Review Literature: Query based Information Extraction

Vishmita Shetty[1] Nikhil Polekar[2] Sandipan Das[3] Prof. Suvarna Pansambal[4]
[1,2,3]Student [4]Assistant Professor
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Atharva College of Engineering, Malad (W), India

Abstract— This paper presents a review of various techniques available in text mining for extracting keywords and key phrase. And also Text data present in multimedia that can contain useful information for automatic annotation and indexing. Keywords and key phrase are the essential part to search the query based on text. In this paper, firstly, it is discussed different techniques for text extraction from images, text data and videos. Secondly, reviews the techniques for indexing and retrieval of data, videos and images by using extracted text. It also discusses some important feature selection metrics which are generally utilized by researchers to rank candidate keywords and key phrases in accordance to their importance.

Key words: Query based Information Extraction, Detection, Localization, Tracking, Extraction

I. INTRODUCTION

Nowadays there is a need to quickly go through large amounts of textual or visual information to find out content of interest and this document space is growing very fast on a daily basis. Huge quantities of data can be analysed easily if important subset of words (Keywords) which gives an idea of main features, concept and theme etc. of the document can be retrieved. Suitable keywords can help summarise a document and thus help in organizing documents and retrieve the documents based on their content. As keywords represent the core of a text, they can be used as a measure of similarity for text clustering. Both single words (keywords) and phrases (key phrases) may be simply termed as “Key” terms. Keyword extraction is a significant task in the field of text mining. Key extraction can be carried out using varied approaches, such as unsupervised and supervised machine learning, statistical methods and other linguistic ones. Linguistic based approaches are generally rule based and are derived from the Linguistic knowledge/features and require field knowledge in addition to language expertise. The linguistic approach includes the lexical analysis, discourse analysis syntactic analysis etc. Statistical approaches are generally based on linguistic corpus as well as statistical feature which are derived from the corpus. This approach is independent of the language on which they are applied. Same technique can be used on multiple languages due to this, but this method is not accurate as linguistic however large amount of datasets has made performance possible. Machine Learning approaches usually employ supervised learning methods, in which keywords are extracted from training documents to learn a model. A testing module is used to test the model. After a satisfactory model is built it is used to find keywords from fresh documents. Naive Bayes, Support Vector Machine, etc are used in this technique. Keyword/key phrase extraction is more efficient than assignment [1]. KEA (Key phrase extraction algorithm) was developed by Frank et al [2]. In this system a classifier is build based on the Bayes theorem from training documents, and then it is used to extract key phrases from fresh documents. KEA analyzes the input document on orthographic boundaries e.g. punctuation marks, newlines etc. to find candidate phrases [1].

II. PROCESS

Text data present in multimedia viz. images and videos contain useful information for automatic annotation, indexing. The Process of Extraction of information involves following: -

- Detection
- Localization
- Tracking
- Extraction
- Recognition, and
- Enhancement of the text from a given image [3].

However there are few issues that make the text extraction process difficult and time consuming

- Differences in text in style
- Orientation
- Size
- Alignment
- Low contrast image
- Complex background [4].

A variety of approaches to text extraction from images and video have been presented for many applications like address block location [5], content-based image/video indexing [6, 10], page segmentation [8, 9], and license plate location [7, 10]. In spite of such in critical studies, it is still not easy to design an all-purpose Text Extraction system.

Text in video images can be classified as caption text or scene text. The difference is that caption text is artificially overlaid on the image whereas the scene text exists naturally in the images. Term ‘graphics text’ is generally used in place of scene text, and “superimposed text” or “artificial text” for caption text. It is documented that scene text is harder to detect [4].The process for extraction caries detection, localization, tracking, extraction, enhancement, and recognition of the text from a given image [3]. Extracted text components are required enhancement because the text region usually has low-resolution and is susceptible to noise. The extracted text images are converted into plain text using OCR technology [4]. Text-frame choice is performed at linked interval of two seconds for caption text within the detected scene frames. Video extraction can be done easily and economical resolution for video indexing applications that only needs keywords from video clips, instead of the whole text [4].

Text localization techniques can be further classified into two types based on utilization of features viz. region-based and texture-based. Region-based techniques use the variations or the properties of the color/gray scale in a text region with the corresponding properties of the background. These techniques can be additionally divided into up to two sub-approaches, namely connected component (CC) and edge-based [4].

III. Methods

All text parts using spatial arrangement will require geometrical analysis whereas on separate non-text parts the limits of the text regions must be marked [4] CC-based methods are widely used mainly due to their relatively simple implementation. Most of Connected Component based methods have four processing stages namely:

- Pre-processing, like color clustering and noise reduction
- CC generation
- Filtering out non-text components and
- Component grouping [11].

Further, the performance of a CC-based method is affected by component grouping, like a projection profile study or text line selection. Additionally, many threshold values are required to filter out the non-text components, and these threshold values are depends on the image/video database

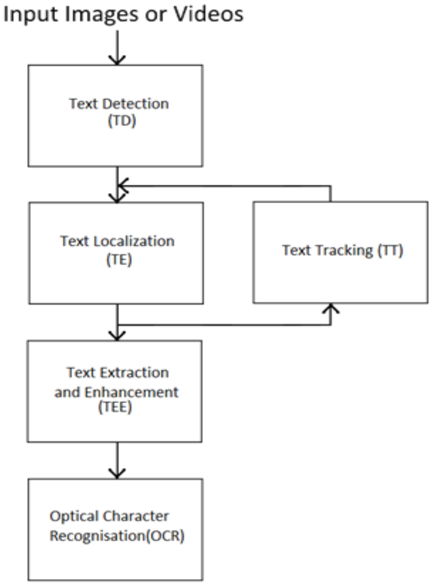


Fig. 1: Text Extraction from Images or Video

IV. Text based query extraction algorithm's for image/video/data

A. Natural Language Processing Based Algorithm:

1) Understanding a statement:

A statement is made up of many grammatical components. These components make up meaning for the statement. A person can express something in various statement formations. Though the make of statements differ the meaning retains.

2) Simplification:

Simplification of statement can involve replacing phrases by words having same meaning. Similarly it is necessary to avoid compound words. A statement is generally converted in CNF for processing

3) Processing:

Once a simplified statement is derived the statement can be broken into subject and predicate. The subject (generally) serves as an important phrase of any statement. The statement is then broken into grammatical components such as nouns, verbs, adjectives with same order of importance. The nouns can be further classified as proper nouns, common nouns etc. Proper nouns mainly serve as keywords of a statement.

Once textual query is received, it needs to be processed to obtain words having higher importance (Keywords). A popular method used for this is KPE (Key Phrase Extraction).

B. Key Phrase Extraction (KPE) Algorithms

This algorithm finds phrases or words that represent keywords in document. Key phrases can be used in several applications like categorizing documents, clustering and summarization. KPE can process documents to produce words that represent the document [12]. These words that serve keywords produced by KPE can be used in combination with string matching algorithms (such as Knutt Morris Pratt) and map the keywords to all possible results by applying suitable pattern matching algorithm.

Steps involved in KPE are:

1) Indexing Keyphrases

- construct decision tree induction algorithm
- Apply GenX algorithm

GenX is a combination of the Genitor steady-state genetic algorithm and the Extractor parameterized KEA. Extractor works by assigning a numerical score to the phrases in the input document. The final output of Extractor is necessarily a list of the highest scoring phrases. The behaviour of the scoring function is resolute by a dozen numerical instructions. Genitor tunes the setting of these parameters, to make best use of the performance of Extractor on a given set of training examples. Key phrase extraction algorithms works by the number of matches between the human-generated phrases and the machine-generated phrases. A machine-generated key phrase matches a human-generated key phrase when they correspond to the same sequence of stems. A stem is what remains when we remove the suffix from a word. By this definition, “computer systems” matches “computer system”, but it does not match “system”. The order in the sequence is important, so “snow skiing” does not match “skiing snow”.

Text extraction from videos/images in two ways:

- Caption text this text is added/superimposed at the time of editing the images and this is the powerful source of finding the text in the images/videos and use text for indexing, retrieval and summarization of information[13].
- And another way is scene text this text is present naturally in the scene shoted by images and videos used to extract information. There are lots of problem extracting scene based text some are resolution, brightness, blurring effects, complex background etc .Because of this caption based technique for images and videos search is the best option for searching information based on text extraction[13].

Algorithm's used in text retrieval from images and videos

C. K Clustering Algorithm

In this the image is broken in to connected blocks and non text block is filtered using component filter and text containing block is again merged using K clustering algorithm.

1) Automatic Text Location and Identification

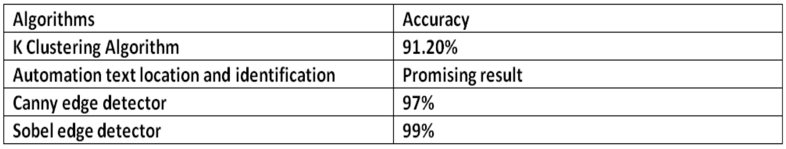
The number of colors was condensed by applying a clustering algorithm. Texts are located using a top-down analysis based on consecutive splitting in horizontal and vertical direction. A bottom-up analysis detected all the same regions using a region growing method; grouping steps applied to the partial output and find subsets of region. Finally text regions and non-text regions were identified.

2) Canny edge detector

When this method is applied to the images it created holes in most of the connected component that corresponds to strings. Connected components without holes were removed. Other non-text components were removed by computing and analyzing the standard deviation of each connected component. An unsupervised local thresholding was devised to carry out fore-ground segmentation in detected text regions. Finally the noisy text regions were recognized and reprocessed to further improve the quality of retrieved foreground.

3) Sobel edge detector

In this approach text was embedded in complex colored document images. They designed a simple edge based feature to perform this task. The image was transformed to gray scale by forming a weighted sum of the R, G, and B components. Then edge detection method was applied on the gray-scale image by convolving the image with Sobel masks, separately for horizontal and vertical edges. Convolution was followed by elimination of non-maxima and thresholding of weak edges. Next, the edge image was divided into small non overlapping blocks of m x m pixels, where m depends on the image resolution. They performed block classification using pre-defined threshold which would differentiate the text from the image.



Algorithms	Accuracy
K Clustering Algorithm	91.20%
Automation text location and identification	Promising result
Canny edge detector	97%
Sobel edge detector	99%

Fig. 2: Accuracy of text extraction algorithm

D. Scene Based Text Extraction Algorithm:

1) Blob Extraction Technique

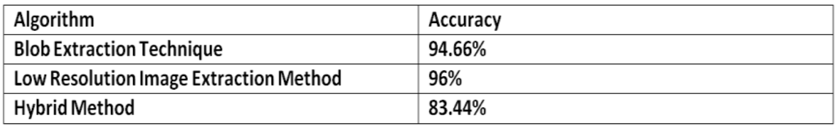
In this approach to detect and extract text from commercial screenshot images. Used implemented edge-based method and connected component labelling method known as blob extraction method. Combination of similar edge detection filter and appropriate threshold number separated the text from the image.

2) Low Resolution Image Extraction Method

In this method detect and extract text regions from low resolution natural scene images. Their proposed work used Discrete Cosine Transform filter to remove and suppress the constant background. The texture characteristic matrix was applied on every 50x50 block of the processed image than discriminate function was used to classify text blocks. The detected text blocks were combined to obtain new text regions. Finally, post processing step used to improve the detection accuracy. This phase used to cover small portions of missed text present in contiguous undetected blocks and unprocessed regions.

3) Hybrid Method

Hybrid method where in a text region detector was planned to generate a text confidence map. A Local binarization technique was used to segment the text components using text confidence map. A Conditional Random Field representation was used to label components as text or non-text which was solved by minimum classification error learning and graph cuts inference algorithm. A learning based method by building adjacent components into minimum spanning tree and cutting off interline edge with an energy minimization representation to group the text components into text lines.



Algorithm	Accuracy
Blob Extraction Technique	94.66%
Low Resolution Image Extraction Method	96%
Hybrid Method	83.44%

Fig. 3: Accuracy of scene based text extraction algorithm

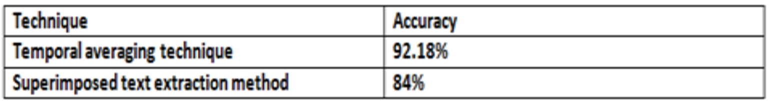
E. Caption Based Text Extraction Algorithms

1) Temporal Averaging Technique

In this technique extract the caption text from various videos. Iteratively temporal averaging approach technique was used in caption extraction process. To improve the image quality and to lessen noise spatial-image analysis was performed. Threshold value was calculated using binarization process based on the global mean and the standard deviation of the gray level of the averaged video image. Binarization technique may lead to holes and disconnectivity on video captions with blurred background. This problem is fixed using morphological processing. Each connected component was used to extract geometrical features to spot the captions. A model-based segmentation approach was applied to accurately extract the caption contents.

2) Superimposed Text Extraction Method

In this technique Key frames from the video were extracted using Colour Histogram technique to minimize the number of video frames and transformed to gray images. Text image portion in the image were cropped. Canny Edge Detection algorithm used to detect edges on the cropped image. From this edge detected images, text region was identified and fed to an Optical Character Recognition system which produces index-able keywords.



Technique	Accuracy
Temporal averaging technique	92.18%
Superimposed text extraction method	84%

Fig. 4: Accuracy of caption based text extraction algorithm

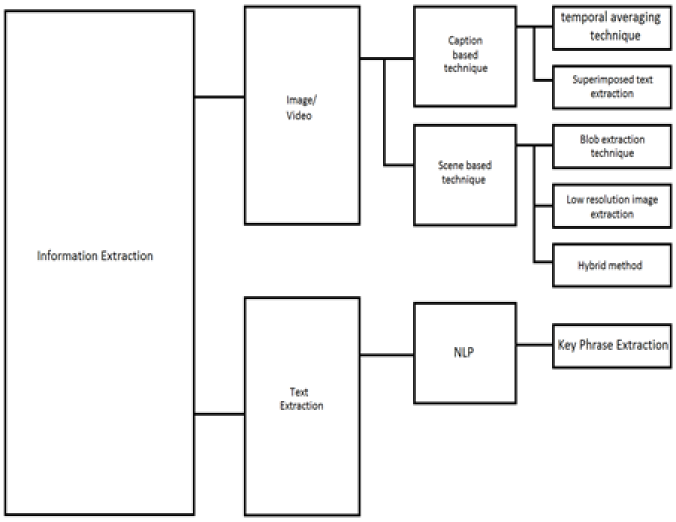


Fig. 5: Information Extraction Technique

V. Conclusion

In this paper a comprehensive literature review of text extraction in media such as images and video as well as text based information extraction from database. The different information Sources such as Color, Texture, Motion, Shape, Geometry, etc are used for text recognition. By merging the different sources of information it is possible enhance the performance of a text extraction system and text based video retrieval systems.

For images/videos retrieval the best method is segmenting the images and videos in blocks and then passing it to OCR and non text block gets filtered and the text blocks get clustered using K Clustering algorithms.

Acknowledgment

Our sincere appreciation to our guide Prof. Suvarna Pansambal for guiding us to achieve a substantial level in our project. We also thank our college for providing us an environment which help us to undertake this project. We express our gratitude for everyone who wholeheartedly supported the project and gave us their valuable time while making this project. All the inputs given by them have found a place in the project. The technical guidance provided by our guide was more than useful and made the project successful. She has always been a source of inspiration for us. It was memorable experience learning under such a highly innovative, enthusiastic and hardworking teacher.

We are also thankful to our Principal Dr. S.P. Kallurkar, our Project co-ordinator Prof. Deepali Maste and all the staff members of the Computers department who have provided us various facilities and guided us to develop a very good project idea.

Finally, we would also like to thank teachers of our college and friends who guided and helped us while working on the project.

References

[1] Sifatullah Siddiqi School of Computer and Systems, Sciences Aditi Sharan School of Computer and Systems Sciences Keyword and Keyphrase Extraction Techniques: A Literature Review International Journal of Computer Applications (0975 – 8887) Volume 109 – No. 2, January 2015

[2] Frank E., Paynter G.W., Witten I.H., Gutwin C., Nevil lManning C.G., ‖ Domain-specific keyphrase extraction‖, Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, pp. 668-673. San Francisco, CA, USA, 1999.

[3] Jung, Keechul, Kwang In Kim, and Anil K Jain. "Text information extraction in images and video: a survey." Pattern recognition 37, no. 5 (2004): 977-997.

[4] Avinash N Bhute and B.B. Meshram VJTI, Matunga, Mumbai-19, Text Based Approach For Indexing And Retrieval Of Image And Video: A Review, Advances in Vision Computing: An International Journal (AVC) Vol.1, No.1, March 2014.

[5] B. Yu, A. K. Jain, and M. Mohiuddin, Address Block Location on Complex Mail Pieces, Proc. of International Conference on Document Analysis and Recognition, 1997, pp. 897-901.

[6] H. J. Zhang, Y. Gong, S. W. Smoliar, and S. Y. Tan, Automatic Parsing of News Video, Proc. of IEEE Conference on Multimedia Computing and Systems, 1994, pp. 45-54. Advances in Vision Computing: An International Journal (AVC) Vol.1, No.1, March 2014 36

[7] Y. Cui and Q. Huang, Character Extraction of License Plates from Video, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 502 –507.

[8] A. K. Jain, and Y. Zhong, Page Segmentation using Texture Analysis, Pattern Recognition, 29 (5) (1996) 743-770.

[9] Y. Y. Tang, S. W. Lee, and C. Y. Suen, Automatic Document Processing: A Survey, Pattern Recognition, 29 (12) (1996) 1931-1952.

[10] D. S. Kim and S. I. Chien, Automatic Car License Plate Extraction using Modified Generalized Symmetry Transform and Image Warping, Proc. of International Symposium on Industrial Electronics, 2001, Vol. 3, pp. 2022-2027.

[11] J. C. Shim, C. Dorai, and R. Bolle, Automatic Text Extraction from Video for Content-based Annotation and Retrieval, Proc. of International Conference on Pattern Recognition, Vol. 1, 1998, pp. 618-620.

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

[13] Manisha Verma, Vasudeva Varma, Applying Key Phrase Extraction to aid Invalidity Search, International Conference on Artificial Intelligence & Law.

[14] C.P. Sumathi1, T. Santhanam2 and G.Gayathri Devi3,A SURVEY ON VARIOUS APPROACHES OF TEXT EXTRACTION IN IMAGES, International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.4, August 2012.