American Sign Language to Audio Conversion using CNN in Python and Website Creation.

Alphonsa Merlin Roy, Alisa Magdeline Rodrigues, Amrita Elsa Vettiyil, Caroline B.Joseph, Ms. S.Santhi Jabarani
Department of Electronics and Communication Engineering
Rajagiri School of Engineering and Technology,Cochin,India
merlinroy05071998@gmail.com, alisa.rodrigs@gmail.com, amritaelsa@gmail.com, carolinebjosph@gmail.com,
santhij@rajagiritech.edu.in

***Abstract*—American sign language to audio conversion system converts the input image of a particular alphabet to an audio output signal. This is implemented with the help of a Convolutional Neural Network(CNN). The network initially processes the reshaping of an image into its required dimensions and scale parameters. These processed images are all collected and then further sent as input to the training stage. Then the trained network is tested using external image inputs. It accurately recognizes and identifies the newly fed images by comparing with the previously trained ones stored in the database. After testing proves to be efficient, final text to speech conversion takes place and thereby receive the speech output. This system is further implemented into a website where it lays a platform for communication for external users.**

***Index Terms*—Database, CNN, Adam, Learning rate, .JSON, pyttsx3, Website**

I. Introduction

Communication via gestures is a widely used method by hearing and speech impaired people. They communicate efficiently with the assistance of the sign languages that are represented by various hand gestures. These comprises of hand shapes, movements and orientations of hands or body, or facial expressions. However it creates a gap with those who do not follow sign language and might require an external assistance. But that is not always practical. Thereby a system to recognize the sign language and automatically interpret it to reduce the gap between the two communicating groups is a necessity. Developing a technical system to support the sign language users to communicate with non-users will provide great assistance and independence to them.

Growth of the technology has greatly proven beneficial for the deaf and mute community. Through this growth they are easily able to communicate with the people around them. With the development in artificial intelligence, several neural network algorithms have been developed to translate the sign language gestures to texts as well as audio. This paper aims at developing such a system to support the the deaf and mute so as to effortlessly communicate with the hearing people.

There are numerous languages communicated in the sign language. Sign languages are not universal and they are not mutually intelligible with each other, although there are also striking similarities among sign languages. It varies from place to place in the matter of gestures, expressions, emotions and ideas. American Sign Language (ASL), Indian Sign Language(ISL), British Sign Language(BSL) are some examples of sign languages used around the globe. Among these ASL stands the most commonly used one.

There are different methods available to detect sign language. They are vision-based recognition, sensor-based recog- nition and glove-based recognition. This paper presents sign language recognition using vision-based method. The aim is to detect the sign language with good accuracy and sensitiv- ity. We have chosen a Convolutional Neural Network(CNN) system implementation as they have a smarter way of looking at the images, zooming in and out on adjacent pixels in small areas and input those readings into several different layers. Through stages of image processing , training and testing of the network it identifies the gestures fed into it thereby outputs the corresponding meaning of the gestures shown.

This paper presents the sign recognition technology to identify the 26 English alphabets and subsequently produce the sound corresponding to each letter.The Fig.1 shows the ASL alphabets.

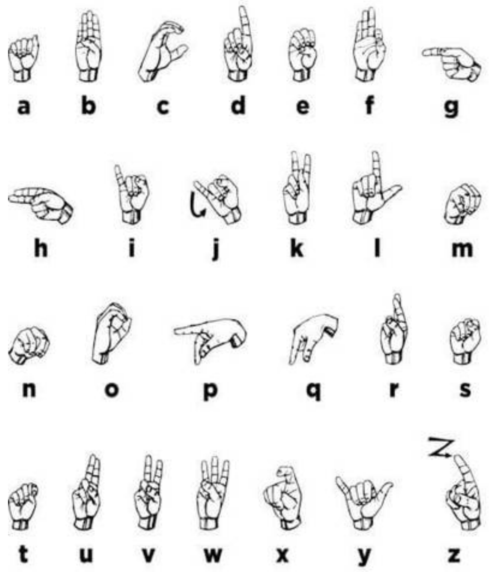


Fig. 1. Representation of ASL-alphabets

A website prototype is also created as an advancement in the paper. This is for the purpose of universal access. After launching the site onto the server any person across any part of the world can access the system to convert the gestures they show to obtain a text and its corresponding audio output.

The paper consist of sections: section II discusses about the researches conducted in sign language recognition. Details about the database is described in section III. The various steps involved in implementing the sign language to audio conversion system is discussed in detail in section IV. Section V describes the experimental results obtained. The summary and conclusion is given in section VI.

II. Literary Survey

Several gesture and image recognition techniques were developed through the years. Especially through deep learning Some of the methods are discussed below:

A method to implement feature extraction through ANN was proposed by Ariya Thongtawee, Onamon Pinsanoh and Yuttana Kitjaidure[1].They use recognition features as location,angle,velocity and motion patterns for American Sign Language alphabets for unconstrained background and accuracy upto 98 percent was achieved without using additional marker.Real time hand tracking was done by using Kalman filter and for gesture recognition pseudo-2 dimension hidden Markov model was used.

Saad Albawi , Tareq Abed Mohammed and Saad AL-Zawi[2] discusees about convolutional neural network in detail. They define all the elements and important issues related to CNN, and how these elements work. In addition, also state the parameters that effect CNN efficiency. The most important layer in CNN is convolution layer Which takes most of the time within the network. Network performance also depends on the number of levels within the network. But in the other hand as the number of levels increases the time required to train and test the network.

A gesture recognition method implemented through CNN was proposed by Rahul Chauhan, Kamal Kumar Ghanshala and R.C Joshi[3]. The paper discusses on how CNN models are built to evaluate its performance on image recognition and detection datasets. The CNN has an excellent performance in machine learning problems. Especially the applications that deal with image data,such as image classification dataset, computer vision and in natural language processing(NLP). The algorithm is implemented on MNIST and CIFAR-10 dataset and its performance are evaluated.The accuracy of models on MNIST is 99.6 %, CIFAR-10 is using real-time data augmentation and dropout on CPU unit. Optimization techniques like Adam(adaptive moment estimation) and Softmax activation function are used by the CNN to reduce the number of parameters.

An Indian sign language recognition system was proposed by Adithya V., Vinod P.R. and Usha Gopalakrishna[4]. In the paper they propose a method for the automatic recognition of finger spelling in Indian sign language. The proposed method uses digital image processing techniques and artificial neural network for recognizing different signs. The signs are identi- fied by the features extracted from the hand shapes. skin colour based segmentation for extracting the hand region from the image. A new shape feature based on the distance transform of the image is proposed in this work. The features extracted from the sign image are used to train a feed forward neural network that recognizes the sign. The method is implemented completely by utilizing digital image processing techniques so the user does not have to wear any special hardware device to get the features of the hand shape.

The method proposed by P. Subha Rajam and Dr. G. Balakrishnan[5] describes a method that provides a basis for the development of Sign Language Recognition system for one of the south Indian languages. In the proposed method, a set of 32 signs, each representing the binary 'UP' 'DOWN' positions of the five fingers is defined. The images are of the palm side of right hand and are loaded at runtime i.e. dynamic loading. The method has been developed with respect to single user both in training and testing phase. The static images have been pre-processed using feature point extraction method and are trained with 10 numbers of images for each sign. The images are converted into text by identifying the finger tip position of static images using image processing techniques. The proposed method is able to identify the images of the signer which are captured dynamically during testing phase. The results with test images are presented, which show that the proposed Sign Language Recognition System is able to recognize images with 98.125% accuracy when trained with 320 images and tested with 160 images.

III. DATABASE

Our database consists of images of all the Twenty-Six alphabets English Alphabets. Collection of these images and storing them in an organised way forms the database.The database can be updated to make the recognition process more efficient.The programming language that we have chosen is Python. For the creation of database we need to import many library files.the imported libraries are :

1) OS
2) cv2

OS

Module

It is a library used in python. Enables us to interface with the operating system that the Python is running on.

Cv2 Module

This is also a library file imported from Python which comes under OpenCV. OpenCV stands for Open Source computer vision. It mainly focuses on real-time computer vision. It is used because of the usage of camera to capture the images of alphabets.The computer vision is enabled through the camera. We have taken fifty images of each alphabet with suitable lighting and a black background. Running the database code, the images are being taken simultaneously and being added to a folder. All the alphabets are stored in a folder where the path has been defined. There is also subfolder for each alphabet where the rest of the 50 images are stored.The images are saved in a matrix form. The randomly generated output is shown in Fig.1

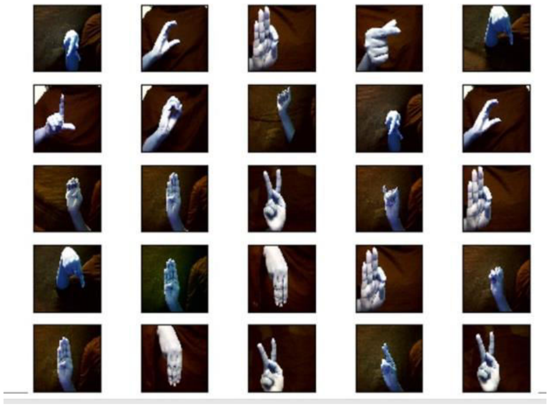


Fig. 2. Randomly generated output

IV. IMPLEMENTATION

The paper proposes the implementation of the sign language to audio conversion system through the following steps:

1)Image Processing
2)CNN Training
3)CNN Testing
4)Text to Speech Conversion
5)Website

A. *Image Processing*

The image before being loaded into the model needs to be resized and reshaped according to the required dimensions. In order to obtain the enhanced image without loosing its necessary features and to extract necessary information certain operations are to be performed on it. The processing basically includes three steps:

- To import the image via image acquisition tools
- Analyze and manipulate the image
- Output based on the results of an altered image or reports based on the analysis.

The computer cannot perceive the input image as done by humans. The image is viewed as pixels which are represented in the RGB format. The computer then feeds it in matrix form. The deep neural networks forms a link between the feed-forward network and adaptive filters which subsequently reduces the image to a form that is easier to process without losing features which are critical for getting a good prediction. The images fed into the neural network must all be of the same size thereby reshaped and resized accordingly. Images larger than the required dimensions can be scaled down through interpolation. The resized images are segregated into training and testing datasets respectively. The dimensions use pixel values. To improve efficiency and accuracy the data is normalized such that the maximum pixel value is 255.

The input 2 dimensional image is converted to 3-dimension with reduced size and converted to its corresponding proba- bilities as shown in the Fig.2

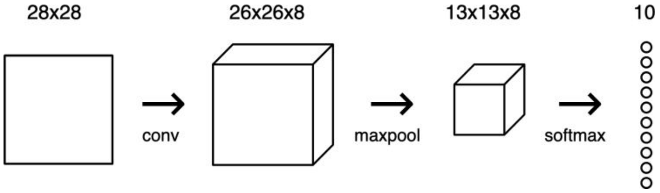


Fig. 3. 28x28 grey-scale image converted to its subsequent probabilities

B. CNN TRAINNING

In a convolutional network we need to train the network. In our project we train our network with images of alphabets which is a large process and time consuming. The main process taking place in this level is teaching our system the alphabets which is shown using hand signs. Each hand sign represents an English alphabet . Training of the model involves its passage through many layers . As it passes through these layers it achieves more efficiency.

Steps in training model

- Creation of the model
  Our model consists of the images of the alphabets.The images are read one by one and saved in the format (28,28,1). Conversion of images into an array takes place. The array is then assigned to a variable X. As the system reads the images it gets appended to another variable
  Y. The images get appended in the form of numbers corresponding to each alphabet. These read images are also converted to an array. Now the X and Y arrays are spilt for training and testing i.e. 70 images are taken for training and 30 images are taken for testing. The next step is Scaling. For this , the divided (xtrain) and (xtest) is divided by 255 so that we get in the range 0 to 1. Now in the next step we take (ytrain and ytest) in a binary format using a code.
  The images are made to run through many layers inside neural networks[6][7].The convolutional layer convolves the original image and the process is considered a fil- ter.The neurons are dropped with a dropout value of 0.2. A flatten layer is applied for the conversion of matrix to vector form.The flattened values are given to a dense layer.Activation functions are also being used here the softmax and Relu layers are used here.both the layers together are used in the input hidden and outout layers. Building of the model- In this step we use a Baseline prediction algorithm.
- Compilation Of the Model We use optimisers in order to compile our model.There are many optimisers, but the one we have chose here. is the Adam algorithm[8]. It is a Backpropagation algorithm .For the network to learn we have given the learning rate as 0.001.

- Final Model This is the process where the training takes place. The pre-processed xtrain and y train images are given into the model. Steps involved in this model are:

  1) Build the Model
  2)Fit the model
  3)Final evaluation of the model

  As a result of the above mentioned steps we get the accuracy ,the error rate.In order to reduce the error and increase the efficiency , the number of epochs must be increased and the Batch size reduced.to see the training progress of each epoch.The value of verbose given here is 2. Epoch-One epoch is a single forward and backward pass through all the images and batch-size is number of samples taken in an Epoch.Next we use a term verbose and assign a value 2 in order to see the result in the way we want. Finally we save the file in .JSON [9]and HDF5[10]formats. All the parameter settings needed to bulid the CNN model are provided in the .JSON file.HDF5 lets you store huge amounts of numerical data, and easily manipulate that data from NumPy.

C. CNN TESTING

Testing is basically checking if the model has learned the alphabets with which it was trained. Each hand sign represented an English alphabet. 30 per cent of the total images were given to the model for testing. Output of the testing part gives a number which corresponds to an alphabet. The efficiency and accuracy at the testing part is a result of how well we had trained the system.In the process the images are being read. For this we need a library file named Tkinter.Tkinter[11] is a standard Graphical user interface library used in python.It is one of the frameworks that Python provides to develop a GUI. It is open source and comes under the python license. A window gets opened and the image can be selected. The selected images is then converted grey scale, image is then resized to a dimension (28,28,1). Then we do the scaling of the image by dividing it with 255. Predictions are made and the output is shown as a label that was given to each alphabet. A dictionary d is defined where each alphabet is given a number. The output is displayed using the print () function. We import files HDF5 and .JSON in the testing stage also.The final result that we get at the output stage is a text.The conversion of image to text takes place here.

D. TEXT TO SPEECH

Text to audio conversion is done through the support of Python library called pyttsx3. Unlike other TTS modules pyttsx3 works offline and is compatible with both python 3 and 2[12]. pttysx3 is an easy tool which converts the text entered, into audio which can be saved as a mp3 file. It supports several languages, voices(male and female),accents etc. Here the audio is being played along with the text . When a text C is shown in the result, the audio for the letter is played saying that the shown alphabet is a C simultaneously.

E. WEBSITE

A prototype of the website is created using a the database we created which is a real life representation of how the systems' outcome is. This is the most advanced feature of the system.Here the users around the world can utilise the exclusive feature of sign to audio conversion , making their lives better than usual.The user can show a letter , its image is captured. The CNN network processes the image and recognizes the letter and converts it into text. The text is then converted into audio which is heard through the speakers connected to the PC.

Website has front end and back end technologies.Front End refers to the website side that the client sees and interacts with on his/her browser.It can also be referred to as the"client-side"[13].It includes everything that the user experiences directly, from text and colors to buttons, images, and navigation menus.The front end technologies include HTML5, CSS3, Javascript, JQuery, AJAX, Bootstrap.

HTML is the fundamental coding language that creates and organizes web content so it can be displayed in a web browser[14].There are various versions of the language.This system uses the fifth version of the markup language HTML ie HTML5.

Cascading Style Sheets (CSS) is a language that goes hand in hand with HTML, and it defines the layout, colors, fonts of a website's content[15].CSS is a style sheet language used for describing the presentation of a document written in a markup language like HTML. CSS is a groundbreaking technology of the World Wide Web.

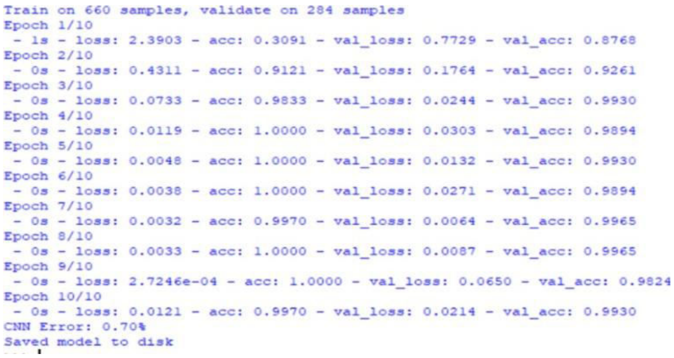
Web sites are mostly driven by a lot of JavaScript code, most of the site's user interface interaction and dynamic effects are implemented by JavaScript.[16][17] Website based on Javascript is better optimised not only to shorten the page load time, but also to improve the user in the page when the functional operation of the response speed.

Back end refers to the behind the work processes. Its the is the portion of the website that the client can't see. It can also be referred to as the "server-side". [13] It's responsible for storing and organizing data, and ensuring everything on the client-side works properly. Its the communication between the back-end and front-end, that is sending and receiving information helps put together a webpage.Back End Technologies include Python, Django

Django is a Python-based free and open-source web framework[17].Django is mainly used to ease the creation of complex, database-driven websites. The framework emphasizes spontaneity, security, scalability, versitality, less code, low coupling, and rapid development.

V. EXPERIMENTAL RESULTS

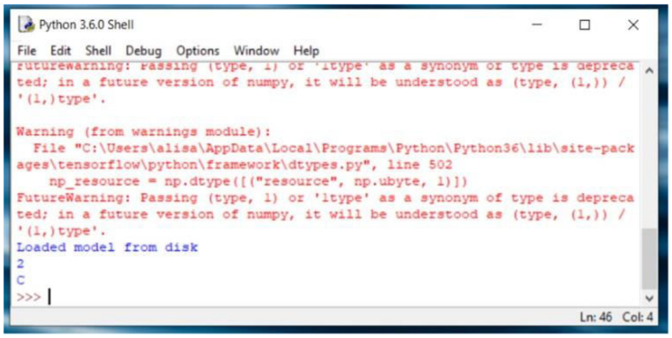
The results obtained shows how efficiently the images were trained. Ten Epochs were considered and batch-size was taken as 660.The CNN error was 0.70 percent and the accuracy obtained was 99.3 per cent,whereas the ANN accuracy was only 91 percent.The output is shown in fig 4.



```
Train on 660 samples, validate on 284 samples
Epoch 1/10
 - 1s - loss: 2.3903 - acc: 0.3091 - val_loss: 0.7729 - val_acc: 0.8768
Epoch 2/10
 - 0s - loss: 0.4311 - acc: 0.9121 - val_loss: 0.1764 - val_acc: 0.9261
Epoch 3/10
 - 0s - loss: 0.0733 - acc: 0.9833 - val_loss: 0.0244 - val_acc: 0.9930
Epoch 4/10
 - 0s - loss: 0.0119 - acc: 1.0000 - val_loss: 0.0303 - val_acc: 0.9894
Epoch 5/10
 - 0s - loss: 0.0048 - acc: 1.0000 - val_loss: 0.0132 - val_acc: 0.9930
Epoch 6/10
 - 0s - loss: 0.0038 - acc: 1.0000 - val_loss: 0.0271 - val_acc: 0.9894
Epoch 7/10
 - 0s - loss: 0.0032 - acc: 0.9970 - val_loss: 0.0064 - val_acc: 0.9965
Epoch 8/10
 - 0s - loss: 0.0033 - acc: 1.0000 - val_loss: 0.0087 - val_acc: 0.9965
Epoch 9/10
 - 0s - loss: 2.7246e-04 - acc: 1.0000 - val_loss: 0.0650 - val_acc: 0.9824
Epoch 10/10
 - 0s - loss: 0.0121 - acc: 0.9970 - val_loss: 0.0214 - val_acc: 0.9930
CNN Error: 0.70%
Saved model to disk
```

Fig. 4. CNN training result

Also at the testing stage it shows how efficiently the images are recognised.After the conversion of image to text at the testing stage,a result was obtained as shown in fig 5.



```
FutureWarning: Passing (type, 1) or '1type' as a synonym of type is depreca
ted; in a future version of numpy, it will be understood as (type, (1,)) /
'(1,)type'.

Warning (from warnings module):
  File "C:\Users\alisa\AppData\Local\Programs\Python\Python36\lib\site-pack
ages\tensorflow\python\framework\dtypes.py", line 502
    np_resource = np.dtype([("resource", np.ubyte, 1)])
FutureWarning: Passing (type, 1) or '1type' as a synonym of type is depreca
ted; in a future version of numpy, it will be understood as (type, (1,)) /
'(1,)type'.
Loaded model from disk
2
C
>>> 
```

Fig. 5. CNN testing result

VI. WEBSITE LAUNCH

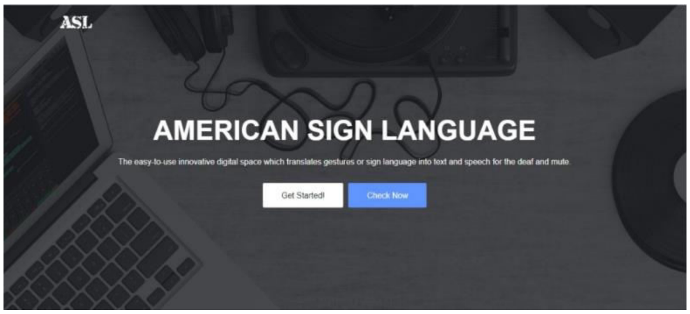
A prototype of the website is created which is a real life representation of how the project’s outcome is.Here in this prototype, the user can show a letter , its image is captured. The CNN network processes the image and recognizes the letter and converts it into text. The text is then converted into audio which is heard through the speakers connected to the PC.Hence making the project usable to a larger audience.

REFERENCES

[1] Ariya Thongtawee, Onamon Pinsanoh and Yuttana Kitjaidure, ”A Novel Feature Extraction for American Sign Language Recognition Using Webcam,” The 2018 Biomedical Engineering International Conference (BMEiCON-2018)

[2] Saad Albawi , Tareq Abed Mohammed anSaad AL-ZAWI ”Understanding of a Convolutional Neural Network,” 2017 International Conference on Engineering and Technology(ICET)

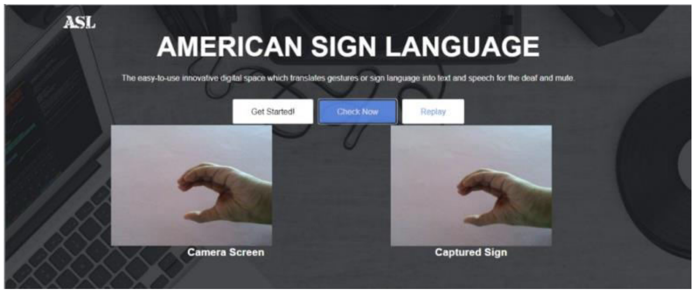
[3] Rahul Chauhan, Kamal Kumar Ghanshala and R.C Joshi,”Convolutional Neural Network (CNN) for Image Detection and Recognition,” 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)



AMERICAN SIGN LANGUAGE

The easy-to-use innovative digital space which translates gestures or sign language into text and speech for the deaf and mute.

Fig. 6. Homepage of the Website



AMERICAN SIGN LANGUAGE

The easy-to-use innovative digital space which translates gestures or sign language into text and speech for the deaf and mute.

Fig. 7. Captured sign and corresponding audio and text as seen in the website

[4] Adithya V., Vinod P.R. and Usha Gopalakrishna, " Artificial Neural Network Based Method for Indian Sign Language Recognition," Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013)

[5]P. Subha Rajam and Dr. G. Balakrishnan, "Real Time Indian Sign Language Recognition System to aid Deaf-dumb People," 2011 IEEE 13th International Conference on Communication Technology

[6] Patel, R., Dhakad, J., Desai, K., Gupta, T., Correia, S. (2018). Hand Gesture Recognition System using Convolutional Neural Networks. 2018 4th International Conference on Computing Communication.

[7] Saad Albawi,Saad Al Zawi,(2017).Understanding of a Convolutional Neural Network.2017/IEEE

[8] Raniah Zaheer,Humera Shaziya (2019).A study of the Optimisation Algorithms in Deep Learning. 2019 IEEE/International conference on inventive systems and control.

[9].JSON file
https://fileinfo.com/extension/js
on [10]HDF5 file
https://www.geeksforgeeks.org/hdf5-files-in-
python/ [11]Tkinter
https://www.tutorialspoint.com
/ [12]PYTTXS3
https://pypi.org/project/pyttsx3/

[13] Hanin M. Abdullah, and Ahmed M. Zeki , "Frontend and Backend Web Technologies in Social Networking Sites: Facebook as an Example
," 2014 3rd International Conference on Advanced Computer Science Applications and Technologies.

[14] HTML AND CSS BASICS
https://frontendmasters.com/books/front-end-handbook/2018/learning/html-css.html films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[15] Bangzhong Cao,Minyong Shi,Chunfang Li, "The Solution of Web Font- end Performance Optimization ,"2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI 2017).

[16] JAVASCRIPT
https://en.wikipedia.org/wiki/JavaScript

[17]DJANGO FRAMEWORK
https://www.djangoproject.com/start/overview/

.