**Sign Language Recognition using Computer Vision**

Tanmay Bagdiya

Roll No.: 3370

Class.: TE-3

*Abstract-*

***Sign language is an effective tool for communication between disabled persons. It is a language that uses manual communication to convey meaning. This can include simultaneously employing hand gestures, movement, orientation of the fingers, arms or body, and facial expressions to convey a speaker's ideas. For normal people to understand sign language a translation from an interpreter is required. Computer vision can eliminate the need for an interpreter, which is helpful for communication between hearing impaired, mute persons and normal persons. This research surveys the various image recognition algorithms and technics including CNN, KNN and SVM, and applies them for hand tracking, finger tracking and gesture recognition with the help of sign language datasets. Identifying of hand pixels in each frame, extracting features and using those features to recognize a particular a hand pose is done in this research to summarise the accuracies of approach and evaluate the result.***

***Keywords: Sign Language, Image Recognition, Computer Vision, Artificial Neural Networks, Support Vector Machines***

**I. Introduction**

Sign Language is a way of communication which uses hands, fingers, arms, head or body to convey information, as opposed to acoustically conveyed sound patterns. Hand gestures are the most common way of communication between the hearing impaired and normal persons. But the contact between a hearing impaired person and a normal person has always faced obstacles such as the majority of the society is unfamiliar with sign language. A person requires an Interpreter when he/she wishes to communicate with hearing impaired, but there is a dearth of competent and experienced Interpreters. A system which can translate sign language into spoken language can greatly help the interaction between hearing impaired and normal people. An automated sign language recognition system can have significant social impacts.

Researchers have been trying to deepen the interaction between humans and computer, and gesture recognition is becoming popular in the field of Human-Computer Interaction.

Various technologies are used for sign language recognition such as glove tracking and computer vision.

Glove based methods use mechanical or optical sensor with a glove for determining the hand gestures [1]. However, this approach limits the simplicity and ease of user interaction as it requires cables and hardware such as instrumented gloves to be carries everywhere.

Vision based approaches are more user friendly and are non-invasive. They only require camera(s) to acquire the images for communication between humans and computers [2]. These approaches are difficult to design because of hurdles like conflicting background colour, low camera pixel and camera position.

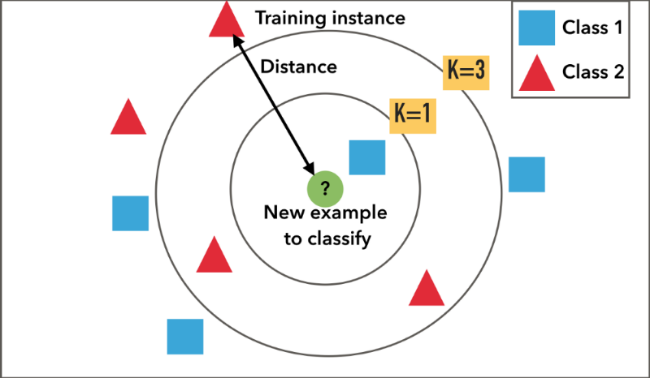
**II. MOTIVATION**

Disabled people have found themselves in the dearth of tools to communicate with people who do not know sign language. This hinders their ability to express their thoughts. With research geared towards image recognition and computer vision technologies increasing by the day, disabled people are not reaping from such advancement. Computer vision can be applied to alleviate their hindrances and boost their communication abilities.

**III. OVERVIEW**

**K-NEAREST NEIGHBOUR (K-NN)**

KNN (K-Nearest Neighbour) is a popular method used for classification and regression. It is a non-parametric lazy learning algorithm. It stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbour.



For KNN the data does not have to be separable with a linear boundary, that is, it can be applied to data from any distribution.

This algorithm is effective in case of large training data, but we need to determine the value of K (number of nearest neighbours) by comparing different accuracies for different K.

Another drawback of this algorithm is that its computation cost is high because we need to compute distance of each query with all training samples.

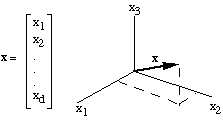
To compare images for similarity for sign language detection we need to define a distance metric. There a different choices such as Euclidean distance and Manhattan distance. In our work we used Euclidean distance to classify images.

Figure 2: The Euclidean distance.

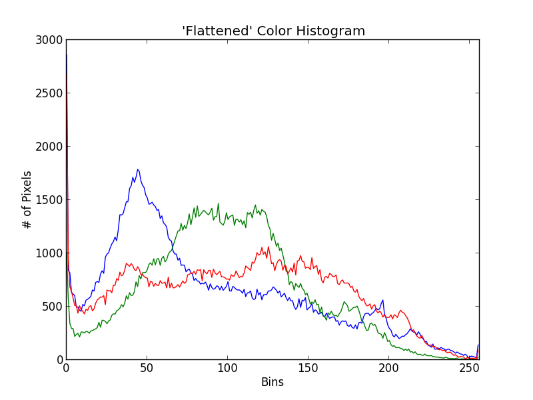
Here, d represents the distance of two points in Euclidean space.

We produced a feature vector from the given input image. In pattern recognition and machine learning, a feature vector is an n-dimensional vector of numerical features that represent some object. Feature vectors are equivalent to the vectors of explanatory variables used in statistical procedures such as linear regression.

For KNN we did not utilize raw pixel intensities as input as they can yield poor results as even small changes in attributes such as rotation, translation, viewpoint, scale can dramatically influence the image itself.



From the feature vector, we constructed a colour histogram to represent the image.  A histogram represents the distribution of colours in an image. It can be visualized as a graph (or plot) that gives a high-level intuition of the intensity (pixel value) distribution. In order to use colour histograms, we make the assumption that images with similar colour distributions are semantically similar.



An implementation of K-Nearest Neighbour on our sign language alphabet dataset shows success by yielding an accuracy of 91%.

**CONVULUTIONAL NEURAL NETWORK (CNN)**

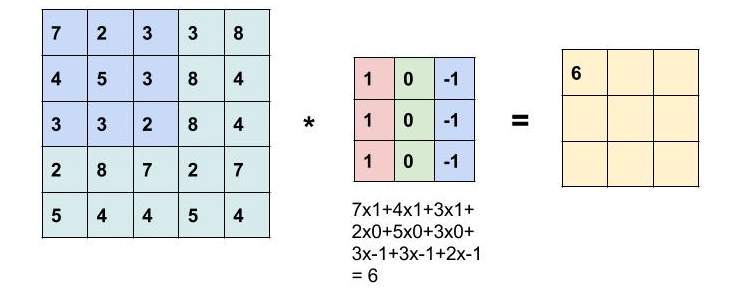
Convolutional Neural Networks have been extremely successful in image recognition and classification problems, and have been successfully implemented for human gesture recognition in recent years.

Convolutional Neural Networks (CNN), are deep neural networks used to process data that have a grid-like topology, e.g. images that can be represented as a 2-D array of pixels.

A CNN model consists of four main operations: Convolution, Non-Linearity (Relu), Pooling and Classification (Fully-connected layer).

Convolutional Layer:

The purpose of convolution is to extract features from the input image. It preserves the spatial relationship between pixels by learning image features using small squares of input data.



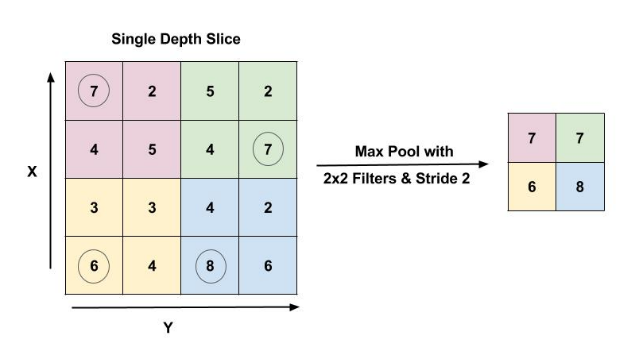
In image processing, to calculate convolution at a particular location (x,y), we extract

(k x k) sized chunk from the image centred at location (x,y). We then multiply the values in this chunk element-by-element with the convolution filter (also sized k x k) and then add them all to obtain a single output. That’s it! Note that k is termed as the kernel size.

Pooling/Sub-Sampling Layers:

Pooling (also called down sampling) reduces the dimensionality of each feature map but retains important data.

The most common form of pooling is Max pooling where we take a filter of size  and apply the maximum operation over the sized part of the image.



Relu:

It is an element-wise operation that replaces all negative pixel values in the feature map by zero. Its purpose is to introduce non-linearity in a convolution network

Fully Connected Layer:

The output from the convolutional and pooling layers represent high-level features of the input image. The purpose of the Fully Connected layer is to use these features for classifying the input image into various classes based on the training dataset.

Training Phase:

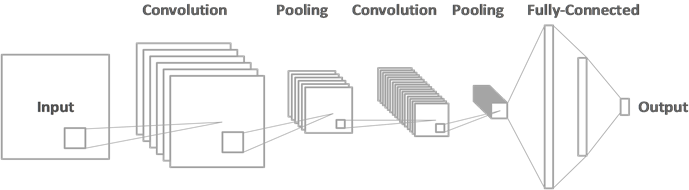
In training sequence all the features of processed image are extracted using CNN and the final feature vector is stored in database associated with that sign.

Testing Phase:

In testing phase the feature vector of the images is created and is matched with the previously maintained database.

On finding the nearest match, the image is recognized as that particular sign.

Architecture:



Most implementations surrounding this task have attempted it via transfer learning, but our network was trained from scratch. Our general architecture was a fairly common CNN architecture, consisting of multiple convolutional and dense layers. The architecture includes 2 groups of 2 convolutional layers followed by a 2x2 max-pool layer, and two groups of fully connected layer followed by a dropout layer and one final output layer

**SUPPORT VECTOR MACHINE (SVM)**

Support Vector Machine, or SVM was first proposed in 1990s by Vapnik and continues to be one of the most famous and high-performing algorithms with little tuning.

Given labeled training data, it constructs a hyperplane or set of hyperplanes (also known as decision boundary) in a high-dimensional space, categorizing the data.

It maximizes the margin from both classes while minimizing the empirical classification error. In spite of being a linear classifier, SVM can classify non-linearly separable data by using kernel trick.

Kernel functions are the functions which can map a given low dimensional input space to a very high dimensional input space by introducing features which are functions of originally provided features, in which the samples are linearly separable.

The linear decision boundary transforms to the equivalent non-linear separator when the mapping is reversed, which solves the non-linear classification problem. Regularization, Gamma, Margin are the other parameters used for tuning the SVM to get desired output.

Highly accuracy output for high dimensional feature space is the major advantage of SVM. But improper choice of kernel would limit its accuracy

Generally, vision-based hand gesture recognition involves 4 steps:

1. Data collection

2. Image segmentation

3 .Gesture modelling

4. Classification of gesture

The diagram shows the workflow in 2 different phases, training phase and testing phase.



1. Data Collection

For training an algorithmic model for supervised machine learning, the dataset is divided in two parts, training data and testing data. This is done after the data collection. The training data is used to train the model and testing data is used to test its progress, which works as feedback for the learning process.

2. Image segmentation

The image segmentation isolates the part of the image containing hand, or the area of our focus from its background which helps to build a precise model.

3. Gesture modelling

The image segment in turn is used in gesture modelling for analyzing the hand posture and gesture pattern. The outcome of gesture modelling is compared to the associated label in training phase. In testing phase, the output of the current gesture model from the previous stage is compared with each model in database.

4. Classification of gesture

The gesture recognition and classification is performed using diverse features such as number of active fingers, finger positions and directions. Algorithmic tools like SIFT and SURF can be used to detect and describe local features in images.

**IIII. COMPARISION**

After applying CNN, KNN and SVM to a sign language alphabet dataset we achieved accuracies as follows

|  |  |
| --- | --- |
| Approach | Accuracy |
| K-NN | 91.48% |
| SVM | 91.38 |
| CNN | 95.86% |

**IV. CONLUSION AND FUTURE WORK**

Sign language recognition is a challenging task. The approaches used in this work try to overcome those challenges. This paper presented a comparative study of three algorithms applied to a sign language alphabet dataset. The results achieved in identifying sign language lead us to conclude that CNN provides better performance than SVM and KNN. The results obtained also proved that the feature extraction and data preparation phase is an important one. However a few limitations exist, such as real time recognition and dealing with a varied set of inputs. Future work will be concerned with eliminating the aforementioned limitations and studying other alternatives if available.

**V. REFERENCES**

[1] K. S. Abhishek, L. C. F. Qubeley and D. Ho, "Glove-based hand gesture recognition sign language translator using capacitive touch sensor," 2016 IEEE.

[2] S. Reshna and M. Jayaraju, "Spotting and recognition of hand gesture for Indian sign language recognition system with skin segmentation and SVM," 2017 IEEE.

[3] P. Trigueiros, F. Ribeiro and L. P. Reis, "A comparison of machine learning algorithms applied to hand gesture recognition," 2012 CISTI.

[4] W. N. Khotimah, N. Suciati, Y. E. Nugyasa and R. Wijaya, "Dynamic Indonesian sign language recognition by using weighted K-Nearest Neighbor," 2017 ICTS.

[5] N. Kumar, "Sign language recognition for hearing impaired people based on hands symbols classification," 2017 CCCA

[6] Pragati Ninawe, Shikha Pandey “A Completion on Fruit Recognition System Using K-Nearest Neighbours Algorithm,” 2014 IJARCET

[7] Jayashree R. Pansare , Shravan H. Gawande , Maya Ingle “Real-Time Static Hand Gesture Recognition for American Sign Language (ASL) in Complex Background,” 2012 JSIP

[8] M. S. Sinith, S. G. Kamal, N. B., N. S., K. Surendran and J. P. S., "Sign Gesture Recognition Using Support Vector Machine," 2012 ICACC.

[9] G. R. S. Murthy & R. S. Jadon, “A review of vision based hand gestures recognition,” 2009 IJITK.