Image Retrieval and Classification using Image Text Extraction

Ms. Pratibha B. Raut[1], Mr. Vikas. R. Marathe[2]
N.B. Navale College of Engineering, Solapur (413304), Maharashtra, India
(Email: - pratibharaut6@gmail.com)

Abstract– The rapid growth of smart phones and online social media have led to the accumulation of large amounts of visual data, in particular, the massive and increasing collection of video on the Internet and social networks. For example, YouTube sends approximately 100 hours of video per minute worldwide in 2014. These infinite videos have triggered research activities in multimedia understanding and video retrieval. Text is a direct source of information in video. Mainly text in images can reflect semantics of images. For images database of mined image; text can add extra metadata which will be used for classification and image retrieval. In this paper business image database, text on images will specify business places of drinks, kid's zone, food and other business places. The aim of this proposed research is to present effective image retrieval mechanism and image classification system. Text extraction from images can involves detection of text, extraction of text, enhancement, and recognition of the text from a given image. The research can be carried out in Detection of text, Extraction of text, Text recognition, Image classification and Image retrieval. Images with text of different fonts, languages, different font size, different styles, alignment, orientation and blurred images can make the problem of automatic text extraction extremely challenging.

Keywords-Smart Phones, Metadata, Text detection, Text extraction, Text Recognition, Image retrieval and Image classification.

I. INTRODUCTION

The multimedia data on web is increasing exponentially due to large development of smart phones and cameras. Everyone wants to share experiences hence, images or video is uploaded on web. Due to this there is need of efficient image retrieval system. The description of image used while storing image can be used for image retrieval system. While existing methods gives irrelevant images as a result. Text extraction methods can play important role in image retrieval system. The extracted text can be used as additional metadata for searching images.

Secondly for the existing business image database, the extracted text data can be used for classification of images. Different business places often have subtle visual differences. As an example, in (a), the main difference between bakery images and pizzeria images is that bakery shop windows have images of bread whereas pizzeria shop windows have pizza images. The information carried by scene text helps to distinguish these two types of business places. (b)Shows an extremely challenging case where the two shops can hardly be distinguished unless the scene text is used.

The proposed work focuses on classification of different business places, e.g., bakery, cafe and bookstore. Various business places have fine differences in visual appearances. For instance, on the shop windows, a pizzeria often has pizza images whereas a bakery usually shows images of bread, see Figure 1.a. In this particular problem, we make use of the domain specific knowledge of business places. We exploit the recognized text in images for fine-grained classification of business places.

Fig. 1: Example images of bakery (left column) and pizzeria (right column).

Automatic recognition and indexing of business places will be useful in many practical scenarios. For occurrence, it can be used to extract information from Google street view images and Google Map can use the information to provide

Generally, the stores use text to show what type of food (pizzeria, diner), drink (tea, coffee) and service (dry cleaning, repair) that they provide. This text information is beneficial for human observers to understand what type of business place it is. For instance, in Figure 1.b, the images of two different business places (pizzeria and bakery) have a very similar appearance. However, they are different types of business places. It is only the way to use text information to identify what type of business places these are. Moreover, text is also useful to identify similar products (logo) such as Heineken, Foster and Carlsberg. Therefore, a multimodal approach which uses recognized text and visual cues for fine-grained classification and logo retrieval is proposed.

Image text or video text can be categorized into scene text and image text. Caption text is also called graphic text or artificial text . Caption text provides excellent directivity and a better overview of the semantic information in captions, subtitles and annotations of the video, while scene text is part of the camera images and is embedded within objects (e.g., trademarks, signboards and buildings) in scenes.

II. PROBLEM STATEMENT

- To develop system for image classification and image retrieval using text data extraction from image.
- Classification of images using fine grained classification
- Efficient image retrieval using textual cues.

III. LITERATURE REVIEW

Boris Epshtein, Eyal Ofek [3] has presented a novel image operator that seeks to find the value of stroke width for each image pixel, and demonstrate its use on the task of text detection in natural images. The specified operator is local and data dependent, which makes it fast and error free enough to eliminate the need for multi-scale computation or scanning windows. Extensive testing shows that the demonstrated scheme getover the latest published algorithms. Its simpleness allows the algorithm to detect texts in many fonts and languages.

Max Jaderberg, Karen Simonyan [4] presented a framework for the recognition of natural scene text. The presented framework does not require any human-labeled data, and performs word recognition on the whole image holistically, departing from the character based recognition systems of the past. The deep neural network models at the centre of this framework are trained on data produced by a synthetic text generation engine – synthetic data that is highly practical and subsequent to replace real data, giving us infinite amounts of training data.

Anand Mishra, Karteek Alahari, and CV Jawahar. [5] propose to use textual cues for query-by- text image retrieval. Given a query text, the method assigns scores to images based on the presence of the query characters. Additional pair wise spatial constraints between characters are used to refine the ranking.

SezerKaraogluy, Ran Taoy, Theo Gevers and Arnold W.M [1, 6] propose to use textual cues in combination with visual cues for fine-grained classification. Bi-grams are computed based on recognized characters in images. These bi-grams are used to encode the textual cues. In contrast, this work performs a word-level textual cue encoding. Moreover, the proposed method aims at high recall word detection which leads to combine state-of-the-art text detectors performed in various color spaces.

Kaveri Pawar, Prof. Priyadarshini. C. Patil [7] worked on restaurant images by which text specifies business places which serve variety of food (e.g. restaurants) or drinks, snacks (e.g. cafeteria, teahouse, bakery and hotel) and what kind of service is provided (e.g. home delivery). Scene text detection and recognition have become popular research area in computer vision. No prior guessing is made regarding the text size, font, language and character set. So, here they have collected some of restaurant, bakery and cafe images which can give a new application to present with the best output and with minimum number of images with mainly use of the mean feature. This concept makes use of textual contents in images for fine-grained classification of business places and logo retrieval. We assessed our system using MSER and Text Saliency map. Merging the proposed textual and visual cues outperforms the visual classification and retrieval by a large margin. The experimental results accuracy 90.35% which indicates the good performance of the proposed method

Tianjun Xiao,Yichong Xu [8] proposed to apply visual attention to fine grained classification task using deep neural network. The pipeline integrates three types of attention: the bottom-up attention that propose candidate patches, the object-level top-down attention that selects relevant patches to a certain object, and the part-level top-down attention that localizes discriminative parts. These attentions were combined to train domain-specific deep nets, and then use it to improve both the aspects. Importantly, expensive annotations like bounding box or part information from end-to-end were avoided. The poor supervision constraint makes our work simpler to generalize. The pipeline delivered significant improvements and achieved the best accuracy under the weakest super vision condition. The performance is competitive against other methods that rely on additional observations.

DISADVANTAGES OF EXISTING SYSTEM -

- Unfortunately, there exists no single best method for detecting words with high recall due to large variations in text style, size and orientation.
- Weak classifiers are used.
- In previous Existing System recall or retrieval value is less as compared to total database size. F- Score value depends on recall and precision

IV. METHODOLOGY

MODULES:

A. Word Level Textual Cue Encoding
B. Visual Cue Encoding
C. Classification and retrieval

A. Word Level Textual Cue Encoding: It has following steps,

a. Image Acquisition
b. Color Channel Generation
c. Character Detection
d. Word Proposal Generation & Recognition

a. Image Acquisition:

Images are acquired from Gallery.

b. Color Channel generation:

In this stage, RGB image is converted into HSV image. After that shade, Saturation and Intensity channels are extracted for further process. Especially intensity channel is used for character detection.

c. Character Detection:

For character detection, two methods are proposed such as MSER region detection and text saliency generation. V channel is used for MSER region detection. In that text region is not detected properly. Other method is saliency map generation for text detection. Finally text saliency was extracted.

d. Word Proposal Generation & Recognition:

Word detection and recognition done by using grammatical operation and optical character recognition method. It has to following steps,

Stage 1:

In this stage text saliency, image capturing and word region detection are performed. Initially, a text saliency image is taken as input. The image taken is in the RGB format, and then this image is transformed to gray scale image.

Stage 2:

In this stage text extraction and character segmentation is performed. For that morphological dilation and erosion operations are performed to fill holes. After morphological operations, local thresholding is applied to transform gray image into binary image. In order to get further contrast enhancement, intensity range of the pixel values should be scaled between 0 and 1. In some situations if unwanted gaps and holes are present in the word region, then region growing segmentation is performed to segment characters from the word region.

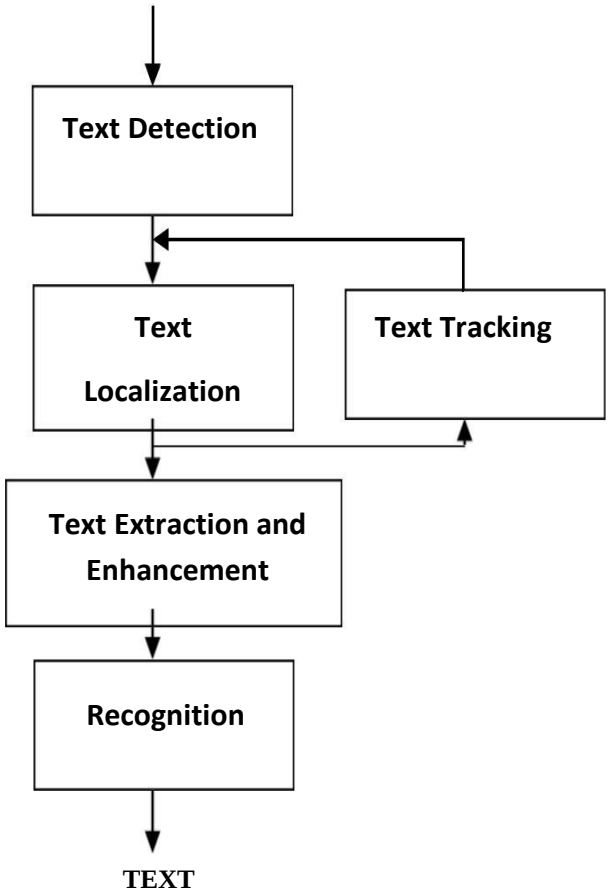

Fig. 2: System Architecture

Stage 3:

Here word recognition is done using template matching. Each segmented character is compared with character templates stored in database. Finally the word was recognized.

B. Visual cue Encoding:

This is implemented for visual features extraction. SURF feature descriptor is used for visual features derivation, strongest key points are derived.

C. Classification and Retrieval:

The classification process is done over the identified word and visual features. Based on the identified word and features, classification and similar images retrieval are explored.

V. DESIGN AND IMPLEMENTATION

Fine grained classification is the problem of assigning images to classes where instances from different classes differ slightly in the appearances e.g., flower types, bird and dog species, and models of a product. In contrast to coarse object category recognition e.g., cars, cats and airplanes, low-level visual cues are often not sufficient to make distinction between fine-grained classes. Even for human observers, fine-grained classification tasks usually require expert and domain specific knowledge.

The common approach to text recognition in images is to detect text first before they can be recognized. The state-of-the-art word detection methods focus on obtaining a high f-score by balancing precision and recall. However, instead of using the f-score, the aim is obtain a high recall. A high recall is required because textual cues that are not detected will not be considered in the next phase of the framework. However, there does exists a single best method for detecting words with high recall due to large variations in text style, size and orientation.

The proposed method uses color spaces containing photometric invariant properties such as robustness against shadows, highlights and specular reflections. The suggested method determine text lines and generates word box proposals based on the character candidates. Then, word box proposals are used as input of a state-of-the-art word recognition method to yield textual cues. At last textual cues are combined with visual cues for fine-grained classification and logo retrieval. The proposed framework is given in figure shown below. In this method word-level textual cues and visual cues are combined for fine-grained classification and logo retrieval.

A. Word-level Textual Cue Encoding

In order to extract the textual cues from the image, a two- step procedure is followed. In the first step, word box proposals are generated to locate the words in the image. In the second step, the word proposals are used as input to a word recognizer to form the word-level representation.

When a word in an image is not detected or localized incorrectly, it is not possible to identify it. The aim is to obtain high recall with the cost of false positives. To this end, the proposed method uses a complementary set of character detection algorithms and color invariant spaces.

B. Low computational cost:

The word box proposal method needs to be efficient especially for large scale scenarios. Further, the number of possible word box candidates should be as low as possible.

C. Generic:

A generic word proposal method is aimed in this proposed work. No need for tuning the method for different alphabets or datasets.

Therefore, an efficient and fully unsupervised bottom-up approach is proposed. First, characters are detected by a text-independent approach. Then, these detected characters are filtered based on geometric and appearance properties. Finally, they are grouped to generate word box proposals.

D. Character Detection:

There exists no single character detection algorithm that is robust against all variations in text style, location and orientation and imaging conditions. Therefore, following methods are proposed to compute character candidates using two methods with different strengths, i.e., text saliency method and Maximally Stable Extremely Regions (MSERs).

A text saliency map is computed using scene background. It is assumed that background pixels are uniformly colored e.g., windows, boards, roads, buildings, fences etc. and that they contrast with text regions. Accordingly, the method uses background homogeneity to form connectivity between background pixels. The method selects initial background seeds and grows these seeds iteratively until all background pixels are covered. Assuming that text regions have strong contrast with the background, text regions will remain uncovered by the region growing algorithm. Finally, the background image is subtracted from the original image to obtain a text saliency map, which is further binarized using to obtain character candidates.

Text saliency computation does not require any tuning for varying text size, style and orientation, and is robust to image noise. However, due to the information loss caused by the image boundary priors and the binarization, the method may miss characters. To compensate for this, MSER as another character detection algorithm is used. MSERs define an extremely region as a connected component of which image values remain stable within the boundary and highly contrast against boundary pixels. MSER regions are widely in use for character detection. MSER is suited for character detection because text regions are usually designed to have uniform appearance. Further, they usually have high contrast with their surroundings. However, MSER has certain shortcomings for character detection such as detecting characters in blurry and noisy images. Moreover, MSER is sensitive to character sizes due to the parameters used to define stable regions. In fact, the MSER and text saliency results are, to a certain extent, complementary. Below figure illustrates complementary properties of MSER and saliency methods.

E. Complementary Color Spaces

Images are captured under uncontrolled illumination conditions. Therefore, text regions may be influenced by different photometric changes such as shadows and specular reflections. A uniformly colored character may vary in intensity due to shadows or highlights. Hence, these shadows or highlights may negatively influence the pixel connectivity for a uniformly colored character. To compensate for this, the proposed method computes the character candidates using a variety of color spaces containing a range of invariant properties. The two channels, (O1, O2), from the opponent color space, Saturation (S) and Hue (H) from HSV, and (I) from gray scale are considered in the proposed method.

F. Word Box Proposal Generation

The next step is to compute word box proposals using character candidates. Combinations of character candidates as potential words are considered. However, it is computationally expensive if all possible combinations are considered. And, due to the nature of text, characters within a word cannot have arbitrary positions and sizes. Therefore, as the first step of computing word box proposals, text lines are generated to restrict the selection of combinations by linking character candidates based on five pair-wise constraints. In Fig. 3, the two boxes stand for two character candidates with (x1, y1), height1 and width1 are being the coordinates of the top-left corner, height and width of the box covering the first character. The box of the second character is defined likewise.

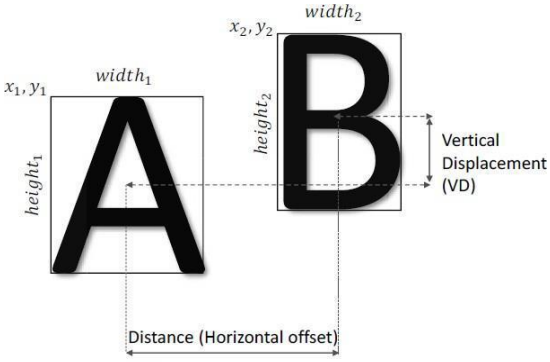

Fig. 3: An illustration on the notions of two character candidates.

G. Fine-grained Classification

Fine-grained classification is the problem of the categorization of subordinate-level categories such as bird species, flower types and building types. The small interclass visual differences and the large intra-class variations make fine-grained classification challenging. In this section, in addition to visual features, Textual cues are used from the images for fine-grained image classification.

VI.RESULT

Here for example purpose input image is Putney school of art color (rgb) image.

Fig. 4 : original image

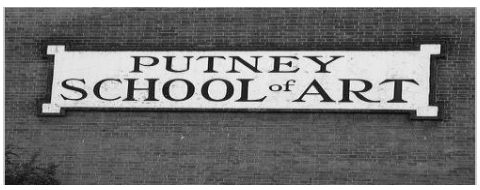

Fig. 5: Grayscale image

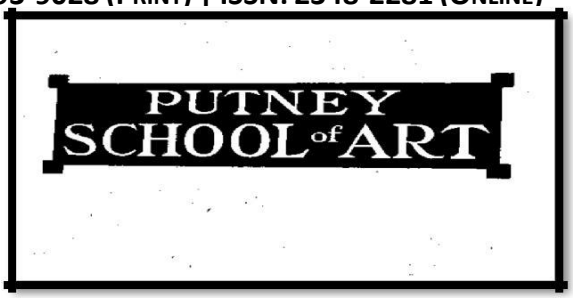

Fig. 6: binary image

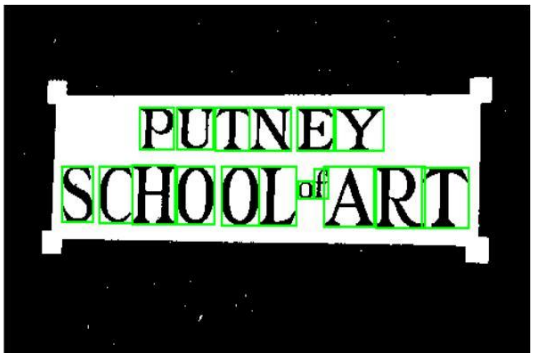

Fig. 7: bounding box detection

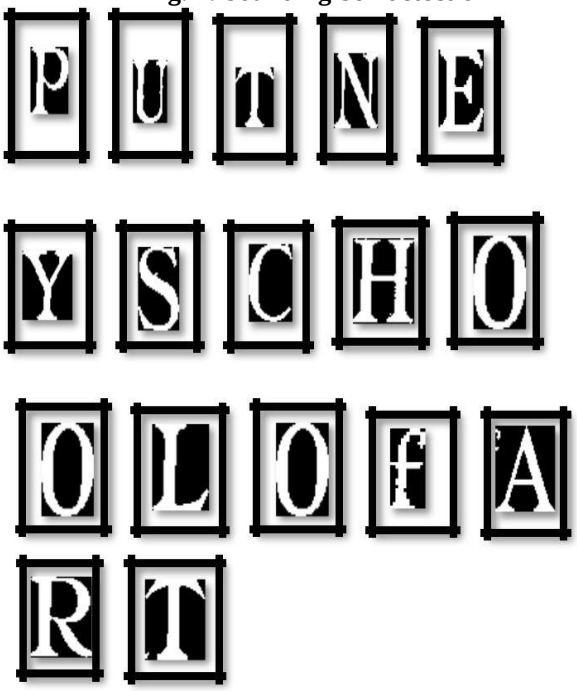

Fig. 8 :Character detection from text images

For the experimental purpose Matlab software is used. The result can be analysed using Word box proposal approach.

- **Evaluation measures:**

The performance is measured in terms of recall, number of proposals and average maximum overlap (AMO)

- **Experiments and Results**

The proposed method generates word box proposals using

different color spaces and character detection algorithms.Word box proposals are generated for each color space independently and then combined. The same candidate regions may be detected for the different color spaces or character detection algorithms. To filter out these duplicate regions, non-maximum suppression is applied.

VII.APPLICATIONS

- Image retrieval with better accuracy in web image Search.
- Criminal or missing persons detection
- Extracted text from image can be used for text to Speech conversion for visually impaired persons.
- Classification of images

VIII. CONCLUSION

This paper provides a review on extensive mechanism of text detection; tracking and recognition in images, where text tracking, tracking based detection, and tracking based recognition will be specifically summarized. It will describe the available datasets, evaluation protocols, technological applications, and grand challenges of image text extraction. The accuracy of image retrieval and classification will increase.

REFERENCES

[1] SezerKaraogluy, Ran Taoy, Theo Gevers and Arnold W.M. Smeulders, "Words Matter: Scene Text for Image Classificationand Retrieval", IEEE Transactions on Multimedia, 2017.

[2] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 37, no. 7, pp. 1480–1500, 2015.

[3] Boris Epshtein, Eyal Ofek, Yonatan Wexler, "Detecting text in natural scenes with stroke width transform" in IEEE Computer Society Conference on Computer Vision and Pattern Recognition in 2010

[4] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman, "Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition" in Computer Vision and Pattern Recognition, 2014

[5] Anand Mishra, Karteek Alahari, and CV Jawahar. Image retrieval using textual cues. In ICCV , 2013

[6] Sezer Karaoglu, Jan C van Gemert, and Theo Gevers. Con-text:Text detection using background connectivity for fine-grained object classification. In ACM MM , 2013.

[7] Kaveri Pawar, Prof. Priyadarshini. C. Patil, Dr. Vishwanath.C.Burkapall, "A Image Classification Technique for Textual Pattern Retrieval in Universal Images" in International Journal of Computer Technology & Applications,Vol 9(3),169-178 2018.

[8] Tianjun Xiao,Yichong Xu,Kuiyuan Yang, Jiaxing Zhang,Yuxin Peng, Zheng Zhang, " Computer Vision and Pattern Recognition" in 2014

[9] X. Qian, G. Liu, H. Wang, and R. Su, "Text detection, localization, and tracking in compressed video," Signal Processing: Image Communication, vol. 22, no. 9, pp. 752–768, 2007.

Ms.Pratibha Raut is pursuing her Master's degree in "Electronics & Telecommunication Engineering" from N.B.N Sinhgad College of Engineering, Solapur Maharashtra. She received her bachelor's Degree in Electronics & Telecommunication Engineering. her area of interest is Image Processing.

Mr.Vikas Marathe has completed Master's degree in Electronics Engineering. He is the Assistant Professor in N.B.N Sinhgad college of Engineering, Solapur. His area of interest is Image Processing.