stand for distances between fingertips; and between fingertips and the wrist. Then there was a need to use a cluster method to determine the best region to express the color area. Even in the case of missing color for one reason or another, such as noise or one fingertip has been covered by another color, this case has been resolved by choosing the default position in the middle of the image.
[35]	Haar wavelets transform	The DWT is applied on the images of the selected PSL word and some features from the wavelet coefficients are extracted.
[36]	Eight features	The authors extract eight features from these gestures: the area, the circumference, the length of two axes of the ellipse to fit the gesture region and their derivatives.
[37]	2D DWT HAAR wavelet	Two-dimensional (2D) DWT was implemented using digital filters and down samplers. Convolution of low and high pass filters were applied to the original data, and the image can be decomposed in specific sets of coefficients at each level of decomposition. For the proposed system, Haar wavelet transform was selected to derive the features from each image because of its simplicity and effectiveness at describing soft parts of human body. In order to extract the approximations and details for each image, nine levels decomposition were employed.

Table-4(b). Feature extraction using specially designed devices.

[30]	Fourier Analysis	The authors used a Fourier analysis approach for periodicity detection, and the VQPCA clustering method for trajectory recognition. VQPCA is a hybrid method of clustering and local PCA which makes no assumptions regarding the underlying data distribution, and finds locally linear representation of the data.

Table-4(c). Feature extraction using video camera.

[9]	Contour mapping	Centroids, Finite State Machine, Canny Edge Detection were used for feature extraction.
[13]	Contour Mapping VOP	Video Object Plane (VOP) generation was used to extract features from video frames. Inter-frame change detection algorithm was used for extracting the VOP using contour mapping.
[38]	Points of Interest, Fourier Descriptor.	Extracted features were divided as two sub-categories, first, hand shape and, second, hand movement. The hand gesture attributes called Point of Interest (POI) of the hands were used. The feature vector consists of 55 features.
[46]	PCA-Hand features without distortion.	The proposed Visual Speaker Alignment (VSA) method strongly improves the results for the appearance-based frame features. The usage of additional training data by Virtual Training Samples (VTS) does not always lead to improvements. But a combination of both methods leads to the best result reported for this data in the literature. Using the word confidences of the recognizer output, multiple recognition systems can easily combined by "rovering" over the system outputs. Four different systems were combined: Sliding window over PCA-Frames, Sliding window over PCA-Frames and hand trajectory (HT) features, Sliding window over PCA-Frames and hand velocity (HV) features and Appearance-based PCA-hand patches
[11]	GFD/GCD	The GFD was obtained by extracting various scaled Fourier coefficients from the 2-D Fourier transform of the polar-raster sampled image.
[12]	The accumulated prediction errors or image differences (AD)	The feature extraction schemes proposed to eliminate the temporal dimension of the video based gestures through the use of weighted and directional accumulated prediction errors. Input video was eliminated and the whole sequence onto one absolute AD image representative of the motion in the video sequence. Different sign gestures can have very similar AD images. The authors proposed the use of weighted directional accumulated image differences. The weighting refers to the manner by which the predication errors are accumulated into one AD image. Such differences can be categorized into positive and negative accumulated prediction errors (AD+ and AD- respectively).
[54]	Orientation histogram	The translational invariance features were extracted.
[46]	Blob extracting	Feature extraction from hand for creating a feature vector to classify a sign in the third stage of the system was carried out. The development of this increment applies blob extracting image processing techniques and creates a vector based on binary value of image.
[41]	Normalized Moment of Inertia (NMI)	The NMI value of image has features of anti-gradation distortion, translation invariability, size invariable and so on.
[34]	Histogram	Histogram Analyzing was used to identify the clearest image.
[25]	Graph parsing	A set of Indexed Edge Unambiguous –PSL (IE-PSL) graphs representing hand postures was treated as a formal language generated with an ETPL graph grammar. The main difference between a standard (string) grammar and a graph grammar consists in an occurrence the so-called embedding transformation in a production of the latter. The embedding transformation allows one to embed the right-hand side graph in a transformed graph during a derivational step.
[23]	Dynamic Time Warping (DTW)	The DTW distance measure feature extraction was used. For the system to identify the most similar database matches to a query video, a distance measure between sign videos was defined
[19]	Four features were used	Projection information, Number of fingers visible and the multi-frame features, Embedded CSS functions, Embedded edge orientation histograms
[14]	Local Linear Embedding	Locally linear embedding with PCA and Supervised LLE were used for feature extraction.
[55]	Kalman Filter	The manual component and extracted features were only from the hand motion, shape and position. The videos were first processed for hand and face detection and segmentation. Then, sign features were extracted for manual signs (hand motion, hand shape, hand position with respect to face). For hand motion analysis, the center of mass (CoM) of each hand was tracked and filtered by a Kalman filter. The features include the width, height and orientation parameters of an ellipse and seven Hu moments calculated on the binary hand image. Hand position features are the normalized horizontal and vertical distances of the hand center to the face center. The feature vector dimensionality is 32 per frame (four hand motion, 10 hand shape and two hand position features for each hand).

Table-5. Classification methods used by researchers.

Method	Description	No. of Papers	References
Neural Network	A verity of neural network classifiers are used	6	[2, 38, 34, 40, 35, 37]
SVM	Support Vector Machine classification	1	[8]
HMM	HMM classifier with its variants	7	[9, 43, 53, 10, 36, 42, 55]
Fuzzy sets	Fuzzy Sets with other classifiers used.	1	[44]
Tensor analysis	Tensor based classification proposed.	2	[33, 17]
FSM and DTW	Finite State Machine and Dynamic Time Wrapping Algorithms.	1	[13]
ROVER	Recogniser output voting error reduction	1	[46]
Euclidean distance classifier with GCD	GCD features of hand shapes in key frames and Euclidean distance classifier.	1	[11]
I, II order Polynomial Classifier	User-independent classification of proposed solution using KNN, 1st order polynomial classifier and 2nd order polynomial classifier.	1	[12]
CAMSHIFT Algorithm	Continuous Adaptive Mean SHIFT Algorithm	1	[54]
HSBN	Handshapes Bayesian Network	1	[18]
BoostMap	A binary classifier and boosting method Ada Boost is employed for embedding optimization. This approach was adopted in Boost Map method.	1	[23]
SVR	Support Vector Regression technique	1	[19]
VQPCA	Vector Quantization Principal Component Analysis	1	[30]

Table-6. Result comparisons.

Results from standard dataset								
Ref.	**Sign lang.**	**Dataset used**			**Standard dataset used**	**Classification methods**	**Recognition rate**	
		Data size	**Training set**	**Test set**				
[9]	ASL	26 130	? ?	? ?	Lifeprint Fingure-spell Library	Dynamic Time Wrapping Static gestures Dynamic gestures	Feature without 85.77 82.97	 with 92.82 87.64
[13]	Same as [7]					Same as [7]		
[17]	Face data	150	Three lighting condition used	1040	CAS-PEAL Database	Recognition Method PCA MPCA-ML MPCA-LV MPCA-JS MPCA-PS PCA+LDA PCA+LPP	 74.79 74.04(83.08) 75.10 92.79 87.79 89.89 88.94	
[17]	Face data	38	3040	9120	Extended Yale B	Recognition Method PCA MPCA-ML MPCA-LV MPCA-JS MPCA-PS PCA+LDA PCA+LPP	 67.48 75.95 (79.77) 49.89 84.20 85.03 85.03 84.30	
[37]	Face data	28	140	140	Weizmann face database	Recognition Method PCA MPCA-ML MPCA-LV MPCA-JS MPCA-PS PCA+LDA PCA+LPP	 86.71 84.52 (86.31) 69.64 77.18 98.21 97.82 91.67	
[18]	ASL	419	?	?	Lexicon Video Dataset using linguistic annotations from SignStream	Handshape Bayesian network(HSBN) Ranked handshapes 1 5 10 15 20 25	Without HSBN 32.1 61.3 75.1 81.0 85.9 89.6	HSBN 26.0 55.1 71.4 80.2 84.5 88.7
[55]	ASL				eNTERFACE			
[23]	ASL	193	?	710	Lexicon Video Dataset	BoostMap embedding	Same (33.1) as brute-force approach but 800 times faster.	
[25]	PSL	48 48	240 144	? ?	Own Database PETS	ETPL(k) graph parsing model	94.31 85.40	
[46]	ASL	201	161	40	RWTH-BOSTON-104	Recogniser Output Voting Error Reduction (ROVER)	12.9 Word error Rate (WER)	

Table-6(b). Results from researchers own dataset.

Ref.	Sign lang.	Dataset used			Classification methods	Recognition rate	
		Size	Training	Test			
[2]	ASL	20	200	100	**ANN(feed-forward BPN)**		
					Without Canny Threshold	77.72	
					With Canny Threshold(0.15)	91.53	
					With Canny Threshold(0.25)	92.33	
[8, 32]	CSL				**SVMs classifier**	95.0256	
[43]	**?**	183	75%	25%	**Hidden Markov Model**		
					Hand Position (0.0) and no movement	49.3	
					Hand Position (1.0) and no movement	70.2	
					Hand position (0.5) and movement (0.5)	70.6	
					Hand Position (0.2) and movement(0.8)	75.6	
[53]	SLN	262	43	43	**Hidden Markov Model**	Test1	Test 2
					Training 1	98.8	91.1
					Training 2	86.6	95.8
					Training 3	98.3	100
			150	150	Training 1	93.7	64.4
					Training 2	58.5	90.7
					Training 3	93.2	92.5
			262	262	Training 1	91.1	56.2
					Training 2	47.6	93.0
					Training 3	89.8	94.3
[44]	VSL	23	?	?	**Fuzzy rule based system**		
					All letters except ‘R’, ‘U’ and ‘V’	100	
					For letters ‘R’, ‘U’ and ‘V’		
					With two-axis MEMS accelerator(ambiguity)	90, 79, and 93	
					After applying Vietnamese spelling rules	94, 90, and 96	
[33]		26	80%	20%	**Viewpoint**	Gray	Binary
					View 1	76.9	69.2
					View 2	73.1	80.8
					Tensor subspace analysis		
					View 3	100	92.3
					View 4	92.3	92.3
					View 5	92.3	88.5
					Mean	86.9	84.6
[10]	UKL			240	**Hidden Markov Model**		
					Static Signs		
		12	?		P2DIDM	94	
					Image distortion, cross-shaped surrounding area	84	
		85	?		Image distortion, square around area	74	
					Pixel-by-pixel	88	
					Dynamic Gestures	91.7	
[38]	ASL	26	26	26	**Combinational neural networks**		
					Without noise immunity	100	
					With noise immunity	48	
[11]	TSL	450	450	-	**Generic Cosine Descriptor (GCD)**	Trained	Test
					Using 3D Hopfield NN	96	91
					GCD	-	100
[12]	ArSL	3450	2300	1150	**KNN and polynomial networks**	87	
[34]	ArSL	30 Gest.	900	300	**Artificial Neural Networks**		
					Elman Network	89.66	
					Fully Recurrent Network	95.11	

Table-6(c). Results from researchers own dataset (contd.).

[40]	SLTSL	300	?	?	**Artificial Neural Networks**	
					Test Results	73.76
					Results for consonants	74.72
					Results for Vowels	71.50
[35]	PSL	8	160	80	**MultiLayer Perceptron Neural Networks**	
					Number of hidden neurons	
					10	98.75
					11	97.08
					12	97.50
[36]	CSL	10	960	240	**Hypothesis comparison guided cross validation (HC-CV)**	88.5
[37]	PSL	20	416	224	**MultiLayer Perceptron Neural Networks**	94.06
[42]	ASL	36	1080	?	**Pseudo two-dimensional hidden Markov models (P2-DHMMs)**	98
[14]	CSL	30	2475	1650	**Local linear embedding**	92.2
[30]	ASL	27	9072	3888	A linear decision tree with Fisher's linear discriminant (FLD)	96.1
					Vector Quantization PCA (VQPCA)	
					Non Periodic Signs	97.30
					Periodic Signs	97.00
					Total	86.80

REFERENCES

[1] Brill R. 1986. The Conference of Educational Administrators Serving the Deaf: A History. Washington, DC: Gallaudet University Press.

[2] Munib Q., Habeeb M., Takruri B. and Al-Malik H. A. 2007. American Sign Language (ASL) recognition based on Hough transform and neural networks. Expert Systems with Applications. 32: 24-37.

[3] Zeshan U., Vasishta M. M. and Sethna M. 2005. Implementation of Indian Sign Language in Educational Settings. Asia Pacific Disability Rehabilitation Journal. (1): 16-40.

[4] Banerji J. N. 1928. India International Reports of Schools for the Deaf. Washington City: Volta Bureau. pp. 18-19.

[5] Vasishta M., Woodward J. and Wilson K. 1978. Sign language in India: regional variation within the deaf population. Indian Journal of Applied Linguistics. 4(2): 66-74.

[6] Suryapriya A. K., Sumam S. and Idicula M. 2009. Design and Development of a Frame Based MT System for English-to-ISL. World Congress on Nature and Biologically Inspired Computing. pp. 1382-1387.

[7] Rezaei A., Vafadoost M., Rezaei S. and Shekofteh Y. 2008. A feature based Method for tracking 3-D Trajectory and the Orientation of a Signers's Hand. Proceedings of the International Conference on Computer and Communication Engineering, Kuala Lumpur, Malaysia. pp. 347-351.

[8] Quan Y. and Jinye P. 2008. Chinese Sign Language Recognition for a Vision-Based Multi-features Classifier. International Symposium on Computer Science and Computational Technology, Shanghai, China. pp. 194-197.

[9] Kshirsagar K. P. and Doye D. 2010. Object Based Key Frame Selection for Hand Gesture Recognition. Advances in Recent Technologies in Communication and Computing (ARTCom). International Conference on. pp. 181-185.

[10] Davydov M. V., Nikolski I. V. and Pasichnyk V. V. 2010. Real-time Ukrainian sign language recognition system. Intelligent Computing and Intelligent Systems (ICIS). IEEE International Conference on. 1: 875-879, 29-31 October.

[11] Pahlevanzadeh M., Vafadoost M. and Shahnazi M. 2007. Sign language recognition. Signal Processing and Its Applications. ISSPA 2007. 9th International Symposium on. pp. 1-4.

[12] Shanableh T. and Assaleh K. 2007. Arabic sign language recognition in user-independent mode. Intelligent and Advanced Systems. ICIAS 2007. International Conference on. pp. 597-600.

[13] Rokade U. S., Doye D. and Kokare M. 2009. Hand Gesture Recognition Using Object Based Key Frame Selection. Digital Image Processing. International Conference on. pp. 288-291.

[14] Xiaolong T., Bian W., Weiwei Y. and Chongqing Liu. 2005. A hand gesture recognition system based on local linear embedding. Journal of Visual Languages and amp; Computing. 16(5): 442-454.

[15] Kuang-Chih L., Ho J. and Kriegman D. J. 2005. Acquiring linear subspaces for face recognition under variable lighting. Pattern Analysis and Machine Intelligence. IEEE Transactions on. pp. 684-698.

[16] Georghiades A. S., Belhumeur P. N. and Kriegman D. J. 2001. From few to many: illumination cone models for face recognition under variable lighting and pose. Pattern Analysis and Machine Intelligence. IEEE Transactions on. 23(6): 643-660.

[17] Rana, S, Liu, W., Lazarescu, M and Venkatesh, S.. 2009. A unified tensor framework for face recognition. Pattern Recognition. 42(11): 2850-2862.

[18] Thangali A., Nash J. P., Sclaroff S. and Neidle C. 2011. Exploiting phonological constraints for handshape inference in ASL video. Computer Vision and Pattern Recognition (CVPR). IEEE Conference on. pp. 521-528.

[19] Tsechpenakis G., Metaxas D. and Neidle, C. . 2006. Learning-based dynamic coupling of discrete and continuous trackers. Computer Vision and Image Understanding. 104(2-3): 140-156.

[20] Aran O., Ari I., Benoit A., Campr P., Carrillo A. H., Fanard F. X., Akarun L., Caplier A., Rombaut M. and Sankur B. 2006. Sign language tutoring tool. In: eNTERFACE, The Summer Workshop on Multimodal Interfaces, Dubrovnik, Croatia. pp. 23-33.

[21] Benoit A. and Caplier A.. 2005. Head nods analysis: Interpretation of non verbal communication gestures. In: International Conference on Image Processing, ICIP, Genova, Italy. 3: 425-428.

[22] Antunes D. R., Guimaraes C., Garcia L. S., Oliveira L. and Fernandes S. 2011. A framework to support development of Sign Language human-computer interaction: Building tools for effective information access and inclusion of the deaf. Research Challenges in Information Science (RCIS). 5th International Conference on. pp. 1-12.

[23] Athitsos V., Wang H. and Stefan A. 2010. A database-based framework for gesture recognition. Personal and Ubiquitous Computing. pp. 511-526.

[24] Athitsos V., Neidle C., Sclaroff S., Nash J., Stefan A., Quan Yuan and Thangali A. 2008. The American Sign Language Lexicon Video Dataset. Computer Vision and Pattern Recognition Workshops. CVPRW '08. IEEE Computer Society Conference on. pp. 1-8.

[25] Flasiński M.and Myśliński S.. 2010. On the use of graph parsing for recognition of isolated hand postures of Polish Sign Language. Pattern Recognition. 43(6): 2249-2264.

[26] Triesch J. and von der Malsburg C. 2002. Classification of hand postures against complex backgrounds using elastic graph matching. Image and Vision Computing. 20(13-14): 937-943.

[27] Triesch J. and von der Malsburg C. 2001. A system for person-independent hand posture recognition against complex backgrounds. Pattern Analysis and Machine Intelligence. IEEE Transactions on. pp. 1449-1453.

[28] Philippe D., David R., Thomas D., Zahedi M. and Ney H. 2007. Speech recognition techniques for a sign language recognition system. In INTERSPEECH-2007. pp. 2513-2516.

[29] ftp://wasserstoff.informatik.rwth-aachen.de/pub/rwth-boston-104/readme.info (accessed on 10 March 2012).

[30] Kong, W. W. and Ranganath, S.. 2008. Signing Exact English (SEE): Modeling and recognition. Pattern Recognition. 41(5): 1638-1652.

[31] Rezaei A., Vafadoost M., Rezaei S. and Daliri A. 2008. 3D Pose Estimation via Elliptical Fourier Descriptors for Deformable Hand Representations. Bioinformatics and Biomedical Engineering. ICBBE. The 2nd International Conference on. pp. 1871-1875.

[32] Quan Y., Jinye P. and Yulong L. 2009. Chinese Sign Language Recognition Based on Gray-Level Co-Occurrence Matrix and Other Multi-features Fusion. Proc. of Industrial Electronics and Applications. 4th IEEE Conference. pp. 1569-1572.

[33] Wang S., Zhang D., Jia C., Zhang N., Zhou C. and Zhang L. . 2010. A Sign Language Recognition Based on Tensor. Multimedia and Information Technology (MMIT), Second International Conference on. 2: 192-195, 24-25 April.

[34] Maraqa M. and Abu-Zaiter R. 2008. Recognition of Arabic Sign Language (ArSL) using recurrent neural networks. Applications of Digital Information and Web Technologies. ICADIWT 2008. First International Conference on the. pp. 478-481.

[35] Kiani Sarkaleh A., Poorahangaryan F., Zanj B. and Karami A. 2009. A Neural Network based system for Persian sign language recognition. Signal and Image Processing Applications (ICSIPA). IEEE International Conference on. pp. 145-149.