

A Review on Hand Gesture Recognition Using Machine Learning

¹Paavan Patel, ²Dheeraj Khathuria, ³Ashish Sharma, ⁴Sachin Sharma and ⁵Dr. Nitin Choubey,

⁵Associate Dean, MPSTME, NMIMS

^{1,2,3,4,5}Department of Computer Science, MPSTME, NMIMS University, Shirpur, Maharashtra, India

Abstract—Sign language is the primary medium of communication for speech impaired people. With a significant proportion of the population being speech impaired, it becomes a necessity to devise a platform to communicate with a person who is speech impaired. As the speaking majority has limited knowledge of sign languages, we can devise a computerized platform which can understand the sign languages, thus enabling a better communication between people. This paper reviews systems and methods that can be used for the development of automatic recognition of sign language.

I. INTRODUCTION

Field of Gesture Recognition is witnessing a good quality of research and efforts these days as the set of its applications is growing at an ever-increasing pace. It has found applications not only to solve prominent problems like communication barrier between speaking majority and speech disabled but also in Virtual Reality and Gaming. Some of the interesting works have used this field to make children learn shapes using hand drawn gestures or to implement gesture navigation in computer applications. The upper layer of every related work lies in 3 main phases starting from acquisition of gesture followed by pre-processing and final classification. This paper discusses these general components and typical methodologies that are used to achieve the goal of gesture recognition.

II. PROCESS WORKFLOW

A. Gesture Acquisition

For the purpose of gesture recognition, the first and most important step is to acquire the input from the user. This input will be in a form of hand gesture representing a meaningful phrase or a single word. There are many possible ways that can perform this task and each method is associated its own set of advantages and also certain undesirable consequences. Most of the related work is centered around the following 2 broad techniques of acquisition[1].

Acquisition can be performed by capturing the image of the hand gesture and then processing this image for further feature extraction and classification. This type of acquisition is known as Image-based acquisition[22]. Clearly, this type of acquisition involves several Image Processing techniques to extract meaning out of a mere image. Also, such techniques are the most common, simplified and sophisticated mode of taking gestures from the user but also has several challenges namely different types of noise, image background, face and hands segmentation, and lighting conditions. Features extracted from images or videos are given as input and the output is simple text. In the recent projects, Microsoft marketed device called Kinect is used in place of cameras which consists of depth sensors making it an appropriate bundle for gesture acquisition[1][19]. Color images are not robust in real life since they are dependent on light conditions. Here, Kinect depth sensor supply images whose pixel value represent how far the camera is from the object[21][23][7].

Apart from the above method, Gesture can also be acquired using sensor-based techniques which usually involve a hardware in the form of data glove that consists of various types of sensors. These gloves can give information like the finger's position, it's rotation, it's movement, it's orientation and the degree of bend of the fingers[27]. A large number of features can be extracted from the data acquired from the gloves. After getting values of these attributes, the gesture can be matched using a model that has undergone supervised training and is associated with a dataset of the gestures represented using the same attributes.

B. Preprocessing & Feature Extraction

Pre-processing is an important step to improve the suppressed details present in an image and enhance image features that will be used for feature extraction. Feature extraction selects high-level features needed in order to perform classification of targets. Features uniquely describe a target in terms of location, size, composition, shape etc[13]. It improves pictorial representation for computer interpretation and processing. All in all, this step takes input from gesture acquisition phase, processes the input to make it usable by the model for classification purpose.

During our literature survey, we came across many preprocessing techniques that are not directly a part of the processing system but tend to make the input more appropriate in terms of generalization of size and quality for the processing. One of the preprocessing techniques is Background subtraction removes undesirable background area from captured frame from video stream and extracting only the desired part which is hand portion. One of the methods used for background subtraction is Gaussian Mixture Model(GMM). This is done by modelling background by classifying the pixels as object or background by calculating posterior probabilities[5].

Another technique Star Skeletonization is used for extracting feature points from the image. These features are the vectors from centroid points to the extremities of human hand in the image. The objective here is to connect these extreme points to the centroid so that a basic skeleton of the image can be obtained of the gesture which can be used for gesture matching and recognition. We also came across Boundary Extraction which can be performed by morphological dilation followed by erosion. This closing operation is done to clean up the anomalies in the targets. Using the dilation result D and the erosion result E, we can compute the boundary of the object using the image subtraction of the D and E[13].

Certain papers also used Skeleton Extraction when the boundary points are known to us, we can compute distance vectors from centroid to boundary points. This distance vector has some kind of noises which can be removed by using Fourier transform and low pass filters. Local maxima of the smoothen distance vectors are calculated and the vectors

between the centroid points and the local maxima(extreme) points are used to get a rough skeleton which will be used for gesture matching and gesture recognition. Another common technique is Histogram of Orientated Gradients (HOG) based feature extraction method which divides the entire image into areas known as “cells” making a dense grid. Histogram of edge orientations across all pixels are determined cell by cell. Final representation is combination of obtained histograms.

Few common preprocessing methods also include Blob analysis ,Noise reduction, Image scaling and Data Augmentation.

C. Classification

Classification in general can be referred as labeling of certain group of pixels on the basis of various features that uniquely divides the data in different different classes. It is divided in two phases that is training phases and testing phase. Years of research has resulted into the implementation of several methods for classification. The selection of a particular method depends upon the type and the extent of datasets.

A 2D convolutional neural network can be used to classify a particular image into a static gesture or a sign. In order to implement a machine learning model which can predict continuous gestures, we have to make some changes to the current neural network[15]. A continuous or a dynamic gesture is composed of the following two major aspects:

- i. Temporal: This includes various factors related to the present state of the hand such as its trajectory or its speed.
- ii. Spatial: This includes the orientation, size or rotation of the hand.

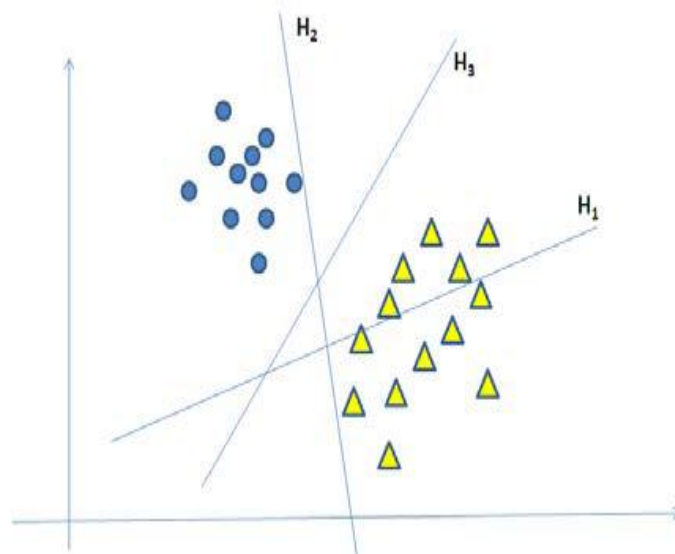
Thus to efficiently and precisely determining a particular gesture we have to take both of these aspects into consideration. Undermining or ignoring the temporal aspects will make the model prone to errors. Therefore, we can use a 3D convolutional neural network to recognize gestures. In contrast with 2D CNN that are fit for taking in the spatial data from single pictures, 3D CNNs can learn both the spatial and the temporal data from a succession of pictures, accordingly taking out the requirement for secondary temporal modelling technique.

While a CNN is particularly used for image classification, other methods such as Support Vector Machine(SVM) is used for classification, regression and outlier detection for large numeric datasets. In this algorithm, we perform classification by finding the hyper plane that differentiates the two classes after each data item is plotted as a point in n-dimensional space (i.e. n is no. of features) with the value of each feature having the value of a specific coordinate[16]. In the below figure we can see that the hyper plane H3 divides 2 different classes more accurately than H1 and H2.

The Euclidean distance between the hand boundary points and the centroid of the hand is used as a feature to train the SVM model. The features are extracted from the training images to train the model.

The dynamic gestures can be identified using models such as a Hidden Markov Model. It is a model with a random sequence of finite hidden states. A hand gesture can also be considered as an HMM with some hidden states[16][20]. In this system the hand trajectory points can be calculated using the Euclidean distances. An HMM is trained using three parameters: The initial state probability vector, a state change

probability matrix and observable probability matrix. The algorithm depends upon the initial states and the probability of the state changes that take place while execution. It is a feedforward neural network that uses back propogation algorithm for supervised training. The linear perceptron is different from a MLP in terms of the nonlinear activation function and multiple layers. Also, this kind of model is preferred when the classification data is not linearly separable.



[Source: Google]

Other methods include Artificial Neural Networks, Naïve Bayes Classification, and Decision Trees. Artificial Neural Network use the structure of our internal brain and its functioning as the base of classification process. This consists of multiple fundamental components called neurons which are associated with an activation function to trigger its input and convert to an output[14].

Naïve Bayes Classifier[22] forecasts the class of a gesture by determining the probability that the given sample belongs to a specific class .There are many Bayesian classifiers among which naïve bayes and Bayesian networks are fundamental methods. They tend to perform real-world problems[22].

In Decision trees[16], the dataset that represents the training data is structured in a tree from root to leaf, this structure makes decision on every level related to the parameter value of the data to be classified. This tree structure is formed using information theoretic called Entropy. A record traversed from the root node to the leaf node is the predicted label for that record[16].

Frameworks and libraries required for Classification:

To make classification robust and efficient various frameworks in several programming languages have been designed and are still regularly updated. Most of these frameworks are open sourced resulting in better community support, low-cost operations and platform independency. These frameworks can be run on most of the operating systems such as Windows, Linux and MacOS. Moreover their optimizations have also helped in running the compiled code on mobile devices too.

Python and R have proven to be the most effective tools used in Machine Learning and data science. Python has numerous libraries and framework for data science while R is specifically used for the same. The most commonly used libraries for machine learning are: Tensor Flow, Theano, SciKit Learn and Keras. These libraries provide tools for both supervised and

unsupervised learning as well as semi-supervised learning(eg. Sparse auto encoders). We can use these libraries for classification and regression depending on the requirement.

Sci-kit Learn provides various functionalities such as Support Vector Machines, DBSCAN, Random Forests etc. This library was created by David Cournapeau while working in Google. This library is commonly used because of its compiled code for efficiency and its simpler dependency: It only depends on two other libraries, Scipy and Numpy. Sci-Kit uses these libraries for data structures, basic arithmetic operations, linear algebra, sparse matrix representation, special functions and basic statistical functions. Sci-Kit also uses the library cython

to integrate c codes in the projects which can be used as efficient as a python code.

Tensor Flow is an open source library designed by google to provide an interface for expressing various deep learning algorithms. It is commonly used to classify data related to image processing. Various operations such as convolutional neural networks can be implemented using its in-built functions or methods. The basic unit in a Tensor Flow environment is a tensor which is a multi-dimensional matrix. Tensor Flow is popular mainly because of its ability to be deployed on large distributed systems, it can utilize numerous computational units present inside a GPU. Several algorithms used by google are Alexnet, Inception v2, and Inception v3.

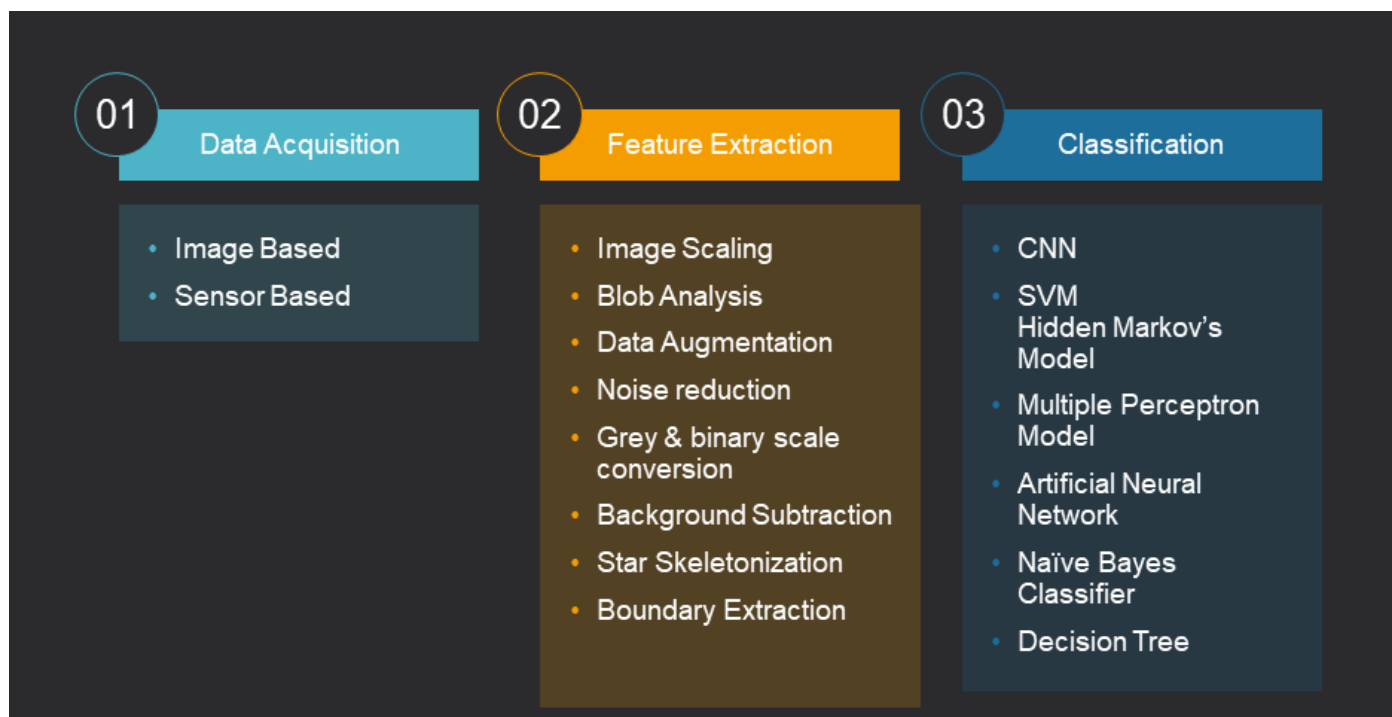


Figure 1: Hand gesture recognition general workflow

III. SUMMARY OF RESEARCH RESULTS

The following tables show summaries of some hand gesture recognition systems:

Table1. Comparison between different hand gesture recognition techniques

Paper	Acquisition	Pre-processing	Classification	Dataset	Extra remarks	Accuracy
Bharatanatyam Hand Gesture Recognition Using Normalized Chain Codes And Oriented Distances[22]	Image Based	Resizing of images, Canny edge detection, Centroid calculation, Chain codes	Naïve Bayes, KNN Classifier, Logistic Regression, Multi- Class SVM	300 images – 50 images per mudra	6 mudras named Alapadma, Araala, Chandrakala, Pataka, Tripataka and Kartarimukha	Naïve Bayes - 88.47%, KNN Classifier - 87.06%, Logistic Regression - 89.83%, Multi- Class SVM - 92.3%

Application Of Convolutional Neural Networks For Static Hand Gestures Recognition Under Different Invariant Features[17]	Image Based	Smoothing, Contrast Enhancement, Delimitation and extraction of ROI	CNN	16200 images for training, 3240 images for testing	Based on the alphabets of sign language of Peru (LSP)	CNN1 – 95.37%, CNN2 – 96.2%
Hand Gesture Recognition Based On Shape Parameters[13]	Image Based	Orientation detection, Centroid calculation, finger region detection, thumb detection	5 bit codes for finger orientation	450 images for testing – 45 different patterns	Used K means clustering for segmenting background and foreground	Approx. 94%
Hand Gesture Recognition Using Machine Learning And The Myo Armband[20]	Sensor based – 8 EMG dry sensors	NA	Pattern Matching	5 gestures x 30 repetitions = 150 Total gestures	5 gestures of the hand: pinch, fist, open, wave in, and wave out	86%
Sign Language Recognition System Using Neural Network For Digital Hardware Implementation[10]	Image Based	Edge Detection, Binarization	Neural Network	47 different pattern images	Employed Multilayer Neural Network	94%
Real-Time American Sign Language Recognition With Convolutional Neural Networks[2]	Image Based	Resizing, Horizontal flipping of dataset images	CNN - GooLeNet	ASL Finger Spelling Dataset from the University of Surrey's Centre for Vision		(a-e) - 98% (a-k) - 74%

CONCLUSION

Based on the concerned study, we observed and conclude that the process of gesture recognition is divided into following phases generally - 1) Gesture Acquisition 2) Preprocessing 3) Input Processing 4) Output. For the phase of gesture acquisition, there are 2 main methods: Image-based acquisition and sensor based acquisition. Former method was observed to be cheap, less complex but less accurate and the later was found to be expensive, complex but more accurate. For Input processing, many techniques were studied both for sensor-based and image-based. These techniques include Convolutional Neural Network(CNN), Neural Network Classification, Pattern Matching etc. Pre-requisite of this phase includes training the model with large and distinct data so that the model can be made more accurate in its predictions. This trained model will be used to test the gestures based on acquired knowledge. The last phase just delivers the output to the end-user in the appropriate form like text or audio(using TTS engine). This output is the classification output of the Input processing phase.

V. PROPOSED SOLUTION



Figure 2: Architecture of hand gesture recognition system

We propose to tackle the problem of hand gesture recognition as follows. For data acquisition, we will be using mobile camera phones to capture the hand gesture during the evaluation phase and during training phase, the hand gesture images will be taken as input. After this phase, the image

obtained will be passed through a series of image preprocessing techniques. The techniques applied in order will be color to grayscale, to binary then image rescaling followed by background subtraction. The preprocessed image will then be passed to the neural network with ReLU activation unit. From the above literature survey it is clearly seen that neural network classification is the most efficient out of all. The predicted gesture label will be taken as output. The user can be shown this gesture label in textual format or through speech.

References

- [1] Cao Dong, Ming C. Leu and Zhaozheng Yin. "American Sign Language Alphabet Recognition Using Microsoft Kinect". Missouri University of Science and Technology
- [2] Brandon Garcia. "Real-time American Sign Language Recognition with Convolutional Neural Networks". Stanford University
- [3] Luis A. Estrada Jiménez, Marco. E. Benalcázar and Nelson Sotomayor. "Gesture Recognition and Machine Learning Applied to Sign Language Translation". Escuela Politécnica Nacional/Departamento de Automatización y Control Industrial, Quito, Ecuador
- [4] Neha Poddar, Shrushti Rao, Shruti Sawant, Vrushi Somavanshi, Prof. Sumita Chandak. "Study of Sign Language Translation using Gesture Recognition". BE, IT Department, Atharva College of Engineering, Mumbai
- [5] Sawant Pramada, Deshpande Saylee, Nale Pranita, Nerkar Samiksha, Mrs. Archana S. "Intelligent Sign Language Recognition Using Image Processing". Vaidya R. H. Sapat College of Engineering, Management Studies and Research, Nashik (M.S.), INDIA
- [6] Caroline Guardino. "American Sign Language Recognition Using Leap Motion Sensor". University of North Florida
- [7] Kanchan Dabre. "Machine Learning Model for Sign Language Interpretation using Webcam Images". Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai.
- [8] Daleesha M Viswanathan, Sumam Mary Idicula. "Recent Developments in Indian Sign Language Recognition: An Analysis". Dept. Of Computer Science, Cochin University of science and Technology, Cochin-22, Kerala, India
- [9] Ellen L. Walker. "Recognizing American Sign Language Letters: A Machine Learning Experience in an Introductory AI Course". Department of Computer Science, Hiram College Hiram, OH, 44240, walkerel@hiram.edu
- [10] Lorena P. Vargas, Leiner Barba, C O Torres and L Mattos "Sign Language Recognition System using Neural Network for Digital Hardware Implementation". Lorena P. Vargas, Leiner Barba, C O Torres and L Mattos, Optic and Computer Science Group – Popular of Cesar University, Km 12, Valledupar- Colombia, E-mail: vargas.lorena@yahoo.com
- [11] A. Nandy, S. Mondal, J.S.Prasad, P. Chakraborty and G. C. Nandi. "Recognizing & Interpreting Indian Sign Language Gesture for Human Robot Interaction". IEEE Xplore Digital Library, September 2010, pp. 712-717.
- [12] Nandy, S. Mondal, J.S.Prasad, P. Chakraborty and G. C. Nandi. "Recognition of Isolated Indian Sign Language gesture in Real Time". In the book of (Information Processing and Management) Springer LNCS-CCIS, March 2010, Vol. 70, pp. 102-107
- [13] Meenakshi Panwar. "Hand Gesture Recognition based on Shape Parameters". Centre for Development of Advanced Computing, Noida, Uttar Pradesh, India,
- [14] Ms. Sonali. B. Maind and Ms. Priyanka Wankar. "Research Paper on Basic of Artificial Neural Network". Department of Information Technology Datta Meghe Institute of Engineering, Technology & Research, Sawangi (M),
- [15] Necati Cihan Camgoz. "Using Convolutional 3D Neural Networks for User-Independent Continuous Gesture Recognition". Simon Hadfield University of Surrey, Guildford, GU2 7XH, UK
- [16] Fernando Ribeiro and Luís Paulo Reis. "Vision-based Gesture Recognition System for Human-Computer Interaction". EEU-
- Escola de Engenharia da Universidade do Minho - DEI, Guimarães, Portugal
- [17] Jose L. Flores C.E., Gladys Cutipa A. and Lauro Enciso R. "Application of Convolutional Neural Networks for Static Hand Gestures Recognition Under Different Invariant Features". Ing. Informática y de Sistemas, Universidad Nacional San Antonio Abad del Cusco - UNSAAC, Peru, E-mails: {121451,efraina.cutipa,lauro.enciso}@unsaac.edu.pe
- [18] S. Mohamed Mansoor Roomi, R. Jyothi Priya and H. Jayalakshmi. "Hand Gesture Recognition for Human-Computer Interaction". Department of Electronics and Communication, Thiagarajar College of Engineering, Madurai, Tamil Nadu, India
- [19] Harsh Vardhan Verma, Eshan Aggarwal and Satish Chandra. "Gesture Recognition Using Kinect for Sign Language Translation". Dept. of Computer Sc. & Engg., Jaypee Institute of Information Technology, Noida, India
- [20] Meiping Tao and Li Ma. "A Hand Gesture Recognition Model Based on Semi-supervised Learning". Xi'an University of Posts and Telecommunications Xi'an, China
- [21] Pichao Wang, Wanqing Li, Song Liu, Zhimin Gao, Chang Tang and Philip Ogunbona. "Large-scale Isolated Gesture Recognition Using Convolutional Neural Networks". Advanced Multimedia Research Lab, University of Wollongong, Australia and School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China
- [22] R. Amrutha and Vandana M Ladwani. "Hand Gesture Recognition Using Normalized Chain Codes and Oriented Distances". CSE Department, PESIT, Bangalore South Campus, Bangalore, India.
- [23] Yue Wang and Ruoyu Yang. "Real-Time Hand Posture Recognition Based On Hand Dominant Line Using Kinect". State Key Laboratory for Novel Software Technology at Nanjing University, China.
- [24] Kanchan Dabre and Surekha Dholay. "Machine Learning Model for Sign Language Interpretation using Webcam Images". Department of Computer Engineering Sardar Patel Institute of Technology Student of M.E.(Computer) Mumbai, India.
- [25] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. "Hand Gesture Recognition with 3D Convolutional Neural Networks". NVIDIA, Santa Clara, California, USA.
- [26] Peijun Bao, Ana I. Maqueda, Carlos R. del-Blanco, and Narciso Garcia. "Tiny Hand Gesture Recognition without Localization via a Deep Convolutional Network".
- [27] M. Mohandes, M. Deriche, and J. Liu. "Image-Based and Sensor-Based Approaches to Arabic Sign Language Recognition".
- [28] Sebastien C. Wong, Adam Gatt, and Victor Stamatescu. "Understanding data augmentation for classification"