 ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

Study of Information Extraction and Optical Character Recognition

Swayanshu Shanti Pragnya
Dept. of Computer Science and Engineering, CUTM, Bhubaneswar, Odisha, India

Abstract
In the intensification rate of techniques and its application towards the convenience of human being is in ceaseless process. While techniques raises the question of storing data and retrieving is common in mind. Text mining is high in demand and been the most interesting way of different data processes. As the name extraction itself shows to retrieve from ancestry of data and information can be any knowledge get by some data. So all together lineage of data is what information extraction does. Here extraction of information will be conducted from images. But information extraction is consisting of different parts like its type, orientation, process and finally a technique for executing the whole process. So this paper is all about the basic of information extraction, its type, condition, process and finally conclude with the Optical Character Recognition (OCR) tool which can make the whole extraction process efficient by the study of different journals. Gathering an overall ideation regarding information extraction from images and its applicable tool is the objective which can make the whole retrieving process convenient.

Keyword
Information Extraction, Text Extraction, Text Information Extraction, Text Detection, OCR

I. Introduction

Information extraction is a part of text mining which is used to extract information to make the data structured from unstructured data. Generally, for getting information, the schemas, relations or PDF are easy to find. Identification of each entity is not structured data or by simply identifying an entity is not the chronological way of information extraction rather it must be a part of machine learning through Natural Language process. We can identify the extraction in two broader manner those are:

A. Hand Coded or Learning Based

This method requires human effort or any set of rules or regular expression for the extraction purposes. Basically these learning based system requires to label unstructured information for training machine learning extraction models. But for choosing various data models, machine learning should possess to choose between robust unseen noisy data, which can be decided from learning based system and hand coded data.

B. Rule- based or Statistical

Hard predicates are responsible for making the rule based extraction easy in terms of interpretation and development. In case of open-ended domain like opinion extraction from any article, statistical method is more appropriate.

II. Geometrical Orientation before Text Extraction

As now-a-days large number of information are stored in image, so extracting text from it is required to store in digital library for further uses. But variation in terms of size, alignment, size, orientation make the extraction more complex. One of the method of indexing like Content based indexing which attach labels in attached image. Only indexing is not the solution of extraction. There are two kind of content i.e. Perceptual (attributes like color, shape or texture) and Semantic content (events and their relations). Text within an image is relatively easy to explain as easily get compared with semantic contents. Extraction of text from images is getting complex due to the probability of getting unstructured data in terms of style, size, texture, alignment or orientation can varies from image to image. Before extraction of text from any image there should be proper geometrical uses of orientation which is going to help in segregation and later on extraction purposes

A. Shape

In case of translation, scaling or rotation cases shape is required to be identified accurately in retrieval process. Generally the shape is categorized in two forms i.e. boundary based and another is region based. 2D shape representation are the main reason behind confusion as identification is relatively poor in comparison to 3D. As now it's a time of 3D where more accurate data can be seen. From identification to other uses also can be performed. Generally identifying a shape just by watching its 2D structure is bit difficult as we see only in terms of height and width. Here extraction process can take so many time as the region and boundary is not properly segregated. All kind of assumption has to be taken for a 2D text in images. If height of the object can be known in terms 3D then the boundary of each text also can identified easily. Structure plays a vital role as by just seeing an image in text, its extraction is not possible but in a live object by the help of 3D printer it can be possible.

B. Size

Text size can varies from text to text depending upon image dimension in terms of height and width. For any kind of domain it should be changed according to domain.

C. Alignment

These characters may arise as a non -planar texts which causes special effects. As alignment plays vital role in finding a single row or column in characters, while extracting data in other dimension.

D. Color

This feature is the most widely used visual characteristic found in image retrieval. Color is totally independent of the complications found in size and orientation of an image. Generally in image retrieval, color histogram is the most used color feature representation format. In statistics it denotes the probability in intensity of three color channels.

Color histogram is not the only feature representation for retrieving image there are other representations like color moments and color sets. For recovering from quantization effects in color histograms another approach of color moments has been used. Mathematically this approach proposed by Stricker and Orengo

is that color distribution to be featured s its moments [2]. For calculation color similarity Weighted Euclidean distance method has been used.

The character in the text may or may not have same colored text in a single line. But inter character distance will provide same distance in each text which will be helpful in component based approach in text detection. But finding monochrome will solve so many problems and will provide very accurate data.

E. Texture

It is a visual pattern that has homogeneity property which result not from a presence of single color rather from various intensified colors. These textures carry information regarding surface arrangements (Structural) and it's relation with surrounding environment.

F. Compression

Digital images are recorded, transferred and processed in compressed format.

III. Text Information Extraction (TIE)

These system receives input in form images. The images can be of any way like in grayscale or colored, text in middle of the image or above, compressed or un- compressed. So it is difficult to extract text from images due to the discontinuity in color, size or orientation. The different kind of problems in TIE can be

1. Detection
2. Localization
3. Track
4. Extraction and
5. Identification (OCR)

Any TIE system consist of following steps i.e.

Images → Text tracking (Text Localization + Text Detection) → Text enhancement and extraction → Recognition (OCR) → Text

IV. Text Detection

This is the initial stage in terms of getting surety of a text in an image. Here the existence of the text in image is determined. For detection of text according to Kim [1], low threshold value is needed for scene change detection as the space occupied by a text in image is relatively small as compare to image dimension. This approach is sensitive but an efficient solution for video indexing application which only need the keywords from video clip than entire text. But in case of invoices this method may not be applicable as the assumption in terms of text size in image may vary from image to image. These textual size assumption can't be followed for other data due to the inconsistency in what kind of image or text is going to be extracted.

V. Text Extraction

A. Text Document Images

Approach given by S. Nirmala [3] for extracting text in the complex background color document images, the method used was canny edge detector for detecting edges. Due to the presence of dilation operation on the edges which creates holes in the nearest components which rather creates a character string. The components which are not nearest for the dilation they are eliminated automatically. But here the problem can be only connected sets will be identified if there will be more conditions based on situation then this method may not provide accuracy. For eliminating other non -text components, the method used was analysis of standard deviation from connected component and computing. Performing segmentation an unsupervised local threshold was reprocessed. At the end text regions are found and reprocessed. The methods like canny edge detector, dilation operation, unsupervised local threshold, connected component analysis are giving 97.12% accuracy in handling degradation as blur or wavy text format.

Based on the report by Davod .et.al [4] a dynamic threshold used in the detection of edges from wavelet coefficient. But for further effective edges were got by the use of alternative heuristic threshold by blurring approximate coefficient. For final text extraction Region of interest was used. Evaluation of 80 pictures were done. Here accuracy rate was 91.20% which is beneficial for robust to noise form by using the methods like wavelet transform and ROI.

An approach by Sachin [5] was for embedded text in complex colored document images. The methods which they used were simple edge detection, Thresholding technique and Block classification. Here conversion of images from gray scale was performed by simply taking the weighted sum of RGB components. For edge detection method used was simple gray scale conversion with simple masks which separates horizontal and vertical edges. After getting edges they are divided into small overlapping blocks of m by m pixels where m is the resolution of image. After these block classification methods was used for differentiating text from image by the help using pre-defined threshold. These approach gives 99% accuracy in the insensitive to color fonts.

B. Scene Text Images

J. Fabrizio has introduced a technique for detection as well as extraction of text from commercially taken screenshot images [6]. For labeling they combined two methods as blob extraction method (Edge based method + connected component labeling method). Extraction process was done by the collation of homogeneity detection filter and threshold number. Here the result of successful extraction was 94.66% which works on complex background.

A new approach carried by Shivkumaraet. al [7] i.e. Boundary growing method (BGM) and multi-color difference (MCD) of multi orientation handwritten scene scene of text from video. MCD was used for increasing gap between text and non –text pixel. For obtaining text clusters the K-means algorithm has been used. By the help of K-means algorithm it obtained text candidates which further helps in eliminating false candidates. BGM used for fixing boundaries in separated clusters. Here accuracy rate was found to be 89.67% in insensitive to contrast.

C. Heterogeneous Text Images

Keechul Jung approached another method for multi oriented graphics and scene text in video images [8]. He has used laplacian operator for highlighting transitions in between background and the text. K-means for classify text and non-text edges or region. For the segregation of artifacts from text cluster the morphological operation has been used. Here accuracy rate is 84.90% for multioriented images.

Miriam Leon has approached a text extraction algorithm which is insensitive to noise skewness and text orientation, color or intensity of any kind of heterogeneous document images. For identifying edge mathematical morphology has been used. The

variance was found by gray levels of components. The text was segregated from connected components with some threshold values. It provides the accuracy of 84.01% insensitive to noise.

D. Edge-based Method

Different kind of methods are available in any image but edge based is focusing on the contrast in between text and in the background. The edges of any text boundary identification is the first role and get merged. Generally edge filter is used for the edge detection. According to Smith [10] the input image will be in different filtered with 3 by 3 horizontal to image and perform threshold for finding vertical edges.After successful completion of smoothing operation it eliminate small edges and adjacent edges by simply finding the connection with each other. For final extraction of text intensity histogram of each cluster is used to find similar shape and texture in characters.

In case of Chen et al. [11], they used an operator named as Canny operator for the detection of edges in an image. Edges of each text are enhanced in terms of scale of information. Morphological dilation is used to connect all the edges into any cluster. Horizontal and vertical aspect is used in finding the filter out of non- text clusters.

E. Texture-based Methods

The problem face in traditional texture based methods are there computational complexity while the stage of texture classification, which increases the processing time. But it is also require to scan the input image for detection and localization of text regions.

According to Sin et al. [12] in case of real scene images by using the frequency features such as number of edges in pixels, frequency features as number of horizontal and vertical lines in scene. By taking the assumption that many of the text regions are rectangular in background, using Hough transform detection of edges is taken place. But it is not clear that these three stages will provide any final result or not.

Wavelet transform is used for text localization approached by Mao et al. [13]. For finding different local energy variation Harr wavelet decomposition method is used. Binary image is acquired after the threshold by local energy variation. By filtering geometric attributes as size and ratio the variation is done. Text regions which are detected in several scales are merged together to form a final result.

Chun et al. [14] used the combination of FFT and neural network for reducing the processing time. FFT computation has taken place for the reduction of overlapped segments of 1 × 64 pixels. The output where each segment is of 32 features. Labeling and noise elimination is done by the help of neural network output. Though author has claimed that to use their system in real time but the processing rate is not reported over there.

VI. Optical Character Recognition

It is a process where specialized software is used for converting scanned images of text or data which will be used while searching digitally. As OCR engines are used for developing and optimizing while extracting data from checks, passports, invoices, insurance and many more. But extraction process will start first if engine will get data so fast. But getting data for a machine or engine is totally dependent on human effort. Providing data frequently by a person is very less accurate as it can be noisy. Solution can be any algorithm for providing accurate data but it again depends upon the complexity and accuracy in terms of time consumption. Overtime working for a person is also not smarter idea. For processing large data sets are cost high and generally for reducing cost low quality output is the beneficial way for an engine. But getting proper data with quality high is as essential from a noisy data. So for scaling data sets efficiently and improving OCR faster there should be a combination of human intelligence and technology.

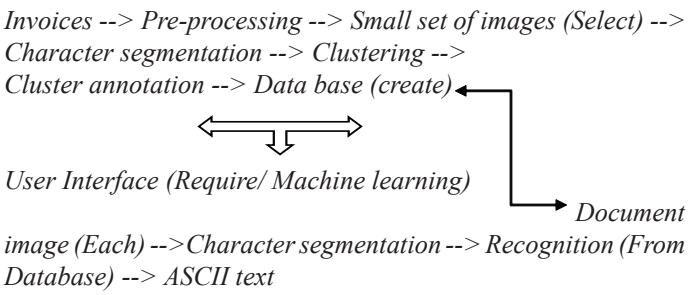

Fig. 1: OCR Method

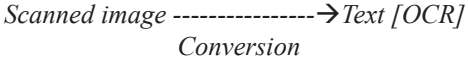

Fig. 2: OCR

Generally after a page scan it is get stored in TIF format (Via bit mapped file). After that while disposure of the image occurs in the screen, we able read. But computer has no eyes like us, it only can identify white and black dots. As no reorganization tool is not directly attached with a computer so it can't identify not a single word from the image. Here the role of OCR comes which attempts to identify the dots (Black and white). OCR has the ability for creating text version from scanned documents. After creating any text file it can be located in any of the pages with given set of words.

For finding any name in the document it has to be solved in two ways if we will try it manually first there will be so many times which will expended for the task, one will review all the documents and pay for the timing. But as error is absolutely ensure in human being so this method can't be applied, as sometimes pages can be skipped or typed twice. So to over -come from such problem OCR will be used. Whole process will be more accurate as it scans all the pages available in the documents. So OCR module will process all the scanned module. There will available text version of each pages. So now if now anyone searched for name in the document then computer will start searching each page to identify. Searching may take time but here human effort is only required just for the start, once the process is started then it will go on. Here no one is going to pay for a search which must have to do if a person will do that. So once OCR process get started then it will collect and assemble every selected or reviewed pages from each document.

OCR is a great tool but not perfect. There are certain limitations. If original document are clean and laser printed still OCR will read up to 97% correct words. But in case of special characters or images or handwriting, it may fail to read over it. If original document will be photocopies or fax or even printed by a dot matrix printer then the readability will be drop down. Very less chance will be there for OCR. The lines and boxes confuse OCR as it tries to read lines as a part of the text. OCR processing can be done in two ways i.e. Layout analysis (Finding elements in

text like line, table or character) and Character recognition. These steps are interlinked. In general recognition of object has been taken place, verification of lines or dots then relations in between them. At the end searching process will take place.

VII. Working of OCR

OCR technology is a replacement over rewriting manually of printed document to electronic form. It is also identify font type, size, formatting of paragraph, and graphic elements like diagrams, chart or any images. For a single A4 page recognition time can vary from a minute to seconds also depending on the hardware and software configuration.

There are several steps has been taken to know the exact steps which OCR performs.

Step 1: Identify the direction of any text. As scanned image can't be aligned in accurate so the direction of text can't be fully horizontal. So by just adjusting the scanned image, the line can be aligned perfectly. So it is the initial work to be done.

Step 2: It includes to know about the dimension of text as it is in single column or in two dimension.

Step 3: Here the position of baseline is decided in every subsequent text lines in each column. Due to baseline allotment the 2D problem will be reduced into single dimension.

Step 4: Make token of each line into single characters by notifying vertical stripes of white pixels. As each token is a rectangular small image of white, black or grey pixels, add one space in average white spaces.

Step 5: Run by tokens and compare it with the unknown characters (letter, numbers or punctuation).

VIII. Drawbacks of OCR

Khodami et al. [15] has done script identification using curvature space feature where they tried to retrieve text from bilingual documents. Their method was to identify scripts in at the level of character and generalize it into word, line or pages. But the problem was common small symbols are not removed in the stage of preprocessing, disconnection in Farsi fonts tends to error.

According to Singh et al. [16] in script recognizer in separating text the used methods was morphological processing of segregating horizontal or vertical stroke. But the threshold value is dependent for the classification and only valid for the machine printed documents.

A. Benefits of OCR

1. **Fast Search:** The OCR software provides the fast retrieval of data. The time which will be used by an employee that will be saved by the help of this software.
2. **Cost efficient:** It reduces the time from an employee as well as helps any organization from cost of hiring any employee for the extraction purposes. Reduces different costs like copying, shipping or printing etc.
3. **Error Rate reduction:** In comparison to any human engagement it's better to get high accuracy by using a software.
4. **Storage space:** As organization need paperless approach without any data loss so the expenses of file cabinets are got saved with OCR.
5. **Efficient Management:** As management is automated which creates an effortless management.
6. **Security:** As documents are scanned and stored in digitally so the threat is less as further access is limited.
7. **Processing:** It provides all the ways to search any document for names, references, numbers or addresses.
8. **It converts** documents which can be edited. If contents need to be changed by the time it can be changed by the help of OCR.
9. **It allows to** copy and paste from any document to other file format.
10. **Less paper** centric.

IX. Conclusion

Information extraction is a process of extracting data but having different types. Extraction can be possible by knowing the geometrical orientation, process like text information extraction, text detection and so on. Due to the above study we came to know that before any digital tool being used for recognition process other processes should be involved as explained in section III. Here we got an ideation of information extraction and the efficient tool along with benefits.

References

[1] H. K. Kim,"Efficient Automatic Text Location Method and Content-Based Indexing and Structuring of Video Database", Journal of Visual Communication and Image Representation 7 (4), pp. 336-344, 1996.

[2] M. Stricker, M. Orengo,"Similarity of color images, in Proc. SPIE Storage and Retrieval for Image and Video Databases", 1995.

[3] P. Nagabhushan, S. Nirmala,"Text Extraction in Complex Color Document Images For Enhanced Readabi", 2009.

[4] Lity,"Intelligent Information Management", pp. 120-133.

[5] Sachin, Grover, Kushal Arora, Suman K. Mitra,"Text Extraction From Document Images Using Edge Information", IEEE India Council Conference, 2009.

[6] J. Fabrizio, M. Cord, B. Marcotegui,"Text Extraction From Street Level Images", CMRT, Vol. 38, Part 3/W4, pp. 199–204, 2009.

[7] Wonder Alexandre Luz Alves, Ronaldo Fumio Hashimoto, "Text Regions Extracted From Scene Images By Ultimate Attribute Opening and Decision Tree Classification", Proceedings of the 23rd Sibgrapi Conference on Graphics, Patterns And Images, 2010.

[8] Keechul Jung, Eun Yi Ki,"Automatic Text Extraction For Content-Based Image Indexing", Proceedings of PAKDD, pp. 497-507, 2004.

[9] Miriam Leon, Veronica Vilaplana, Antoni Gasull, Ferran Marques,"Region-Based Caption Text Extraction", 11th International Workshop on Image Analysis For Multimedia Interactive Services (Wiamis), 2010.

[10] M.A. Smith, T. Kanade,"Video Skimming for Quick Browsing Based on Audio and Image Characterization", Technical Report CMU-CS-95-186, Carnegie Mellon University, July 1995.

[11] D. Chen, K. Shearer, H. Bourlard,"Text Enhancement with Asymmetric Filter for Video OCR", Proc. of International Conference on Image Analysis and Processing, 2001, pp. 192-197.

[12] B. Sin, S. Kim, B. Cho,"Locating Characters in Scene Images using Frequency Features", Proc. of International Conference on Pattern Recognition, Vol. 3, pp. 489-492, 2002.
[13] W. Mao, F. Chung, K. Lanm, W. Siu,"Hybrid Chinese/ English Text Detection in Images and Video Frames", Proc. of International Conference on Pattern Recognition, Vol. 3, pp. 1015-1018, 2002.
[14] B. T. Chun, Y. Bae, T. Y. Kim,"Automatic Text Extraction in Digital Videos using FFT and Neural Network", Proc. of IEEE International Fuzzy Systems Conference, Vol. 2, pp. 1112 –1115, 1999.
[15] M Khoddami, A. Behard,"Rarsi and Latin script Identification using Curvature scale space features", 10th IEEE symposium on neural network application in electrical engineering, pp. 213-217, 2010.
[16] S. singh, A. Kumar, D. K Shaw, D. Ghosh,"Script separation in Machine printed Bilingual Documents using morphological approach", 20th IEEE National conference on communications, pp. 1-5, 2014.

Swayanshu Shanti Pragnya : She is in final year of Engineering student of Computer science branch in CUTM. She is highly ebullient for technical research related to clustering and writing papers. Always chasing to apply mathematics in live problems and trying to solve by its application in terms of Fourier series, transformer, derivative and logarithms. Being a research intern in every internship she has learnt different techniques of analysis, survey, gap finding, literature review, content development, application of algorithms, analysis of live data in R and proper documentation. She has published 3 papers in the journals like IJSRD and IJERT in 2016 and 17 respectively.