DESIGN OF TRAFFIC AMERCEMENT AUTOMATION USING COMPUTER VISION

[1]Mrs Mangala HS, [2]Abhijeet Kumar, [3]Abhishek Mishra, [4]Aditya Kumar Agarwal, [5]Anagha Shree CA

1Assitant Professor, 2Student, 3Student, 4Student, 5Student
1Department of Computer Science and Engineering
1Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India

Abstract

In order to transform various types of electronic documents, such as scanned documents, digital images, and PDF files into fully editable and searchable text data, optical character recognition (OCR), a classic machine learning challenge, has been a long-standing topic in a variety of applications in the healthcare, education, insurance, and legal industries. OCR is a crucial and fundamental technique for data analysis due to the daily, fast creation of digital pictures. We have been able to save a respectable amount of time and effort when developing, processing, and storing electronic documents and customising them for various uses with the use of OCR technology. There are now several alternative OCR systems that, in addition to making theoretical advances to various practical disciplines, have shown effective implementations in the actual world. . Four well-known OCR services—Google Docs OCR, Tesseract, ABBYY FineReader, and Transym—have been used in this work to conduct a number of qualitative and quantitative experimental assessments. We use a dataset of 1227 photos from 15 distinct categories to analyse the accuracy and dependability of the OCR software. We also examine contemporary OCR uses in healthcare informatics. The current assessment is anticipated to promote optical character recognition (OCR) research by offering fresh perspectives and attention to the study field and helping researchers choose the most accurate and effective optical character recognition service. Nowadays web is becoming the main channel for reaching customers and prospects; Clickstream data generated by websites has become another important enterprise data source. As simple as it sounds for recording every click a customer made so that we can use clickstream data for modelling user behaviour, and gaining valuable customer insights. Clickstream analysis commonly refers to analyzing click data and website optimization. Such analysis is typically done to extract insights into website visitor behaviour especially social-media or e-commerce websites. Also, nowadays online learning became a trend in the education system. We can see many online learning portals which are providing live training on various technologies. To identify potential customers or to identify recommendations for existing customers. Clickstream analysis can be used to figure out which geographies and time zones most of the traffic comes from, and which devices, Browsers (such as their name, versions), time spent, Operating Systems, are used to access the websites, which common paths users take before they do something in the site. Analysis of clickstream data in real-time (streaming) has more value than batch mode(stored). We analyze and visualize online- learning portal's clickstream data on the fly for business intelligence purposes. In this paper data pipeline creation is proposed using tools such as Apache Kafka, Apache Spark for streaming and Apache Cassandra and Flask to query and visualize clickstream data respectively.

Index Terms- *Batch, Clickstream, Data, Pipeline*.

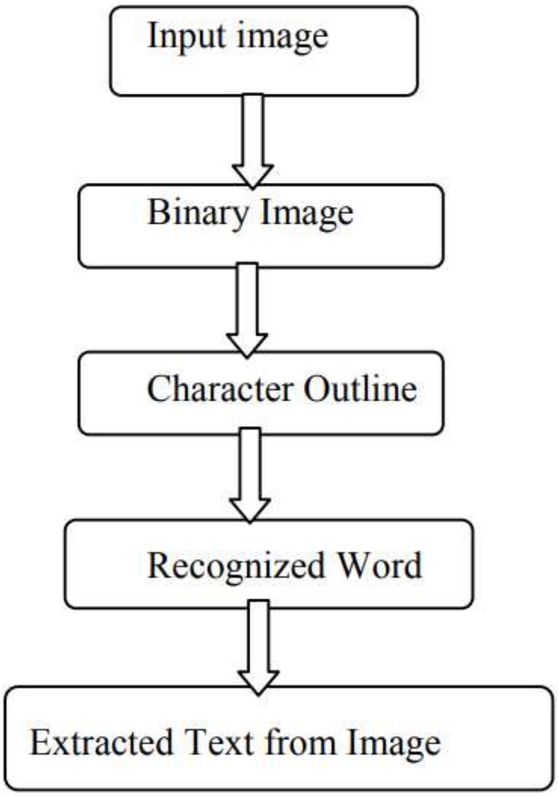
I. INTRODUCTION

Natural language processing (NLP), computer vision, biomedical informatics, machine learning, and others have all found optical character recognition (OCR) to be a particularly useful study topic. In order to facilitate additional processing activities, this computational technique has been used to transform scanned, handwritten, or PDF files into editable text formats (such as text files or MS Word/Excel files).

For instance, dealing with a sizable number of patient paperwork (such as insurance forms) has been necessary in the healthcare industry. It is essential to enter the patient data in a standard format into a database so that it may be accessed later for analysis in order to analyse the information in such formats. We can quickly capture each patient's data by using OCR technology to automatically extract it from the forms and enter it into databases. OCR makes the procedure incredibly simple by converting those documents into text that can be searched and edited with ease. In terms of software engineering, "Software as a Service" (SaaS) has evolved as both a design pattern and a delivery mechanism for the architectural concept of centralised computing.

OCR has significantly aided in the process improvement of several real-world applications in the fields of education, banking, insurance, and healthcare. A clickstream is the recording of the client taps on while perusing site or utilizing other programming applications. As the client clicks any place in the website page or applications, the activity is logged inside the web server or on a customer, and also conceivably the router, proxy server or web browser.

Tesseract is an open-source OCR toolkit that was first created by HP and made available under the Apache Licence [35] for a variety of operating system platforms, including Mac OS X, Linux, and Windows. Tesseract development has been maintained by Google since 2006 [36], and it is one of the most widely used OCR systems globally [29]. In phase one of the Tesseract method, the input picture is transformed into a binary image using adaptive thresholding algorithms [37]. Then it applies linked component analysis to extract letter layouts, converting those layouts into blobs—regions in an image data that have different qualities, such colour or intensity, from nearby pixels. Analysis of clickstream data is valuable for web movement investigation, statistical surveying, programming testing and for dissecting representative profitability. Starting clickstream or snap way information must be gathered from server log documents. As it were, data was gathered just from "genuine people" tapping on webpages through browsers.



The primary purpose of clickstream following is to comprehend client conduct and give website admins understanding into what guests are doing on their website. An worldwide firm called "ABBYY" [23] has invented and developed ABBYY FineReader as a sophisticated OCR software solution to offer top-notch OCR services. For many years, it has been enhancing optical character recognition's core capabilities and delivering encouraging outcomes for text retrieval from digital pictures [28]. ABBYY FineReader is a commercial software application, and its source code is not publicly accessible, therefore the academic community has not yet seen examples of its underlying algorithms. This information itself is "neutral" as in any dataset is neutral. The information can be utilized for different reason, for marketing. Furthermore, scientist, any website admin, blogger or individual with a site can find out about how to enhance their webpage. Utilizing clickstream information can raise security concerns, particularly since some Internet specialist co-ops have turned to offering clients clickstream information as an approach to improve income. Analyzing the information of clients that visit an organization or association's site can be imperative with a specific end goal to stay focused. Retrospective OCR experiments have mostly concentrated on obtaining medical data for research purposes. For a significant genome-wide association research, Peissig et al. [42] employed OCR to supplement existing electronic health record data by extracting cataract subtypes and severity from handwritten ophthalmology forms

II. PROBLEM DESCRIPTION

. Optical Character Recognition (OCR) is an established problem statement in machine learning and artificial intelligenceA well-known issue in artificial intelligence and machine learning is optical character recognition (OCR). The issue arises when the available data is somewhat vague and uncontrolled, which is precisely the case with handwritten text recognition, despite the fact that most people think the situation is clear-cut.

There are many issues of intense research that still unsolved or solved with limited success for example: The optimization of feature processing time and query response time are important for a huge image database. The selection of algorithms, parameters etc. is very important because specific one is not necessary suitable for all application e.g. the segmentation algorithm used for natural images may not be suitable for medical images. An appropriate index structure that allows efficient searching of large image database is still a problem under research. On the other hand, the object recognition and detection is one of the most challenging problems in image retrieval. The addressing of this problem is urgent. In general, the object should be recognized regardless of the illumination changes, changes of size, rotation, background clutter, viewpoint change, and occlusion. However, this cannot be obtained in most CBIR because it does not have adequate ability to capture important properties of object. In support of better object recognition, the interest point’s detectors were introduced to represent the local features of images in image retrieval systems. Binary histogram had also been proposed that can be used for image classification and retrieval without image segmentation which provides good retrieval performances, but it only works on gray level images

III. RELATED WORKS

Sequence alignment has been widely applied in various domains to study the similar and different properties of sequences from the same resource, for example, aligning protein sequences or DNA sequences in bioinformatics and aligning sentences from different languages in machine translation. Dynamic programming is the core of many sequence analysis methods, e.g. dynamic time warping, edit distanceand linear HMM . Alshawi et al proposed an alignment algorithm to search pairings of words from bitexts (source language sentences with their translations) for machine translation, which makes use of dynamic programming to learn a mapping function minimizing the total costs of a set of pairings. Needleman-Wunsch algorithm and Smith-Waterman algorithm are well-known pairwise sequence alignment algorithms for protein and DNA alignments, both of which are extensions of edit distance with a predefined linear gap penalty and a similarity matrix to specify the scores for aligned characters. Hobby created ground truth for OCR’s by using a machine readable description to print the document and then matching character bounding boxes with bounding boxes derived from a scanned image of the document. Xu et al. aligned an imperfect transcript obtained from a scanned image of a printed page with the characters in unsegmented text image. Neither of these are really appropriate since we do not have the approximate mapping that is required nor are we aligning im ages with text. HMM is a model widely used for alignment tasks in different domains, e.g. for sequence alignment in speech recognition the alignment of synthesized speech with speech machine translation aligning parallel corpora in machine translation the alignment of speech recognition output with captions in video. Krogh et al. proposed to use a linear HMM as a structure generating protein sequences by a random process. It is basically a hidden Markov chain with three kinds of state nodes: match, insert and delete, in which all transitions and character distance costs are position-dependent, i.e. different distributions are associated with the same kind of states or transitions at different positions. Unlike Krogh’s linear HMMs, the HMM at each level of our hierarchical alignment approach directly takes positions as states and calculates the probability of generating a sequence of OCR output given any possible sequence of positions in the ground truth. That is, there is a state corresponding to every position in the ground truth sequence. This structure is very similar to the HMM model proposed by J.

Rothfield et al. for word by word alignment of scanned handwritten document images with ASCII transcripts. This model is not hierarchical and is not practical for aligning large sequences. In this paper we seek to align text to text not text to images. Given the problem and domain differences, the transition probabilities and generative probabilities have to be and are defined differently. The details are given in section 2.1. For the same task of handwriting alignment, Kornfield employs dynamic time warping (DTW) to align feature sequences extracted from word image series with ASCII transcripts, which is essentially an edit distance based global alignment method with deletion, insertion and match costs uniformly defined as the dissimilarity between corresponding items from two time series. Compared with edit distance based alignment algorithm, the HMM based alignment allows one to learn the domain knowledge through training over aligned or even unaligned sequences and formulate the probabilities of alignments using arbitrary distributions and is more flexible and powerful.

IV. MATERIALS AND METHOD

Many feature like translation and rotation cannot be retrieving by CBIR. So Scale Invariant Features transform or widely known as SIFT is one of the techniques that have been successfully used for local interest point detector and its descriptors. Scale-invariant feature transform (or SIFT) is an algorithm in computer vision to detect and describe local features in images. Local image features are specific image properties of local image region such as edges, corners, lines and curves. SIFT is not just scale invariant but also invariant to rotation and translation.

At the same time we give feature for image retrieval on character based. It search image on message contain in the image. For that Optical Character Recognition is used. OCR the mechanical or electronic conversion of images of typed, handwritten or printed text, whether from scanned document, a scene-photo. For that Tesseract will use. After recognizing text from image by OCR, it will store that message in any file. After that we used Boyer–Moore string search algorithm, for searching string stored in file. The Boyer–Moore string search algorithm is an efficient string searching algorithm that is the standard benchmark for practical string search literature. It was developed by Robert S. Boyer and J. Strother Moore in 1977. The algorithm preprocesses the string being searched for (the pattern), but not the string being searched in (the text). The algorithm runs faster as the pattern length increases.

For this software requirements are java and jdk 1.7. Ide, netbean 6.9 and operating system is windows x.From the following architecture shows the how OCR work with other feature for image retrievals. From the diagram it shows that, image can be retrieve with different feature of text, colour and texture along with it can be retrieve with the text contain in that image. This architecture shows with the help of OCR, it can be recognized characters. The Fig 3: shows the process flow of the image retrieval system and is described as follows:

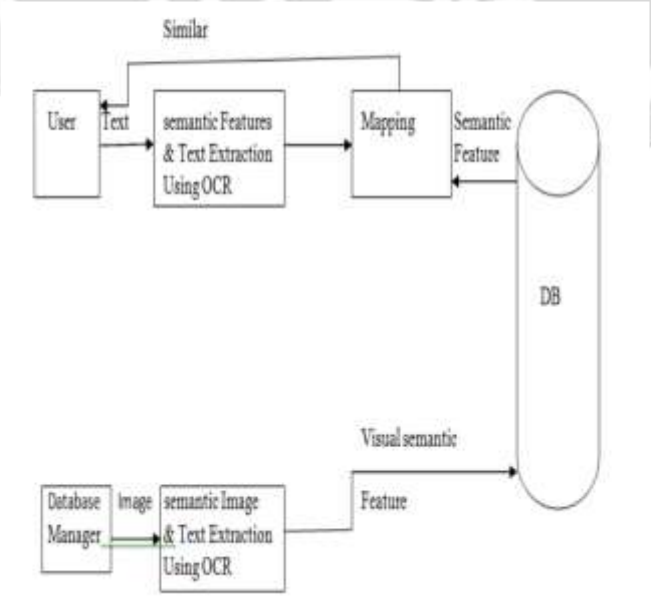

Similar
User
Text
semantic Features & Text Extraction Using OCR
Mapping
Semantic Feature
DB
Visual semantic
Feature
Database Manager
Image
semantic Image & Text Extraction Using OCR

Fig1: flow diagram of proposed methodology

- Collection of Database: The database contains the collection of images which are stored in any one of the formats .jpg, .bmp, .tiff

- Query: The user provides the sample image or the text as a query for the system.

- Feature extraction: There are various kinds of visual features to represent an image such as colour, texture, size and spatial relationship. Single feature can represent the part of the image property. Hence the combination of all the features of an image is used for effective image retrieval Similarity mapping:

- Text Extraction using OCR: In this it retrieves the text from the image and stored in data base.

- Retrieval: The system retrieves the images based on the sequence of ranked images, or given text from the database.

The alignment at the upper level aims at providing a rough alignment between two sequences on a larger scale and allows us to break up the original problem of aligning long sequences into the problem of aligning much shorter subsequences. These subsequences are aligned at a lower level. Given the ground truth and the OCR output for a book, the hierarchical approach works as follows:

1. At the top level, we look for and align a set of unique words in order to partition an entire book into small portions. It is done in 3 steps.

(a). we first extract all the unique words in the ground truth, each of which occurs only once in the book, and create a word list A which is sorted according to the order that they appear in the book. According to the Zipf's law on the distribution of word frequencies in a natural language document, almost half of the distinct words are unique.

(b).For each unique word in the ground truth, we look for the same words in the OCR output. Because of OCR errors and duplicate pages, it is possible that a unique word has no correspondence or more than one correspondence in the OCR output. We, therefore, filter out those words from the list A, which have no correspondences in the OCR output and whose immediate neighbors do not match. The words in the OCR text which correspond to the filtered outputs in A form a sorted word list B which is ordered according to their position in the OCR text.

(c).Using our alignment model(section 2.1), we filter out those repeated correspondences caused by redundant texts in OCR output from list B and finally get a one to one mapping from the unique words in the ground truth to those in the OCR output. The unique words after filtering and alignment are called anchor words.

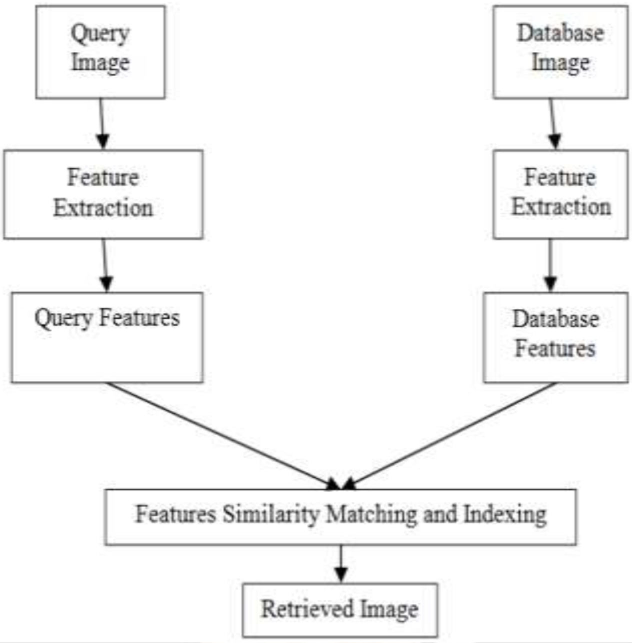


Fig2. Activity in fine generatoin

2. At the middle level, we use anchor words as boundaries to partition the OCR output and the ground truth of the whole book into smaller corresponding subsequences. Using our alignment model, we align each pair of subsequences at the word level.

3. After word alignment, exactly matched words are directly mapped to the character level. Using exactly matched words as boundaries, we align the texts between every pair of these boundaries at character level.

The first step in the hierarchical alignment framework is quiet robust. Even if there are large chunks of texts missed, reduplicated, or wrongly recognized, the anchor words can be correctly located and aligned. After these three steps of alignment, we finally get the character by character alignment between OCR output and ground truth . The next subsection describe the details of our HMM-based alignment model at each level of the hierarchical framework.

Text Based image retrievals (TBIR) TBIR is very popular framework. The text based image retrieval was introduced in the year 1970s. TBIR was first annotated the images by text and then used text-based database management systems to perform image retrieval. TBIR is used to manually annotate the image in the database with Annotations keywords or Descriptions. This process is used to describe both image contents and other metadata of the image.

Optical Character Recognition (OCR) OCR is Optical Character Recognition (Optical Character Reader, OCR) is the mechanical or electronic conversion of images of typed, handwritten or printed text, whether from scanned document, a scene-photo. TesseractOCR, optical character recognition software, to image spam mail filter. Tesseract-OCR is free OCR software and licensed with the Apache 2.0 License. Tesseract was probably the first OCR engine able to handle white-on-black text so trivially. At this stage, outlines are gathered together, purely by nesting, into Blobs. Blobs are organized into text lines, and the lines and regions are analyzed for fixed pitch or proportional text. Text lines arebroken into words differently according to the kind of character spacing. Fixed pitch text is chopped immediately by character cells. Proportional text is broken into words using definite spaces and fuzzy spaces.

IV CONCLUSION

In this paper, we proposed a hierarchical alignment method for aligning OCR output and ground truth for books. Our hierarchical alignment approach partitions the alignment problem for an entire book into the problem of aligning many shorter subsequences. A HMM-based model is employed for alignment at each level. Experimental results show that even on OCR output with high error rate, our alignment method works very well.

In this paper we have discussed various method and techniques of image retrievals with different parameter. CBIR method has been widely used in various areas to improve the performance of the system and achieve better results in different applications. More key points of SIFT features will be detected for image that have corners and edges which make the retrieval results better than image that has lesser corners. Our experiment show that OCR is search image which contain text message on it, it is alternative way of image retrieval by using OCR. Text retrieval is the basis of image retrieval, Many techniques come from this domain. Text and image feature of retrievals combined, have biggest chances for success.

In our study we proposed a system which is talked about in the past area is assessed utilizing clickstream data. This depends on a normal extraction of clickstream data continuously. One-click on the website will produce complete information about users. We'll both benefit from batch data analysis and real-time data analysis using Big Data tools. The advantage of analysing the real-time clickstream data and stored data can be used for prediction purposes. Also, we can able to detect what is happening at the moment on our site. The results will be shown in will be a complete open-source solution to analyze and process real-time streaming clickstream data. The solution is basic and assessed utilizing the technical support clickstream data collected from the online learning portal or website. In this solution, we discuss about the clickstream data and analyze user behaviour from them, also this technique and procedure are relevant to any real-time data analysis. The information can be utilized for a different reason, for marketing. Furthermore, scientists, any website admin, blogger or individual with a site can find out about how to enhance their webpage.

In this paper we have discussed various method and techniques of image retrievals with different parameter. CBIR method has been widely used in various areas to improve the performance of the system and achieve better results in different applications. More key points of SIFT features will be detected for image that have corners and edges which make the retrieval results better than image that has lesser corners. Our experiment show that OCR is search image which contain text message on it, it is alternative way of image retrieval by using OCR. Text retrieval is the basis of image retrieval, Many techniques come from this domain. Text and image feature of retrievals combined, have biggest chances for success

REFERENCES

[1] Pal, G., Atkinson, K. & Li, G. Real-time user clickstream behavior analysis based on apache storm streaming. Electron Commer Res (2021).

[2] Baker, R., Xu, D., Park, J. et al. The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: opening the black box of learning processes. Int J Educ Technol High Educ 17, 13 (2020).

[3] Abdul Jabbar, Pervaiz Akhtar, Samir Dani, Real-time big data processing for instantaneous marketing decisions: A problematization approach, Industrial Marketing Management, 2020, ISSN 0019-8501.

[4] Frhan, Amjad. (2017). Website Clickstream Data Visualization Using Improved Markov Chain Modelling In Apache Flume. MATEC Web of Conferences. 125. 04025. 10.1051/matecconf/201712504025.

[5] Dasgupta, A., Arendt, D.L., Franklin, L.R., Wong, P.C. and Cook, K.A. (2018), Human Factors in Streaming Data Analysis: Challenges and Opportunities for Information Visualization. Computer Graphics Forum, 37: 254-272.

[6] Birjali, M., Beni-Hssane, A., & Erritali, M. (2017). Analyzing social media through big data using infosphere biginsights and apache flume. Procedia Computer Science, 113, 280–285.

[7] Hanamanthrao, R., & Thejaswini, S. Real-time clickstream data analytics and visualization. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 2139–2144.

[8] Ichinose, A., Takefusa, A., Nakada, H., & Oguchi, M. A study of a video analysis framework using kafka and spark streaming. In 2017 IEEE International Conference on Big Data (Big Data), pp. 2396–2401.

[9] Liu, Z., Wang, Y., Dontcheva, M., Hoffman, M., Walker, S., & Wilson, A. (2017). Patterns and sequences: Interactive exploration of clickstreams to understand common visitor paths. IEEE Transactions on Visualization and Computer Graphics, 23(1), 321–330.

[10] Pal, G., Li, G., & Atkinson, K. (2018). Multi-agent big-data lambda architecture model for e-commerce analytics. Data, 3(4), 58.

[11] Pulkit Sharma, Komal Mahajan, Vishal Bhatnagar, “Analyzing Cilck Stream Data Using Hadoop” Second International Conference on Computational Intelligence & Communication Technology (CICT).

[12] Nikitha Johnsirani Venkatesan, Earl Kim, Dong Ryeol Shin, “PoN: Open source solution for real-time data analysis” Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC).

[13] Rajat Kateja, Amerineni Rohith, Piyush Kumar, Ritwik Sinha, “VizClick visualizing clickstream data” International Conference on Information Visualization Theory and Applications (IVAPP).

[14] Zikopoulos, Paul, and Chris Eaton. “Understanding Big Data: Analytics for enterprise class hadoop and streaming data.” McGraw-Hill Osborne Media.

[15] Scholz, M., et al. (2016). R package clickstream: analyzing clickstream data with markov chains. Journal of Statistical Software, 74(4), 1–17.