A Project/Dissertation ETE Report

on

Human Action Recognition Using Machine Learning

Submitted in partial fulfillment of the

requirement for the award of the degree of

Bachelors of Technology
in
Computer Science and Engineering

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Dr. T. Ganesh Kumar
Associate Professor

Submitted By:

Prashant Katiyar
Enroll no: 19021180065
Adm no: 19SCSE1180072

Kumar Skand Kartik
Enroll no: 19021180056
Adm no: 19SCSE1180062

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

DECEMBER,
2022

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"Human Action Recognition"** in partial fulfillment of the requirements for the award of the **Bachelors of Technology** submitted in the **School of Computing Science and Engineering** of Galgotias University, Greater Noida, is an original work carried out during the period of **JULY-2022 to DEC - 2022**, under the supervision of **Dr. T. Ganesh Kumar (Associate Professor),** Department of Computer Science and Engineering, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Prashant Katiyar - 19SCSE1180072

Kumar Skand Kartik Shukla– 19SCSE1180062

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

(Supervisor)

Dr. T. Ganesh Kumar

Associate Professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of **Prashant Katiyar – 19SCSE1180072, Kumar Skand Kartik Shukla – 19SCSE1180062** has been held on________ _____ and his/her work is recommended for the award of **BACHELOR OF TECHNOLOGY INCOMPUTER SCIENCE AND ENGINEERING.**

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Program Chair **Signature of Dean**

Date:

Place: Greater Noida

Abstract

People with speech disabilities communicate in sign language and therefore have trouble in mingling with the able-bodied. There is a need for an interpretation system which could act as a bridge between them and those who do not know their sign language. A functional unobtrusive Indian sign language recognition system was implemented and tested on real world data. A vocabulary of 26 symbols was collected. The vocabulary consisted mostly of two-handed signs which were drawn from a wide repertoire of words of technical and daily-use origins.

Our project aims to create a computer application and train a model which when shown a real time video of hand gestures of Indian Sign Language shows the output for that particular sign in text format on the screen.

Human Action Recognition (HAR) has achieved a remarkable milestone in the field of computer vision. In this context, this survey mainly deals with the various categories of approaches that have been proposed for HAR in the last ten years. To be specific, HAR techniques range from conventional machine learning methods to recently popular deep learning methods, and this field is growing fast. Human action recognition targets recognizing different actions from a sequence of observations and different environmental conditions. A wide different application is applicable to vision-based action recognition research. However, accurate and effective vision-based recognition systems continue to be a big challenging area of research in the field of computer vision. The basic features of communication between human being characterized by human language. Many people are disabled due to hearing impairment, and they lose the ability to communicate through human language, so where deaf people communicate through sign language. One specific field of interest is sign language recognition.

Table of Contents

CHAPTER-1
Introduction

Human beings have long found it necessary to solve issues that threaten their survival or well-being. As a result, there has always been a need for them to communicate. A significant component required for successful communication is language. Language has been used for a very long time in the expression of ideas, feelings, and emotions. This can be accomplished using written symbols, gestures, or vocalizations[1]. Although the use of language for communication has helped solve problems, it often faces challenges as well. For instance, effective communication generally requires that all involved parties understand and respond to at least one common language[2]. This is not always the case in specific instances, communicating parties may rely on different written symbolic, sign, or vocal languages[5,6,9,12]. Alternatively, people may have limitations in terms of not being able to read or understand written symbols or vocals, and can be taught to communicate using these methods.

In other cases, human beings may be born with or develop a disability that may limit them from sharing certain forms of language[3]. For instance, people with hearing impairment and people who cannot physically speak due solely to certain disabilities may be limited to the use of gestures and sign language. However, it should be noted that the use of specific gestures or sign languages is not universal, and varies from one region to another and among different ethnic communities worldwide. In addition, learning multiple sign languages is complex and may not be possible for a majority of the public[20]. It is impossible for people with speech impairment to learn spoken language; this means that it is a problem for hearing-impaired people both to communicate with other people who are not conversant with sign language and to communicate among themselves.

Human beings have adopted various methods to solve this challenge[15]. For instance, human sign language translators are commonly used in public places and TV channels to communicate spoken messages in a form that people living with these disabilities can understand. However, in certain cases human sign language translators may not be available, or may not be efficient[16] and reliable. In such cases there is a need to adopt more reliable means of translating sign language into a written or spoken language.

Problem Statement

Scientists, researchers, and scholars are responsible for propelling humanity by solving problems, eliminating barriers to problem-solving, and promoting cohesion and development in society. A significant problem or barrier to problem-solving lies in ineffective communication and high communication barriers. This problem exists between speech impaired people, who can only use sign language, and other community members who cannot understand sign language. The same issue exists with respect to hearing-impaired people from different regions in the world which use foreign sign languages[7,15,25]. This communication barrier problem commonly occurs in public institutions when speech-impaired people seek services from a public service worker unfamiliar with a particular sign language. It limits speech impaired people from working in different places to communicating with sign language interpreters who are not conversant. This limits speech-impaired people from accessing or offering public services and from working in various industrial sectors. While efforts have been made to solve this challenge[19], for example, the use of human translators, they are not very efficient[21]. The advancement of technology has made it possible to use more advanced systems such as machine learning. This includes the use of special algorithms that facilitate sign language translation into written text that is universally understandable. However, research remains in progress, and it has not been established which model can best solve the problem[5,13]. Therefore, there is a need to develop and compare the performance of different machine learning models in the recognition and translation of sign language and human actions for effective communication. This is important because it can enhance the effectiveness of communication between speech-impaired people, regular community members[23], and those who use different sign languages. It can enable speech-impaired people to enjoy equal opportunities for work in public institutions and various industrial sectors with other people without this disability[25].

Tools and Technologies Used:

- Python 3.8.2
- Tensorflow 1.11.0
- OpenCV 3.4.3.18
- NumPy 1.15.3
- Matplotlib 3.0.0
- Mediapipe
- Keras 2.2.1
- VS Code/Pycharm

CHAPTER-2
Literature Survey

A Review- Deaf Mute Communication Interpreter [1]:

The purpose of this essay is to discuss the various currently used deaf-mute communication interpreter systems. Wearable communication devices and online learning systems are the two primary categories of communication approaches employed by the deaf-mute.[1] There are three types of wearable communication systems: glove-based, keypad-based, and Handicom touch-screen. Each of the three approaches discussed above uses a combination of sensors, an accelerometer, a suitable microcontroller, a text-to-speech module, a keypad, and a touch-screen.[3] The second option, an online learning system, can replace the requirement for an external device to translate messages between a deaf-mute and non-deaf-mute persons. The Online Learning System makes use of a number of methods. The five segmented approaches are the SLIM module, TESSA, Wi-See Technology, SWI PELE System, and WebSign Technology.[1]

Hand Gesture Recognition Using PCA in [3]:

In this paper, the authors present a method for database-driven hand gesture recognition based on a thresholding approach, a skin colour model approach, and an efficient template matching approach.[2] This method can be used successfully for human robotics applications and related applications. The segmentation of the hand region begins with the use of the YCbCr colour space skin colour model. Thresholding is used in the following stage to distinguish between foreground and background. Principal Component Analysis (PCA) is then used to construct a template-based matching approach.[26]

The Dumb People's Hand Gesture Recognition System[4]:

Authors described their digital image processing-based system for recognizing static hand gestures.[4] SIFT technique is used to the feature vector for hand gestures. The edges where the SIFT features were generated were invariant to scaling, rotation, and noise addition.[21]

Hand Gesture Recognition for Sign Language Recognition: A Review in [6]:

Authors described several methods of hand gesture and sign language recognition put out in the past by many scholars in their article, "Hand Gesture Recognition for Sign Language Recognition: A Review in [6]." Sign language is the only means of communication for the dumb and deaf. These physically disabled people communicate their feelings and thoughts to others by using sign language. [6]

An Automated System for Recognizing Indian Sign Language in [5]:

The approach for automatically identifying signs using shape-based features is presented in this paper. Otsu's thresholding approach, which selects an ideal threshold to reduce the within-class variance of threshold black and white pixels, is used to segment the hand region from the images. [5] Hu's invariant moments are used to partition the hand region and calculate its features, which are then fed into an artificial neural network for categorization. Accuracy, Sensitivity, and Specificity are used to assess the system's performance.[9]

A Review of Indian and American Feature Extraction Sign language in [9]:

The paper discussed current findings and creation of sign language based on hand-to-hand interaction and nonverbal cues.[17] Typically, a sign language recognition system comprehensive three-step feature extraction, preprocessing, and classification. Recognization classification techniques include Support Vector Machine (SVM), neural network (SVM), Scale Invariant Feature, Hidden Markov Models (HMM) (SIFT) transform, etc.[17]

Sign Pro: A Deaf and Blind Application Suite. in [10]:

The author described a programme that aids the dumb and deaf. a person who uses signs to communicate with the rest of the world language.[10] Real-time gesture recognition is this system's fundamental component. convert from text. The steps in processing include: gesture extraction, gesture matching, and speech to text conversion. Gesture Use of several image processing algorithms is required for extraction. such as the International Journal of (online) ISSN 2454-2024, Science & Technical Research,[20] page www.ijtrs.com 433 IJTRS-V2-I7-005 Volume 2 Issue VII, www.ijtrs.org Histogram for August 2017 @2017, IJTRS All Rights Reserved matching, computing the bounding box, and segmenting skin tones and the area expanding.[22] Techniques Useful in Gesture correlation-based and feature point matching are examples of matching.[18] Vocalization is one of the application's additional features from text and the translation of text into gestures.[23]

An Effective Wavelet Transform Framework for Indian Sign Language Recognition [2]:

The suggested ISLR system is regarded as a pattern recognition method that has two crucial modules: feature extraction and classification. An Efficient Framework for Indian Sign Language Recognition Using Wavelet Transform [2]: To recognise sign language, discrete wavelet transform (DWT)-based feature extraction and closest neighbour classifier are combined. The experimental findings demonstrate that the suggested hand gesture recognition system, when using a cosine distance classifier, achieves a maximum classification accuracy of 99.23%. [2]

Sign language:

The word sign language is similar to the language phrase, many of both are spread around different world territories [7]. Similar to language, sign language evolves over a long period of period sign language grammar and vocabulary, so it is considered a legitimate language. Because no perception of hearing is needed to understand sign language and no voice is needed to produce sign language, it is the common language among the deaf. Sign languages are usually constructed by using simultaneous compilation associated with hand shapes, orientations, and moves of the hands, palms, or body, along with facial expressions to fluidly explicit a speaker's thoughts.

Sign capturing methods:

The signs must be captured to provide input for the sign language recognition system. To capture images of hand gestures through a Microsoft Kinect camera which handles single-hand signs, double-hand signs, and finger-spelling [4,12,13], Microsoft Kinect sensors to capture multimodal data [6], Microsoft Kinect (RGB-D) sensor handled by the Nui Capture Analyze application [7], front cameras and mobile cameras [5,8,11], Sony video cameras [9], and Cannon 600 D camera RGB videos [10] are used. Microsoft Kinect was initially designed as a Gaming console peripheral device. The three sensors, that is, RGB, audio, and depth allow movements to be detected and user faces/speeches to be recognized. Microsoft Kinect sensors use a variety of useful computer vision applications, including gesture recognition, motion recognition, robotics, and virtual reality.

Chapter-3

Functionality/Working of Project

Appearance-based sign language recognition system

A simple block diagram of the appearance-based SLR system is shown in Fig. 1.

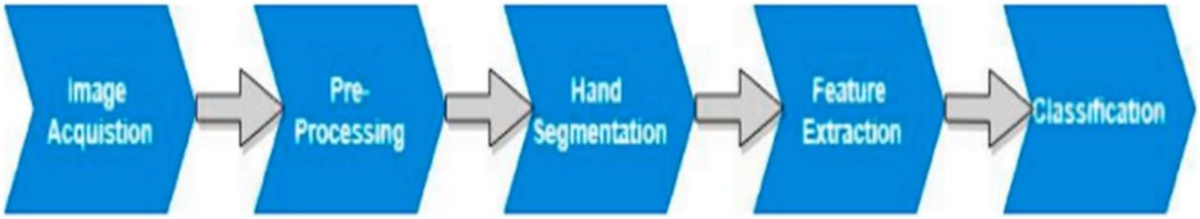

Fig. 1. Appearance-based SLR system.

- **Image acquisition**

The important element used in the sign language recognition method (SLR) as the input method is the camera. The input data for the SLR are in the form of a moving image that can easily be recorded by a camera. Nevertheless, some researchers use normal cameras to capture images [1,2,4,7,11]. Some researchers claim that they are using cameras and no gloves to reduce the challenge of using sensor-based gloves. Cameras usually support many video formats, so we need to define the default format and the format that we want to utilize by using Digitizer Configuration Format (DCF) file. Some researchers have used higher-quality cameras because the image of the web camera is blurred. A camera was used to capture 30 frames per second of real-time video and then analyzed frame by frame for dynamic gestures. The system uses a skin filter to extract the skin region and is then converted into HSV color space for each frame to an image.

There is also another device named Microsoft Kinect [14] that is used to capture images. Nowadays, because of its feature, Kinect is commonly used by researchers. Kinect can simultaneously have colour video streams and depth video streams. Background segmentation can be easily done with depth data and can be accomplished with Kinect by using signal language recognition.

Many researchers have used predefined datasets from the American Sign Language Image Dataset (ASLID) ,ASL Gesture Dataset 2012, ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) – 2010 , ChaLearn Looking at People 2014 (CLAP14) , RWTH-Phoenix-Weather 12 , RWTH-Phoenix-Weather Multisigner 2014 , SIGNUM and ArSL databases , ASLU [9], Myo Armband, and RWTH-BOSTON-50 . Few researchers create their own datas for their training of data. Because of the lack of availability of sign language datasets in particular region languages. Researchers record the data from the signer to create a dataset. ASL signs represent letters of the English alphabet [1] shown in Fig. 2 and ISL two-hand signs are shown in Fig. 3.

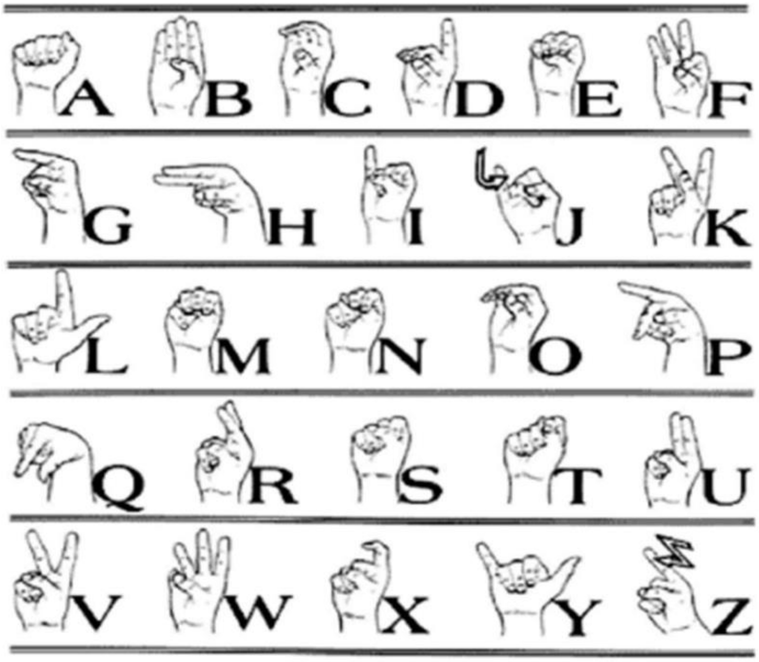

Fig. 2. ASL one hand signs.

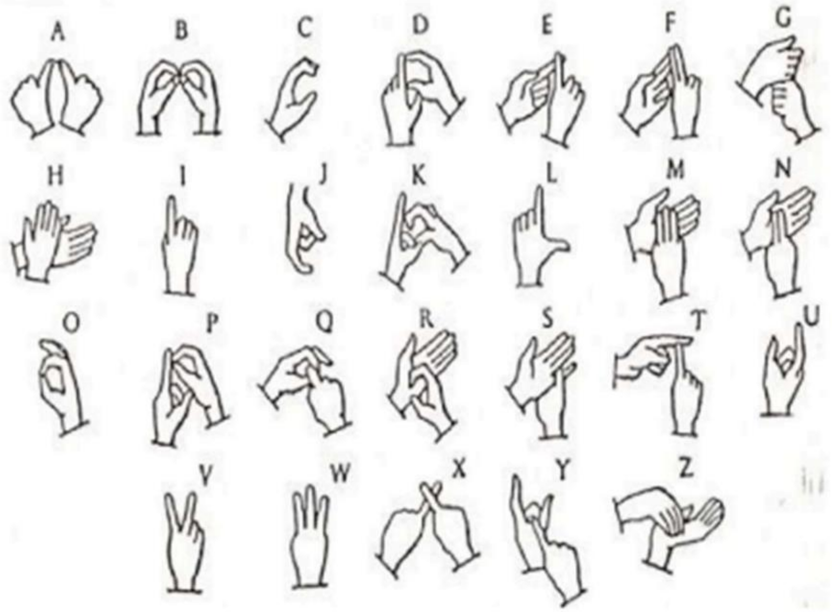

Fig. 3. ISL Two hand signs.

- **Preprocessing and segmentation:**

The image pre-processing step improves the system input to modify the image and videos. Median and Gaussian filters are some of the most frequently used methods to reduce noise in input images or videos. In research [7,9] median filtering is exclusively used for the image pre-processing stage and morphological operations [2] are also broadly used to remove unwanted information from the input. For instance, Badhe et al. [1] and Krishnaveni et al. [11] threshold the input image into binary and then K-means clustering with morphological operations in the pre-processing stage to remove noise. An adaptive histogram [6] is used to improve the contrast of input images acquired in different environments.

The segmentation approach can be contextual or non-contextual. Contextual segmentation considers the spatial connection between highlights, for example, edge recognition strategies, while a non-context-oriented division does not consider spatial relationships but bunches pixels dependent on global attributes. Skin detection also applies hand movement tracking with skin detection to produce more specific end results. Similarly to skin detection, coloured gloves are used to provide distinct characteristics to the hands, thus aiding in hand segmentation.

Skin colour segmentation is processed in the RGB model, HSV model, HIS model, and YCbCr colour models [5], while colour segmentation follows difficulties because we may face sensitivity to illumination, cameras, and skin tone. The HSV colour model is famous because the hue of the hand used differentiates palm from arm easily. In research [9], the Palms and faces of people were segmented according to HSV and YCbCr colour models. Ahmed et al. [4] used the RGB colour model to carry out hand skin colour segmentation. Match and compare with the RGB colour model used to find skin colour in a given image or video. According to research [22], the YCbCr colour model was found to be useful for colour segmentation under various light conditions.

The discrete cosine transform [6], the Viola Jones algorithm [7] and The Gaussian distribution low pass filter (grayscale) gives powerful colour segments to the face and hand in the grayscale model. In research proposed the use of K-means clustering in YCbCr colour space to isolate the frontal area from the background in an image. Badhe et al. introduced a hand tracking movement in Ref. [1] to track hands in the video. A Canny edge detector is achieved by erosion and dilation. The edge traversal algorithm segments the hand motion from the back ground in the video. Hand segmentation is a technique to segregate hands and different features from the rest of the image in vision-based systems. Rao et al. [6] employed the DCT & Viola Jones algorithm to do the frame pruning and utilized coloured gloves to aid adaptive histogram hand-head segmentation. Many hand segmentation techniques have been proposed in computer vision. The Canny edge detector is used to detect the palm edges of an image. The Canny edge detector is known for its ideal performance in identifying edges [7].

The other strategy for hand segmentation is Harris corner detection, which is used to find articulation points and motion of hands [14]. Boulares et al. [4,7] discover hand segmentation with 2D hand signature analysis using motion data matrices. Morphological operations [20] extract the elements of an image, which are helpful in the representation and description of region shape, i.e., skeletons, boundaries, and convex hulls.

- **Feature extraction**

Feature extraction is the process of extracting multiple features from an image. The features are image background, translation of image, scaling, shaping, rotation, angle, and coordinates. To extract the external boundary of objects in images, Fourier descriptors [1,7,9] are used. The sequence of coordinates forms boundaries to identify objects in an image. The Horn-Schunck optical flow algorithm [5] extracts tracking points for both arms in every frame. In Almeida et al. (2014) [14], the Speeded Up Robust Features (SURF) algorithm has been used as a feature extraction strategy in prior research. SURF is a patented descriptor for finding local features in a video.

In the Hough transform [14], the elements are arranged in into pairs (q, h) since we utilize polar directions to identify lines. It is used to find two-hand communication features for the recognition of SL. HOG [17], which is broadly utilized in the segmentation stage. Haar classifiers have been used for object recognition and used for initial real-time face detectors [9]. Local binary patterns (LBPs), find the surface and shape in grayscale images. LBP is, by all accounts, good with different facial expressions and rotation of an image. Therefore, it is reasonable for extraction in gesture-based recognition. Other feature extraction methods are tracked particle filters , 121 points used as basic descriptors , Zernike moments for keyframe extraction , and the distance algorithm [20], which are used to extract features for classification in the final step.

- **Classification**

Classification is the final stage and an essential level in the popularity of gestures. Words or sentences in sign language are made from continuous gestures, with modifications over time. Consequently, a reputation approach must be capable of handling sequential information. A few problems occur when the device handles noisy facts and uncontrolled surroundings. The method of popularity is to pick out the model from the set of fashions that could properly represent the phrase series. There are two varieties of gesture popularity processes. A few researchers have used the extracted functions for gesture recognition, which include template matching, and some have used machine learning classifiers consisting of Hidden Markov models (HMM).

- **Template matching**

An appropriate symbolic similarity measure is studied to establish matching among test and reference signs, and a simple nearest neighbor classifier is used to recognize an obscure sign as viewed as one of the recognized signs by indicating a preferred level of threshold. Euclidean Distance [1,4,14] Every gesture image in the testing dataset is compared against each gesture in the training dataset by using Euclidean distance. The gesture with a minimum distance is considered a match.

Traditional machine learning-based approaches

- **Data acquisition/image acquisition**

The important element used in signal language recognition (SLR) as the input method is the camera. The input data for the SLR is in the form of a moving image that can easily be recorded by a camera. Nevertheless, some researchers use simple cameras to capture images. Some researchers still use simple cameras [8,10,12,15] to capture images. There is also another device named Microsoft Kinect, which is used to capture images. Nowadays, Kinect is widely used by researchers because of its features. Kinect can offer colour video streams and depth video streams concurrently. With depth data, background segmentation can be carried out easily [13]. used Kinect for sign language recognition.

- **Datasets**

Most researchers create their own datasets for the training of their data. Because of the non-availability of sign language datasets in particular regions, researchers record the data from the signer to create a dataset. Researchers prepare their own sign language datasets because they usually do not have enough datasets to use for research. Numerous researchers have used predefined datasets from the American Sign Language like Image Dataset (ASLID), ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) – 2010 [3], ChaLearn Looking at People 2014 (CLAP14) [8], RWTH-Phoenix-Weather Multisigner 2014T [16,18] and The ArSL database [6], the Massey University dataset [19], and the AND VIVA challenge dataset [21].

The latest update on the pre-processing method and experiments using active sensors was conducted in Ref. [3]. They suggested a feature extraction method using the data obtained by using a Leap Motion Controller (LMC) [24]. An LMC is a device that can identify hand movements at 200 fps and gives an identity whenever it detects hand movements. The particular LMC API can directly map the detected data to hand fingertips and movements. However, LMC is not perfect and is still developing. It has some difficulties in implementing the API when hands are flipped over. The Leap Motion controller [4] is a smaller and more commercialised sensor for hand and finger motions in a 3D space of approximately eight cubits above the device. The sensor reports data such as the position and speed of the hand and fingers, primarily based on the sensor's coordinate system. Data is transferred to a computer using a USB connection.

- **Preprocessing**

The image pre-processing stage is done to modify the photo or video inputs to improve the standard overall performance of the system. Median and Gaussian filters are some of the most regularly used techniques to reduce noise in images or videos obtained. The pre-processing method for extracting features from the training image and these features are stored in neural networks to classify images in testing images.

Pre-processing methods are: Haar feature classifier , Nearest neighbor interpolation [21], Image background subtraction [15], Median filtering [14], Bandpass filter [22], Gabor filter [24], Savitzky-Golay filter [24], RGB to HSV colour space [12] and HOG [2].

Neural network model

- **CNN** [3,15,19,20,23]

A convolutional neural network (CNN) is perceived as the most important deep learning neural network model to perform with regards to recognising and classifying images. It uses multilayer superposition to extract low-level features into relevant features, resulting in a hierarchical structure similar to simulations of human brain activity [3,15,19, 20,23]. The procedure of lengthy manual feature extraction can be prevented because the recent features are passed from the past layer. CNN combines both feature learning and classification. A convolutional neural network is made up of many layers in general. In the convolution layer, a convolution operation is used to extract features from an input layer or a prior layer. The pooling layer can constantly shrink the data’s space size, reducing the number of features and computations. In the CNN, the fully connected layer serves as a “classifier.” A simple CNN diagram is represented in Fig. 4.

CNN automatically learns the values from these layers. In the context of image classification, our CNN may learn to identify edges, identify shapes, and help boundaries identify higher-level features such as face structures, respectively with the first, second, and highest layers of the network by applying convolution filters, nonlinear activation functions, pooling, and back propagation.

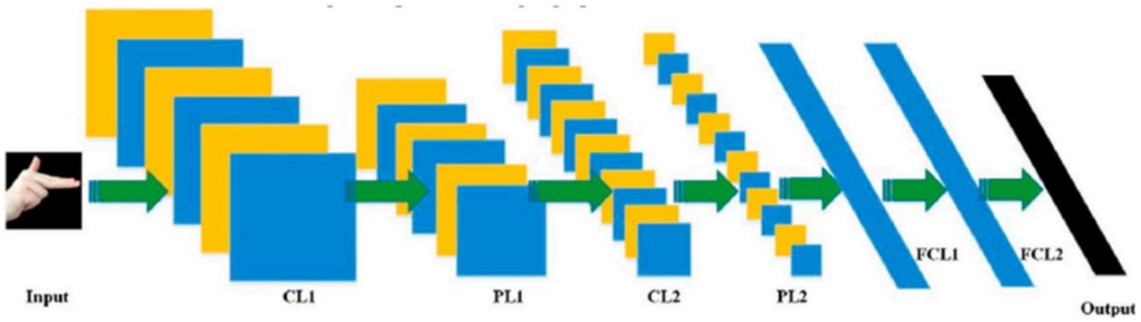

Fig. 4. Convolutional neural networks.

5

- **CNN-RNN(LSTM)** [10,12,16,18,24]

A CNN architecture can be built to gather spatial features from the video for SLR and then extract temporal features from the video by utilizing an LSTM (long-term memory) and an RNN (recurrent neural network) model. LSTM identifies gesture classes using the sequence information in SLR video.

- **Deep-CNN**

Deep convolutional neural networks (DCNNs) to learn conditional probabilities for the presence of components and their spatial relationships within image patches. Video sequences in which the temporal structure provides very helpful data where this is missing or in ways less obvious in static images. Other neural network models work similar to CNN, like faster R–CNN [8] CNN-dynamic Bayesian network (DBN) [13], stream CNN [26], and attention-based RNN [3], which are used to extract spatial features from videos and long sequences of the pose.

- **Classification**

Classification finds a function to determine the category to which the input data belongs. It can be two-category or multi-category. Factors including the classification construction method, the properties of the data to be classified, and the number of training samples all influence classification accuracy.

ReLU [23,25] are also known as “Ramp functions” due to how they look when plotted. Notice how the function is 0 for negative inputs but then linearly increases for positive values. Although the ReLU is not always saturable, it is tremendously computationally efficient. ReLU has no vanishing gradient problem compared to nonlinear functions because the gradient is constant within the nonnegative region. Learning and optimising the ReLU function is significantly easier.

Softmax function [3,8,10,12,13,15,18,21,23], the Softmax classifier finds classes based on final output probabilities in an efficient manner, and it implies multi-classification networks. In Softmax, every class is assigned decimal probabilities in the multiclass problem, and probabilities add up to 1.0. The categorical cross-entropy loss function is mostly used with the softmax activation function. The output of the model is most effective because of its logarithmic output value. The softmax activation rescales the model output so that it has the right properties. Because of this, it is common to append a softmax function because of the final layer of the neural network.

Chapter – 4

Results and Discussions

Methodologies:

It uses a vision-based methodology. The issue of using any artificial gadgets for interaction is eliminated because all of the indications are represented with just the naked hands. [13]

- ### Training and Testing:

To compare the photographs taken while utilising this technology for communication, a comprehensive database of sign language motions must be created. The procedures we used to produce our data set are listed below.[21] To create our dataset, we utilised the Open Computer Vision (OpenCV) library. First, for training purposes, we took around 800 photographs of each ASL symbol, and for testing purposes, we took about 200 images of each symbol.[25] We begin by taking a picture of each frame produced by our computer's webcam.[24] As seen in the image below, each frame has a region of interest (ROI) that is indicated by a blue-bounded square. We retrieved our ROI, which is RGB, from the entire image and converted it into a greyscale image.[15] In order to extract different aspects from our image, we then apply our gaussian blur filter to it.

- ### Gesture Classifications:

In order to forecast the user's final symbol, we used two levels of algorithms in our approach for this project. First Algorithm Layer: 1. Apply the gaussian blur filter and threshold to the opencv captured frame to obtain the processed image after feature extraction. 2. The CNN model is given this processed image for prediction, and if a letter is found in more than 50 frames, it is printed and taken into account while creating the word. 3. The blank symbol is used to indicate spaces between words. Second Algorithm Layer: We identify various symbol sets that yield similar outcomes when recognised. 2. Using classifiers designed specifically for those sets, we then categorise between those sets.

- **Challenges Faced:**

We experienced a lot of difficulties while working on the project. The dataset was the very first problem we ran into. Since working with merely square photos was much more practical, we wanted to handle raw images and that too in Keras. We opted to create our own dataset because we couldn't discover any existing ones for it. The second challenge was choosing a filter that we could use on our photos to extract the proper features, allowing us to subsequently input that image into the CNN model. We experimented with a variety of filters, such as binary threshold, canny edge detection, gaussian blur, and others, but we ultimately opted on the gaussian blur filter. More problems with the accuracy of the model we trained in previous stages were encountered, but we eventually resolved them by enlarging the input image size and also by enhancing the dataset.

Chapter – 5

Conclusion and References

Conclusion:

This article describes the development of an effective real time vision-based American sign language recognition system for Deaf and Dumb individuals using asl alphabets. On our dataset, we finally achieved a final accuracy of 62.3%. After implementing two layers of algorithms that allow us to check and predict symbols that are increasingly similar to one another, we are able to enhance our prediction. In this method, as long as the symbols are adequately shown, there is no background noise, and the illumination is sufficient, we can nearly always recognise the symbols.

References:

1] International Journal of Applied Engineering Research, Volume 11, pp. 290–296. "Deaf Mute Communication Interpreter- A Review." Sunitha K. A., Lingam Sunny, Anitha Saraswathi, P. Aarthi, and K. Jayapriya

Volume 7, Pages 1874–1883 of Circuits and Systems (2016).

[2] Mathavan Suresh Anand, Angappan Kumaresan, and Nagarajan Mohan Kumar. A Wavelet Transform-Based Efficient Framework for Indian Sign Language Recognition

[3] "Hand Gesture Recognition Using PCA," International Journal of Computer Science Engineering and Technology, Volume 5, Issue 7, July 2015, pp. 267–27. Amardeep Singh and Mandeep Kaur Ahuja.

[4] Sagar P. More and Prof. Abdul Sattar's "Hand gesture detecting system for dumb people,"

[5] S Janarthanan, T Ganesh Kumar, S Janakiraman, RK Dhanaraj, MA Shah

Journal of Sensors 2022

[6] Chandandeep Kaur and Nivit Gill, "An Automated System for Indian Sign Language Recognition," International Journal of Advanced Research in Computer Science and Software Engineering

[7] "Hand Gesture Recognition for Sign Language Recognition: A Review," International Journal of Science, Engineering, and Technology Research (IJSETR), Volume 4, Issue 3, March 2015. Vinay Jain and Pratibha Pandey.

[8] Drs. Design Issue and Proposed Implementation of Communication Aid for Deaf & Dumb People, Pankaj Agrawal, Dr. Arun Mitra, and Nakul Nagpal, International Journal on Recent and

Innovation Trends in Computing and Communication, Volume: 3 Issue: 5, pp. 147-149.

[9] Sign Language Recognition System. 2020 IRJET Vol 3. March, 2020 S. Shirbhate, Mr. Vedant D. Shinde, Ms. Mayuri A. Khandge, Ms. Sanam A. Metkari, Ms. Pooja U. Borkar.

[10] Kshitij Bantupalli and Ying Xie's master's thesis in computer science, American Sign Language Recognition Using Machine Learning and Computer Vision (2019).

[11] M. Krishnaveni and V. Radha, "Classifier fusion based on Bayes aggregation approach for Indian sign language datasets," Procedia Engineering, 30, no. 11, 1110–1118 (2012).

[12] Yang Su and Qing Zhu, Continuous Chinese sign language identification with CNNLSTM, in Proc. SPIE 10420, Ninth International Conference on Digital Image Processing (ICDIP 2017), July 21, 2017, p. 104200F.

[13] Q. Xiao, Y. Zhao, and W. Huan, "Multi-sensor data fusion for sign language recognition based on dynamic Bayesian network and convolutional neural network," Multimed. Tool. Appl. 78 (2019), 15335–15352, doi:10.1007/s11042-018–6939–8.

[14] S.G.M. Almeida, F.G. Guimar aes, and J.A. Ramrez, "Feature extraction in brazilian sign language identification based on phonological structure and employing RGB-d sensors," Expert Syst. Appl. 41 (16), 7259–7271 (2014), doi:10.1016/j. eswa.2014

[15] P. Kumar and A. Wadhawan, Deep Learning-Based Sign Language Recognition System for Static Signs, Neural Comput & Applic, 2020, doi:10.1007/s00521-019-04691-y

[16] N.C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in Computer Vision and Pattern Recognition, 2018 IEEE/CVF Conference on, Salt Lake City, UT, pp. 7784–7793.

[17] H. Lilha and D. Shivmurthy, "Analysis of pixel level features in detection of real life dual-handed sign language data set," IEEE, Recent Trends in Information Systems (ReTIS), 2011 International Conference on, December, pp. 246-251.

[18] N.C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: end-to-end hand form and continuous sign language detection," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, pp. 3075–3084.

[19] A. Kika and A. Koni, Hand gesture detection using a convolutional neural network and a histogram of oriented gradient features, in CEUR Workshop Proceedings, vol. 2280, CEUR-WS, 2018, pp. 75–79.

[20] R. Akmeliawati, M.P. Ooi, and Y.C. Kuang, Real-time translation of Malaysian sign language using neural network and colour segmentation, 2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007, Warsaw, pp. 1-6.

[21] J. Kautz, K. Kim, S. Gupta, P. Molchanov, Hand gesture identification using 3D

convolutional neural networks was published in 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, Boston, MA, pp. 1–7, doi:10.1109/CVPRW.2015.7301342.

[22] TI Manish, D Murugan, GT Kumar

Communications in information science and management engineering 3 (8), 402

[24] A Kumar, S Sagar, TG Kumar, KS Kumar

CRC Press

[25] JBS Loret, TG Kumar, A Hemathadhevi, DR Thirupurasundari

Solid State Technology 64 (2), 2181-2191

[26] TI Manish, TGK D Murugan, K Rajalakshmi

Asian Academic Research Journal of Multidisciplinary 1 (31)

[27] TG Kumar, D Murugan, K Rajalakshmi

G eofizika 32 (2), 179–189-179–189