Journal of Engineering Sciences

TEXT RECOGNITION AND RETRIEVAL IN NATURAL SCENE IMAGES

AMRUTHAM SRINU, K. RAMBABU

Branch Email id:-dnrpgcollege2023@gmail.com

PG STUDENT OF MCA, D.N.R. COLLEGE, P.G. COURSES (AUTONOMOUS), BHIMAVARAM-534202.

Email id:amruthamsrinu10@gmail.com

ASSISTANT PROFESSOR IN THE DEPARTMENT OF MASTER OF COMPUTER APPLICATIONS ,D.N.R. COLLEGE, BHIMAVARAM 534202.

Email id:kattaramababudnr@gmail.com

ABSTRACT

In the past few years, text in natural scene images has gained potential to be a key feature for content based retrieval. They can be extracted and used in search engines, providing relevant information about the images. Robust and efficient techniques from the document analysis and the vision community were borrowed to solve the challenge of digitizing text in such images in the wild. In this thesis, we address the common challenges towards scene text analysis by proposing novel solutions for the recognition and retrieval settings. We develop end to end pipelines which detect and recognize text, the two core challenges of scene text analysis.

For the detection task, we first study and categorize all major publications since 2000 based on their architecture. Broadening the scope of a detection method, we propose a fusion of two complementary styles of detection. The first method evaluates MSER clusters as text or non-text using an adaboost clas- sifier. The method outperforms the other publicly available implementations on standard ICDAR 2011 and MRRC datasets. The second method generates text region proposals using a CNN based text/non- text classifier with high recall. We compare the method with other object region proposal algorithms on the ICDAR datasets and analyse our results. Leveraging on the high recall of the proposals, we fuse the two detection methods to obtain a flexible detection scheme.

For the recognition task, we propose a conditional random field based framework for recognizing word images. We model the character locations as nodes and the bigram interactions as the pairwise potentials. Observing that the interaction potentials computed using the large lexicon are less effective than the small lexicon setting, we propose an iterative method, which alternates between finding the most likely solution and refining the interaction potentials. We evaluate our method on public datasets and obtain nearly 15% improvement in recognition accuracy over baseline methods on the IIIT-5K word dataset with a large lexicon containing 0.5 million words. We also propose a text query based retrieval task for word images and evaluate retrieval performance in various settings.

Finally, we present two contrasting end to end recognition frameworks for scene text analysis on scene images. The first framework consists of text segmentation and a standard printed text OCR. The text segmented image is fed to Tesseract to get word regions and labels. This case sensitive and lexicon free approach performs at par with the other successful pipelines of the decade on the ICDAR 2003 dataset. The second framework combines the CNN based region proposal

method with the CRF based recognizer with various lexicon sizes. Additionally, we also use the latter to retrieve scene images with text queries.

1. INTRODUCTION

In the last decade, we saw a surge in multimedia content generated across the world. With the arrival of digital equipments like cameras, camcorders, etc., it became feasible for a large section of the world's population to generate such content. These devices became even more popular especially after their improvement in performance, gradual decrease in prices as well as their integration into cell phones. In fact, after the increase in mobility of such devices via cellphones, coupled with the advances in cheap storage, it was possible to capture media content on the fly in the form of images and videos.

The multimedia content generated at large scale by the population provided us the opportunity to tag, categorize and make them browsable [37, 71, 95, 115]. This scenario gained more significance when the content was uploadable to the internet where millions of people can access them. Human driven tasks like annotating the content to identifying people in a security camera feed became challenging as the amount of content grew. To compensate that, automated systems were developed to mimic human perception of the content. Several use cases came into the picture such as, (i) annotation systems which tagged the image and videos based on various properties and its content, (ii) retrieval systems which utilized such annotations to develop scalable solutions to index the content and, (ii) real time systems to extract certain kind of information e.g., understanding the surroundings in case of navigation, analysing sports videos, etc.

Several kinds of information can be extracted from multimedia content with varying levels of com- putation [110]. At the the lowest level we have trivial information which require no computation like meta-data provided with the image (e.g., date of picture taken, camera model etc.) [65]. With a little computation, we can perform some simple processing and generate tags from text associated with the image (e.g., image name and description) or also around the text in case of web pages via keywords finding or text summarization techniques [35, 122]. At the highest levels of computation, we try to in- terpret the image cognitively were we 'look' into the image and identify objects, persons, places, etc. in them [61, 108]. We see that as the higher levels of computation gives us more relevant information as compared to its lower counterparts, making the information extraction process more useful. For exam- ple, from a holiday picture, it would be most informative to tag the image with number of people, their facial expressions and objects around them than with text associated which describe the image as family on vacation or the date and location where the picture was taken.

Incompatibility with Scene Images

The methods devised by the document analysis community mostly exploited the patterns found in printed text. In printed text we come across characteristic features like dominance of text content in a page, standard layouts, adherence to a single font style, simple black on white text etc. which can only be found in printed text. However, in scene images we see that text is sparse and can be found in any style with varying foreground and background color complexities.

They follow layouts which can be cognitively perceivable by humans thus generating a large scope of possibilities as compared to a limited layout set present in printed text content. In Figure 1.1(a) we see a typical scanned book page with a fixed layout consisting of page number, book title and paragraphs with even line spacing and font type. The consistency can be exploited and the same method can be applied to the other pages of the book to recognize the text. On the other hand, the scene image consists of text with varying sizes with a layout that is only applicable to this particular image. Owing to this variation/non-uniformity present in scene text, a direct application of OCRs were expected to perform poorly on them, thus requiring alternative strategies to address the challenges.

2. LITERATURE SURVEY AND RELATED WORK

1. Faster R-CNN:

Title: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.

Authors: Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun.

Year: 2015.

This paper introduced the Faster R-CNN architecture, which is a widely used framework for object detection, including text detection in scene images. Many subsequent works have used this as a backbone for text detection.

2. EAST: An Efficient and Accurate Scene Text Detector:

Title: EAST: An Efficient and Accurate Scene Text Detector.

Authors: Xinyu Zhou, Cong Yao, He Wen, and Yuzhi Wang.

Year: 2017.

The EAST (Efficient and Accurate Scene Text) detector proposed a novel architecture that achieved state-of-the-art results in text detection tasks. It's known for its efficiency and accuracy.

3. TextBoxes: A Fast Text Detector with a Single Deep Neural Network:

Title: TextBoxes: A Fast Text Detector with a Single Deep Neural Network.

Authors: Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu.

Year: 2017.

TextBoxes introduced an end-to-end trainable framework for text detection, which is efficient and achieves competitive results. It focuses on detecting oriented text, which is common in scene images.

4. **CRNN: Convolutional Recurrent Neural Networks for Text Recognition:**

Title: An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition.

Authors: Baoguang Shi, Xiang Bai, and Cong Yao.

Year: 2016.

The CRNN model combines Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for end-to-end text recognition. It's a fundamental paper for recognizing text within detected regions.

5. **ASTER: An Attentional Scene Text Recognizer with Flexible Rectification:**

Title: ASTER: An Attentional Scene Text Recognizer with Flexible Rectification.

Authors: Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai.

Year: 2018.

3. EXISTING SYSTEM

Text can play an important role in understanding street view images. In light of this, many at- tempts have been made to recognize scene text [9, 29, 78, 94, 117, 120]. Scene text recognition is a challenging problem and its recent success is mostly limited to the *small lexicon setting*, where an image-specific lexicon containing the ground truth word is provided. Typically, these lexicons contain only 50 words [117]. This setting has many practical applications, but it does not scale well. As an example consider the scenario of assisting visually-impaired people in finding books by their titles in a library. Here the lexicon is populated with all the book titles. In this case, the small lexicon setting becomes less accurate as the lexicon sizes can range from a few thousands to a million. For instance, when lexicon size increases from 50 to 1000, the recognition accuracy drops by more than 10% [77,94]. In other words, the general problem of scene text recognition, i.e., recognition with the help of a large lexicon (say a million dictionary words) is far from being solved. In this chapter, we investigate this problem.

One way to address the task of recognizing scene text is to pose the problem in conditional random field (CRF) framework and obtain the maximum a posteriori (MAP) solution as proposed in [77, 78, 86, 101, 117, 121]. In these frameworks, an energy function consisting of unary and pairwise potentials is defined, and the minimum of this function corresponds to the text contained in the word image. These methods demonstrated successful results in a small lexicon setting primarily due to the fact that the pairwise terms are computed with this lexicon have a positive bias towards the ground truth word. However, when the pairwise terms are computed from large lexicons, they become too generic, and often in such cases the MAP solution does not correspond to the ground truth. Besides this, MAP solutions suffer from drawbacks, such as (i) approximation errors in inference, (ii) poor precision/recall for character detection, (iii) weak unary and pairwise potentials. Consider the word "PITT" shown in Figure 3.1 as an example. The MAP solution for the word is "PITA", which is incorrect. Our approach addresses this problem by using the top-M solutions to ultimately

Journal of Engineering Sciences

4. PROPOSED SYSTEM

We model the scene text recognition task as an inference problem on a CRF model, similar to [78], where unary potentials are computed from character classification scores and pairwise potentials from the lexicons. Small lexicon based pairwise potentials often help to recover from the errors made by character classification [99, 113]. However, when the pairwise potentials are computed from large lex- icons, they become too generic, and the overall model cannot cope with erroneous unary potentials. To overcome this issue, starting from a large lexicon recognition problem, we automatically refine the problem statement and convert it to a small lexicon inference task.

The framework has the following components, as shown in Fig. 3.2: (i) Candidate word generation module, where we generate multiple words with each word as a set of characters spanning over the image, (ii) CRF inference module, where each word is represented as a CRF and inferred to obtain diverse solutions, and (iii) Lexicon reduction module, where we prune the lexicon by removing distant words after re-ranking the lexicon with a novel group edit distance computed using the diverse solutions. It is accompanied by re-computation of pairwise potentials which become image specific as the lexicon size decreases. We use different stopping criteria for recognition and retrieval tasks as we alternatively reduce our lexicon and infer solutions.

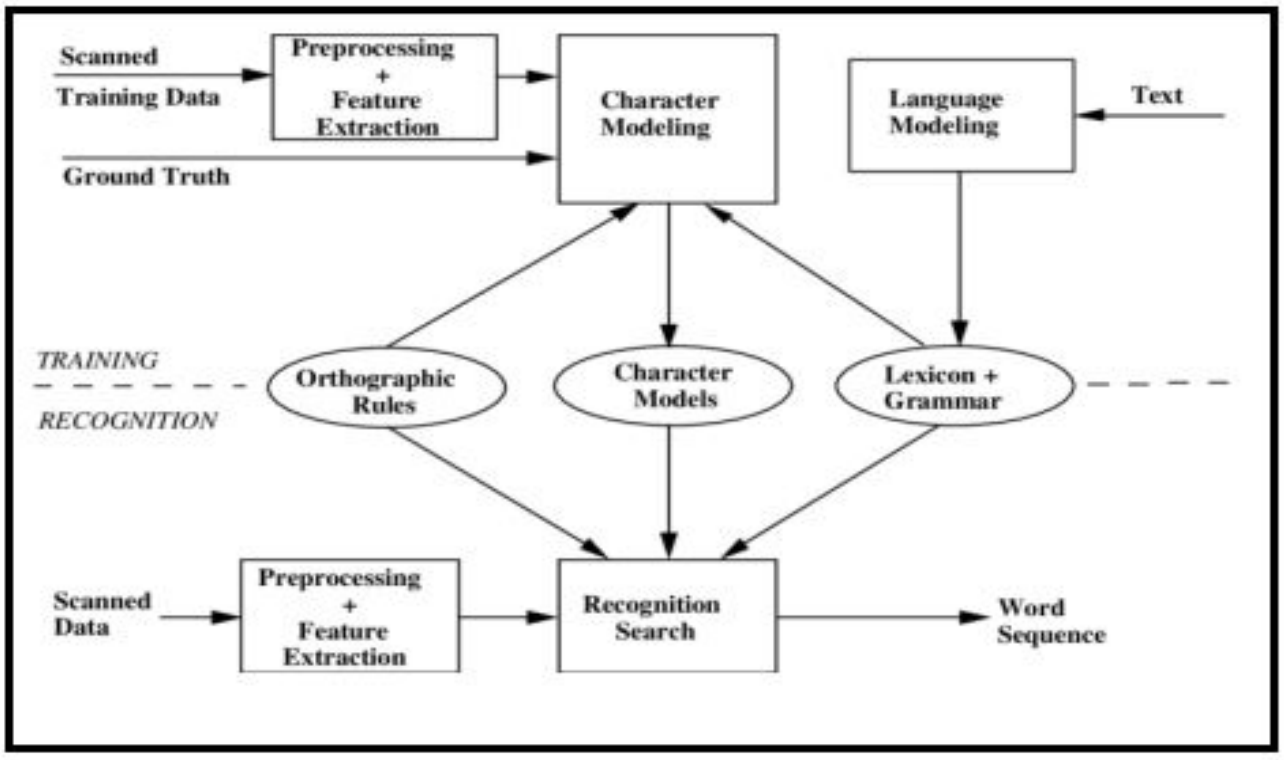


FIG 1 – SYSTEM ARCHITECTURE

5. METHODOLOGIES MODULE

MODULES

i) imutils:

Imutils is a Python library that provides a series of convenience functions to make basic image processing functions such as translation, rotation, resizing, skeletonization, displaying Matplotlib images, sorting contours, detecting edges, and much more easier with OpenCV and both Python 2.7 and Python 3 [12]. It was developed by Adrian Rosebrock, a computer vision expert and the author of the book "Practical Python and OpenCV" [2].

You can install **imutils** using pip by running the following command in your terminal:

pip install imutils

ii)numpy

NumPy is a Python library created in 2005 that performs numerical calculations. It is generally used for working with arrays.

NumPy also includes a wide range of mathematical functions, such as linear algebra, Fourier transforms, and random number generation, which can be applied to arrays.

What is NumPy Used for?

NumPy is an important library generally used for:

- Machine Learning
- Data Science
- Image and Signal Processing
- Scientific Computing
- Quantum Computing

Why Use NumPy?

Some of the major reasons why we should use NumPy are:

1. Faster Execution

In Python, we use lists to work with arrays. But when it comes to large array operations, Python lists are not optimized enough.

Numpy arrays are optimized for complex mathematical and statistical operations. Operations on NumPy are up to 50x faster than iterating over native Python lists using loops.

Here're some of the reasons why NumPy is so fast:

- Uses specialized data structures called numpy arrays.
- Created using high-performance languages like C and C++.

2. Used with Various Libraries

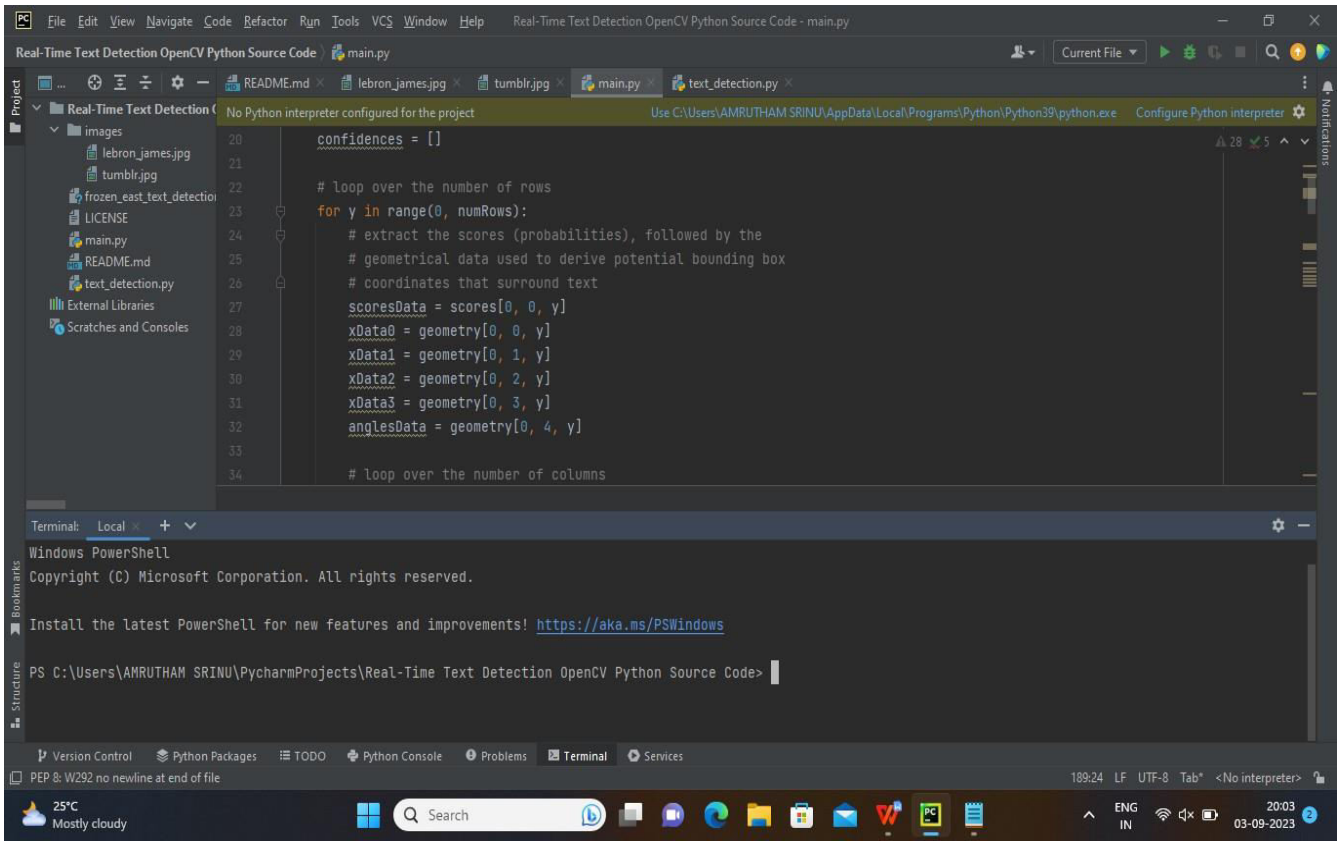
NumPy is heavily used with various libraries like Pandas, Scipy, scikit-learn, etc.

Import NumPy in Python

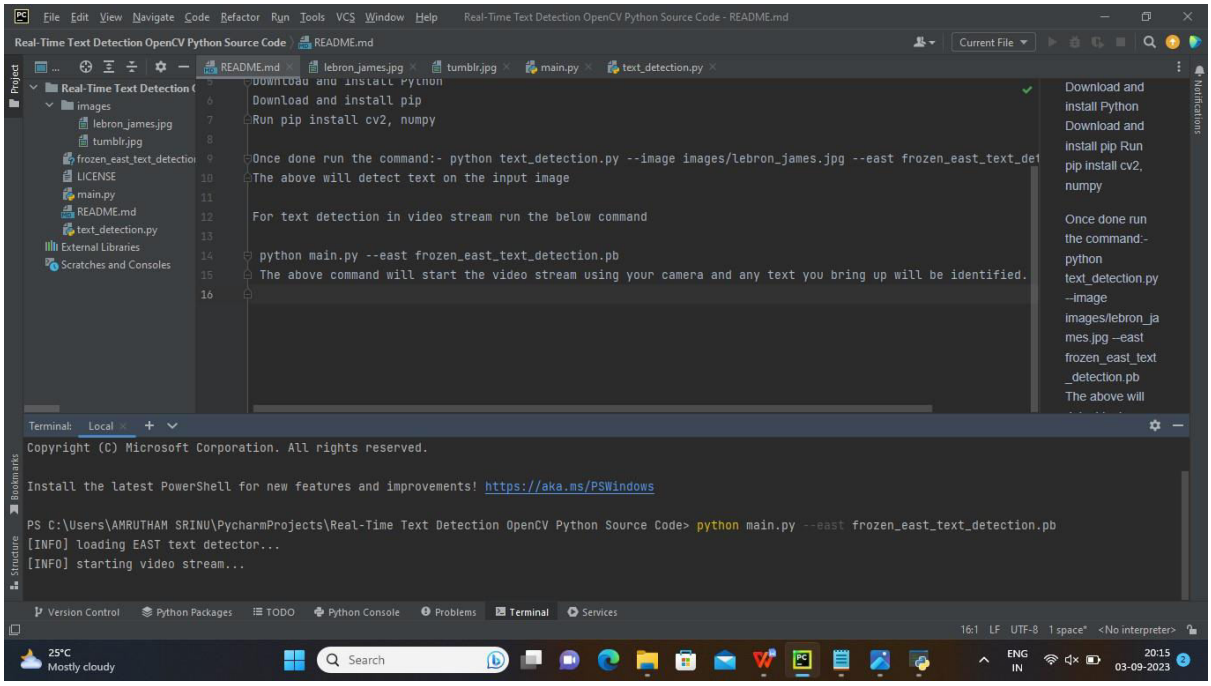
Journal of Engineering Sciences

6. RESULTS AND DISCUSSION SCREENSHOTS

HOME SCREEN



Journal of Engineering Sciences



DESCRIPTION:

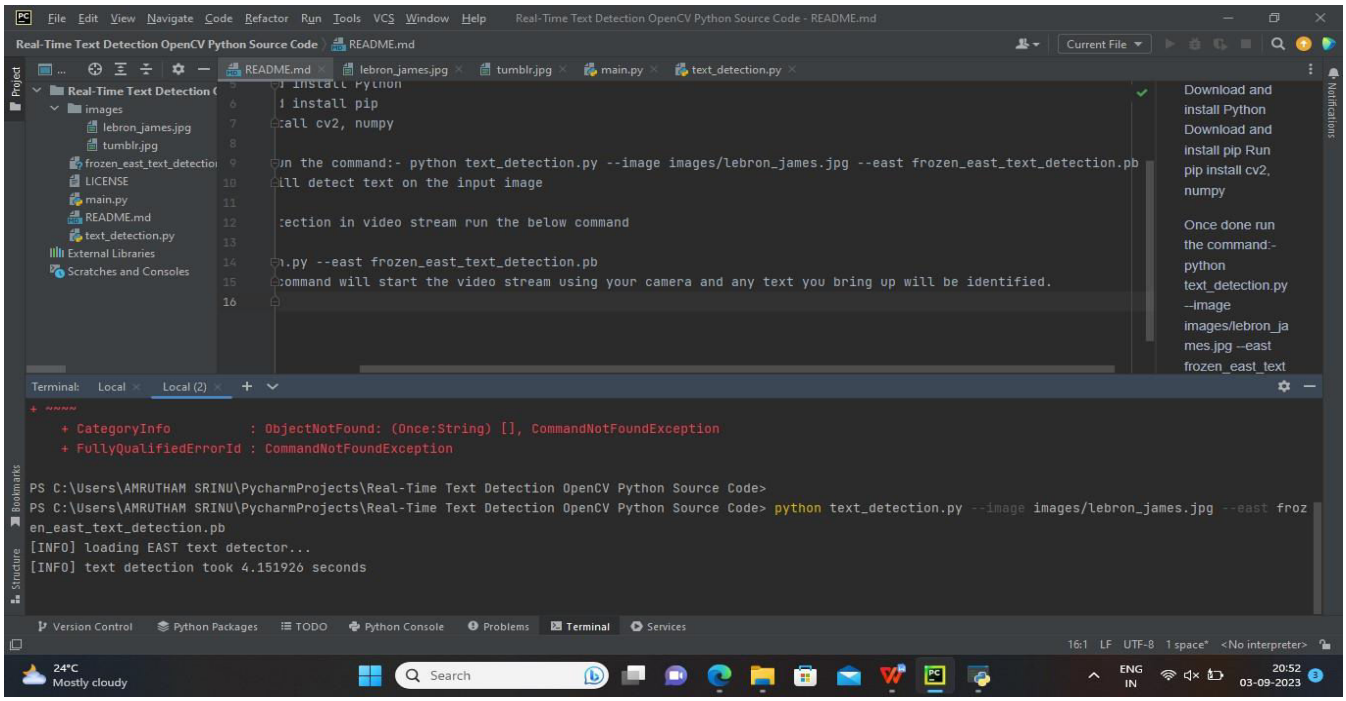
Dedecting text for video

OUTPUT SCREEN 1

Journal of Engineering Sciences

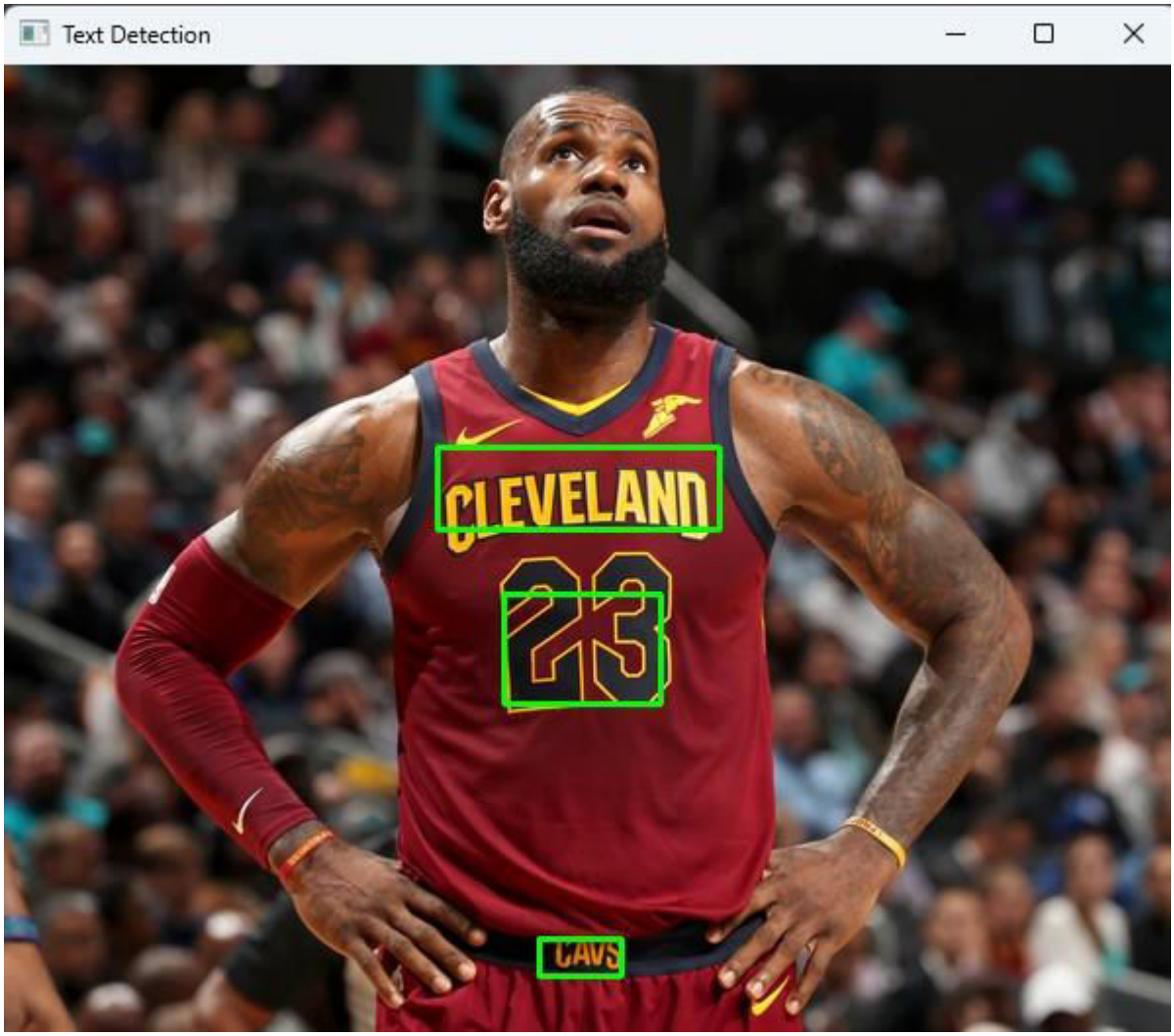
DESCRIPTION:

Detecting text from an image



OUTPUT SCREEN 2

Journal of Engineering Sciences



7. CONCLUSION AND FUTURE SCOPE

In this thesis, we proposed solutions to standard challenges in scene text analysis. We discussed and implemented two kinds of text detection methods in Chapter 2 and improved the performance of lexicon based cropped word recognition, especially for large lexicons comprising of 0.5 million words in Chapter 3. Finally, we combine the detection and recognition modules, resulting in two contrasting end to end pipelines which are discussed in Chapter 4.

For the text detection challenge, we pursue a detection via segmentation approach using hierarchical clustering. We find MSERs on the scene image which are clustered using the single linkage cluster- ing method, thus generating several overlapping text candidates. The candidates are classified using a text/non-text adaboost classifier and the positively classified ones are used to create a binarized image consisting of text pixels as foreground. We also implement a text region proposal method which gen- erates several possible word level regions on the scene image with high recall. The method relies on a patch level text/non-text CNN classifier which generates a score map with per pixel probability of text occurrence. The score map is thresholded and components are combined using RLSA to generate the proposals.

For the text recognition task, we improved an existing CRF based framework for word recognition using a lexicon. We generated several word candidates through multiple binarization of the cropped word image and represented each of them in a CRF framework with characters as nodes. Diverse solutions were inferred form the the word candidates using a pairwise prior computed from the lexicon associated which in turn was used to reduce the lexicon. The process of inference on word candidates and lexicon reduction was done iteratively, thus reducing the lexicon to a single word. We also used the iterative lexicon reduction method to enable retrieval on cropped word images. Given, a set of cropped word images with a lexicon, we partially reduced the lexicon, thus associating a small set of words to each image with high probability of ground truth amongst them. At retrieval stage, we searched for our text query among the associated lexicons and ranked the retrieved word images based on edit distance and energy scores. We found our method to perform well in .

8. REFERENCES

[1] ICDAR 2003 datasets, http://algoval.essex.ac.uk/icdar.

[2] Street View Text dataset, http://vision.ucsd.edu/˜kai/svt.

[3] M. Agrawal and D. Doermann. Clutter noise removal in binary document images. In ICDAR, 2009.

[4] B. Al-Badr and S. A. Mahmoud. Survey and bibliography of arabic optical text recognition. Signal Processing, 1995.

[5] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012.

[6] O. Alsharif and J. Pineau. End-to-end text recognition with hybrid hmm maxout models. arXiv, 2013.

[7] P. Arbela´ez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In CVPR, 2014.

[8] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse m-best solutions in markov random fields. In ECCV, 2012.