Available Online at www.ijcsmc.com/journal

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 11, Issue. 1, January 2022, pg.209 – 213

Vision Pro– A Lens Application

Dr. A.A. Jaiswal[1]; Abhinav Nema[2]; Akshay Mokalwar[3]; Kalyani Paraye[4]; Prajwal Shivarkar[5]
[1]HOD Computer Department, KDK College of Engineering, Nagpur, Maharashtra, (India)
[2-5]B.E. Final Year Student, Computer Department, KDK College of Engineering, Nagpur, Maharashtra, (India)
[1] hodct@kdkce.edu.in; [2] 23abhinavnema@gmail.com; [3] aamokalwar@gmail.com;
[4] parayekalyani@gmail.com; [5] prajwalshivarkar786@gmail.com
DOI: 10.47760/ijcsmc.2022.v11i01.028

Abstract— Object and text detection are very common in day to day life. Real world applications such as Google Lens, Bixby Vision and many other apps are available in our mobile phone. The single app which can Translate, Search and detect text in real time is popular in past years. Our Vision Pro is similar to them and basically an image recognition software which can translate the text, detect handwritten text, search for that text, and search for the objects. This technology uses Camera from the device and built on neural network.
The idea of Vision Pro mainly comes from the Google Lens. It uses the Convolution Neural Network along with some major python libraries for translation, Tensorflow or Keras for model training and OpenCV for reading image and processing it. Also Vision API can also be used to make this type of program.

Keywords— Text translation, Text detection, Image search, QR (Quick Response) code, Cartoonify.

I. INTRODUCTION

When directing the phone's camera at an object, Lens will attempt to identify the object by reading barcodes, QR codes, labels and text, and show relevant search results, web pages, and information. The service is similar to Google Lens, a previous app that functioned similarly. Lens uses more advanced deep learning routines in order to empower detection capabilities, similar to other apps like Bixby Vision and Image Analysis Toolset (available on Google Play); Vision Pro can read QR code/Barcode, translate Hindi to English and vice versa, Extract text from image and search on the internet, Convert image to cartoon sketch and perform visual image search.

A. Objective

The project can be used in day-to-day life. It can be utilized by any normal user having mobile/PC. The system is used to perform copy pasting quickly and searching on the internet

B. Uses

Vision pro Lens can be used in different parts of education such as Biology, Mineralogy, Architecture and history and Marketing to achieve additional information about the object and increase the motivation of the learner. Image can be converted into extra pixelated image or into a sketch. Pointing towards the common object, once picture is clicked it identifies and display object information. Users have to point the camera towards code then it will open the information related to that. Used for translation and text search.

II. LITERATURE REVIEW

Mobile phone nowadays is a powerful scientific instrument. However, the potential of it still not fully understood and presented. Vision pro is an image recognition technology based on neural networks. The main positive aspects of using Vision pro Lens in our opinion are:

- Provided by the possibility to use personal phones to search any time.
- Interaction with any objects include education
- Creation of interaction between real and virtual worlds.

TABLE I: USING VISION PRO IN DIFFERENT FIELDS OF EDUCATION [1]

Field of Science	Way Of Using
Biology	Nowadays Lens is characterized by the possibility of biology objects recognition (animals, plants, etc.)
Mineralogy	lens can use the color and the structure of the minerals to analyse it (not available now, but we think it will be provided in the future)
Architecture and history	Analysing the building and monuments
Marketing	Analysing and checking out different real-life products like clothes

Vision pro is an image recognition technology designed to bring up relevant information related to objects it identifies using visual analysis based on the convolution neural network.

Object and text detection are very common in day to day life. Real world applications such as Google Lens, Bixby Vision and many other apps are available in our mobile phone. The single app which can Translate, Search and detect text in real time is popular in past years. Our Vision Pro is similar to them and basically an image recognition software which can translate the text, detect handwritten text, search for that text, and search for the objects. This technology uses Camera from the device and built on neural network. When directing the phones camera at an object, vision pro lens will attempt to identify the object, by reading barcodes, QR codes, labels and text , and show relevant search results , webpages and information.

Related Work

Xiangyu Zhang, Jianhua Zou, Kaiming He, and Jian Sun (Accelerating Very Deep Convolutional Networks for Classification and Detection)(2015)[2] studied about the test computation of convolutional neural networks (CNNs), especially very deep CNNs that have substantially impacted the pc vision community. Unlike previous methods that are designed for approximating linear filters or linear responses, our method takes the nonlinear units under consideration. They developed an efficient solution to the resulting nonlinear optimization problem without the necessity of stochastic gradient descent (SGD). More importantly, while previous methods mainly specialize in optimizing one or two layers, our nonlinear method enables an asymmetric reconstruction that reduces the rapidly accumulated error when multiple (e.g., $\geqslant$10) layers are approximated.

Norliza Katuk, Ku-Ruhana Ku Mahamud, Nur Haryani Zakaria (A Review Of The Current Trends And Future Directions Of Camera Barcode Reading)(2018)[3] studied about the use of camera to read and scan QR Code and barcode. Practically, a smartphone camera can be used as an imaging device for reading a barcode. Although barcode has been used for identifying products and items, the use of a smartphone camera as a reading device has not been explored thoroughly.

Dr. A. Anushya Assistant Professor (Google Lens as an Image Classifier) (2019) [4] studied about how Google lens classify the images and the algorithms used. The mobile app, Google Lens can replace all the image processing steps such that there is no need of training phase, algorithm and software. This paper provides a proposal about Google Lens. We can extract the features from images and classify the images as it belongs to which group. Google Lens can do the classification tasks very efficiently.

Viktor B. Shapovalov, Yevhenii B. Shapovalov, Zhanna I. Bilyk, Anna P. Megalinska, Ivan O. Muzyka (The Google Lens analyzing quality: an analysis of the possibility to use in the educational process) (2020)[1] studied that how lens software can be used in different fields of education and what are the other possibilities of its uses. Nowadays Lens is characterized by the possibility of biology objects recognition (animals, plants, etc.), Analyzing the building and monuments, and Analyzing and searching for different real-life products such as clothes.

III. METHODOLOGIES

The lens application is built from scratch which contains various algorithms and small function contributing to a single application. A web application is developed using python and Django able to access camera from web. This contains five different modules for each feature.

1. Text extraction using pytesseract
2. Text translation (Hindi English) and language detection.
3. QR code/Barcode scanner
4. Image to Cartoon sketch
5. Image search (Visual Search)

A. Text Extraction

Pytesseract is the python module helps to identify and extract text from images it works fine with images but it have some difficulties in live camera text detection and handwritten text detection. Pytesseract is a wrapper for Tesseract-OCR Engine. It is also useful as a stand-alone invocation script to Tesseract, as it can read all image types supported by the Pillow and Leptonica imaging libraries, including jpeg, png, gif, bmp, tiff, and others. The pre-processing of image is required before passing it to pytesseract using OpenCV [5][7]. Tesseract can identify multiple language from image. In this application we specified two languages (Hindi and English). The image first captured, pre-processed and saved in local server. Then it is accessed and text is displayed on screen.

B. Text translation (Hindi⇔English) and language detection [6], [7]

The model is trained on the dataset containing Hindi and English translations. First the language is detected and identified, then it is translated. The data is pre-processed by handling outlier's sentences, basic text cleaning like bad data (other languages), emails, copyright text, punctuation, digits, etc.
This data is tokenized and indexed so it is easily readable by machine. This is done by NLTK Tokenizers [6], every word is split into unique words and given a specific integers. Word embedding is done next, this helps to feature the most common words, and the idea is to learn a set of features and their values, so that we have dense vector representations for words.

$$\mathbf{E_{50\times 10K}} = \begin{array}{c} \begin{matrix} \text{Man} & \text{Woman} & \text{King} & \text{Queen} \end{matrix} \\ \begin{bmatrix} -1 & 1 & -0.99 & 0.98 \\ 0.02 & 0.01 & 0.94 & 0.97 \\ \vdots & \vdots & \vdots & \vdots \\ 0.78 & 0.84 & 0.91 & 0.92 \end{bmatrix} \end{array} \begin{matrix} \\ \text{Gender} \\ \text{Royal} \\ \\ \text{Kind} \end{matrix}$$

Fig. 1 Word Embedding Matrix [6]

This matrix is initialized randomly. 'Gender', 'Royal' and 'Kind' are 3 out of 50 features. So, each word from this matrix will be a 50-dimensional vector. For example, if the male gender is considered as -1 and female as 1 (only for differentiating in this example, this is a completely unbiased and random assumption), in the embedding matrix, the word 'Man' has a value -1 for the 'Gender' feature and 'King' has -0.99 for the same feature. This states that 'man' and 'king' are very similar in gender. So, when these word vectors will be placed in a 50-dimensional vector space, there are high probabilities of a greater cosine similarity in between vectors, if all other features are similar too. Now, to find the word embedding of 'Man', the embedding matrix (E) is multiplied with the one-hot encoded vector ($O_{5\ 5\ 4\ 5}$) for 'Man':

$$\begin{matrix} \text{Man} & \text{Woman} & \text{(All words)} & \text{King} & \text{Queen} \end{matrix}$$
$$\begin{bmatrix} -1 & 1 & \dots & -0.99 & 0.98 \\ 0.02 & 0.01 & \dots & 0.94 & 0.97 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.78 & 0.84 & \dots & 0.91 & 0.92 \end{bmatrix}_{50\times 10K} \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \boxed{1} \\ \vdots \\ 0 \end{bmatrix}_{10K\times 1} = \begin{bmatrix} -1 \\ 0.02 \\ \vdots \\ 0.78 \end{bmatrix}_{50\times 1}$$

Fig. 1 Word embedding for specific word [6]

The result is a 50-dimensional vector representation of 'Man'. This is called the 'embedding' for the word 'Man'. Similarly, all the words in this corpus are represented as 50-dimensional vectors. Some well-known pre-trained word embeddings are Word2Vec, Glove and fasttext.

As a next step, Sequence to sequence models are used when the output is to be predicted after reading the whole sequence. This has encoder-decoder architecture and two LSTMs each for encoder and decoder. After building and combining every step a model is ready to train and prediction. Prediction is done using beam search, which gives the advantage of displaying top two probabilities with the predictions, and then verifying if the translation makes sense accordingly. Prediction can be also done without beam search.

C. QR code/Barcode scanner

QR code/Barcode Scanner uses Pyzbar [3] module. It can decode the codes and show the data attached with it. The code is read by OpenCV and redirected to the link using webbrowser. No image have to be stored in server, just point camera towards code. Pyzbar library is self-sufficient to read any QR code/Barcode easily. The data can be stored in a file or directly displayed.

D. Image to Cartoon sketch

Image can be directly captured from camera or can be uploaded already stored image. This image is first converted into RGB vector image. This image then reduced to grayscale, then medianblur, edges are highlighted using adaptiveThreshold. Once the edges are identified, a bilateral filter is applied to the original image which gives the color matrix. Now the image is inverted to the RGB using bitwise_and and mask is applied of the edges which is highlighted in previous step. All this features are in OpenCV.

E. Image Search

This feature help to search for the similar type of images. This is also known as Content-Based image retrieval. Content-based image retrieval (CBIR) [9] is a system for retrieving relevant images based on a given image. The system consists of an image query and an image database. Features are extracted from all the images whether from query or from database by feature extraction algorithm.

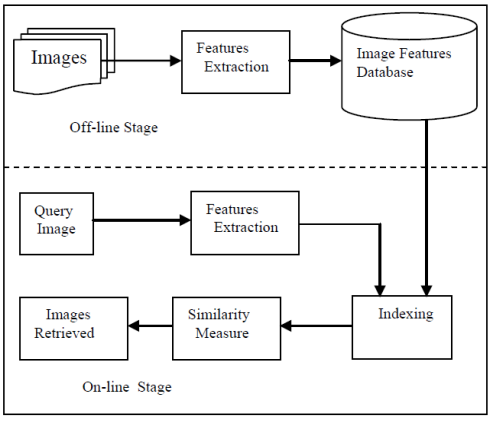

Fig. 3 Flow of the CBIR (Alkhawlani et al. 2015) [10]

For extracting the features from the images we use Convolution Neural Network (CNN), Convolutional Neural Network consists of layers, such as convolutional layer for feature extraction, pooling layer for sampling the features, and fully-connected layer for doing prediction.

$$dist(q, img) = ||q - img||_2 = \sqrt{\sum_{i=1}^{n} (q_i - img_i)^2}$$

Fig. 4 Euclidean distance

Where: *q = the query, img = the image, n = the number of feature vector element, i = the position of the vector. [9]*
After feature extraction, we calculate distance between the query and all images using Euclidean distance, smaller the distance more similar the image is. Tensorflow is used for feature extraction and Numpy for calculating distance. This image search works for small scale. Since the Google never made its algorithm public, this is the best way to implement image search. After we retrieve all of the images, now we can extract the features from all images using CNN [14] and save those features on .npy format for later use. To specify the architecture, we will use VGG-16 architecture and pre-trained weight from the ImageNet. [9]

IV. RESULT AND CONCLUSIONS

After combining each methodology into a web application using Python and Django, this application can be accessed from any device having internet connection and a good quality camera once this application is hosted on web.

There are many ways of performing the same tasks but we tried to choose the optimal methods which require less time and require less space. The language translation can be done sing Google translate library but it is better to do it manually which allow to add more data in the dataset according to the condition. LSTM [13] is best way to do it. For text extraction from images and real-time camera, it is best to use Tesseract instead of EasyOCR or KerasOCR, since it can also detect handwritten text and multilingual text.

Future scope of image search is very vast. Not only similar images but it can also be upgraded to search for similar objects in the image. Translations can also support new languages with training the models with different languages. This whole application can be converted into a mobile app in the future with more optimizations and can be used in large scale datasets

REFERENCES

[1] Viktor B. Shapovalov, Yevhenii B. Shapovalov, Zhanna I. Bilyk, Anna P. Megalinska, Ivan O. Muzyka 2020. The Google Lens analyzing quality: an analysis of the possibility to use in the educational process.
[2] Xiangyu Zhang, Jianhua Zou, Kaiming He, and Jian Sun 2015. Accelerating Very Deep Convolutional Networks for Classification and Detection.
[3] Norliza Katuk, Ku-Ruhana Ku Mahamud, Nur Haryani Zakaria 2018. A Review of the Current Trends and Future Directions of Camera Barcode Reading.
[4] Dr. A. Anushya Assistant Professor 2019. Google Lens as an Image Classifier.
[5] A comprehensive guide to OCR with Tesseract, OpenCV and Python. Available: https://nanonets.com/blog/ocr-with-tesseract/
[6] Richa Ranjan, rahiakela 2020. Neural Machine Translation for Hindi-English: Sequence to sequence learning. Available: https://medium.com/analytics-vidhya/neural-machine-translation-for-hindi-english-sequence-to-sequence-learning-1298655e334a
[7] Maharishi Roy 2021. English to Hindi Neural Machine Translation. Exploring basic NLP preprocessing and Encoder-Decoder architecture ubiquitous in all Seq2Seq tasks. Available: https://towardsdatascience.com/english-to-hindi-neural-machine-translation-7cb3a426491f
[8] Harald Scheidl 2018. Build a Handwritten Text Recognition System using Tensorflow. A minimalistic neural network implementation which can be trained on the CPU. Available: https://towardsdatascience.com/build-a-handwritten-text-recognition-system-using-tensorflow-2326a3487cd5
[9] Irfan Alghani Khalid 2020. Build An Image Search Engine Using Python. Available: https://towardsdatascience.com/build-an-image-search-engine-using-python-ad181e76441b
[10] Alkhawlani M, Elmogy M, El Bakry H. 2015. Text-based, Content-based, and Semantic-based Image Retrievals: A Survey. International Journal of Computer and Information Technology.
[11] LeCun Y, Bottou L, Bengio Y, Haffner P. 1998. Gradient Based Learning Applied to Document Recognition. Proc. Of The IEEE.
[12] Matsui Y. Simple Image Search Engine. Available: https://github.com/matsui528/sis
[13] Xiaoyan Zhu, Yequan Wang, Li Zhao, Minlie Huang 2016. Attention-based LSTM for Aspect-level Sentiment Classification.

