Name: Sumit Mishra

Email address: sm27598@gmail.com

Contact number: 7400255701 / 8689907119

Anydesk address: **Nil**

Date: **07th September 2020**

Self Case Study - 2: Scene Text Detection, Recognition & Translation

---

Overview

1. The need for computer vision in text detection and recognition from an image or video is getting very popular these days.
2. Because, text is to reliably and effectively spread or acquire information across time and space. In this sense, text constitutes the cornerstone of human civilization.
3. This approach can be used for handwriting recognition, natural scene text detection and recognition, vehicle number detection and recognition and many more.
4. But multiple challenges may still be encountered when detecting and recognizing text in the scene.
   Few of the challenges are as follow:
   - ➢ Text in images exhibit much higher diversity and variability especially for natural scene images.
     For example, instances of scene text can be in different languages, colors, fonts, sizes, orientations, and shapes.
   - ➢ The backgrounds of natural scenes are virtually unpredictable.
     There might be patterns extremely similar to text (e.g., tree leaves, traffic signs, bricks, windows, and stockades), or occlusions caused by foreign objects, which may potentially lead to confusion and mistakes.
   - ➢ In some circumstances, the quality of text images and videos could not be guaranteed. That is, in poor imaging conditions, text instances may be of low resolution and severe distortion due to inappropriate shooting distance or angle, or blurred because of out of focus or shaking, or noise on account of low light level, or corrupted by highlights or shadows.

5. These difficulties run through the years before deep learning showed its potential in computer vision as well as in other fields.
6. But nowadays researchers turn to deep neural networks for automatic feature learning and start with more in-depth studies.
7. Researchers are now turning to more specific aspects and challenges. Against difficulties in real-world scenarios, newly published datasets are collected with unique and representative characteristics.
   For example, there are datasets featuring long text, blurred text, and curved text respectively.
8. This project aims to build a system that can detect the text region from the natural scenes(like a street board, banner, traffic signs, store board, etc.) and recognize the text of that detected region that contains textual information and then that text can be translated to another language that end-user can understand.
9. This type of system is very useful in various scenarios like in foreign countries, the sign board or other boards will be in the language of that country.
10. So, the whole objective of this case study is to build a system that can detect and recognize a text from a natural scene image and then can be translated to another language.

Research-Papers/Solutions/Architectures/Kernels

Research paper:

Link to research paper: https://arxiv.org/pdf/1811.04256.pdf

Summary of research paper:

1. This paper explains the need of scene text detection and recognition applications in today's world, approaches that were used before deep learning era, current approaches, different dataset available for this kind of application, and their references.
2. Before deep learning came into light, most text detection methods either adopt Connected Components Analysis (CCA) or Sliding Window (SW) based classification.
3. CCA based methods first extract candidate components through a variety of ways (e.g., color clustering or extreme region extraction), and then filter out non-text components using manually designed rules or classifiers automatically trained on hand-crafted features.

4. Whereas in sliding window classification methods, windows of varying sizes slide over the input image, where each window is classified as text segments/regions or not. Those classified as positive are further grouped into text regions with morphological operations, Conditional Random Field (CRF) and other alternative graph bases.
5. For text recognition, one branch adopted the feature based methods and utilizes label embedding to directly perform matching between strings and images.
6. Strokes and character key-points are also detected as features for classification.
7. Another decomposes the recognition process into a series of sub-problems.
8. Various methods have been proposed to tackle these sub-problems, which includes text binarization, text line segmentation, character segmentation, single character recognition and word correction.
9. In summary, text detection and recognition methods before the deep learning era mainly extract low level or mid-level handcrafted image features, which entails demanding and repetitive preprocessing and postprocessing steps. Constrained by the limited representation ability of handcrafted features and the complexity of pipelines, those methods can hardly handle intricate circumstances, e.g. blurred or noisy images in the ICDAR 2015 dataset.
10. Methods after the evolution of the deep learning era are characterized by the following two distinctions:
    (1) Most methods utilize deep-learning based models;
    (2) Most researchers are approaching the problem from a diversity of perspectives, trying to solve different challenges.
11. The detection of scene text has a different set of characteristics and challenges that require unique methodologies and solutions.
12. Thus, many methods rely on special representation for scene text to solve these non-trivial problems.
13. The evolution of scene text detection algorithms undergoes three main stages:
    (1) In the first stage, learning-based methods are equipped with multistep pipelines, but these methods are still slow and complicated.
    (2) Then, the idea and methods of general object detection are successfully implanted into this task.
    (3) In the third stage, researchers design special representations based on sub-text components to solve the challenges of long text and irregular text.

14. In the deep learning era, scene text recognition models use CNNs to encode images into feature spaces. The main difference lies in the text content decoding module.
15. But still there are so many pre-trained models available for text detection and recognition like EAST for text detection, Pyserract, Azure API for text recognition and many more techniques and models.
16. An end-to-end text detection and recognition systems (also known as text spotting systems) can be directly built or both independent text detection and recognition models can be grouped together to build an end-to-end system for text detection and recognition.

First Cut Approach

1. The main objective of this case study is to build a system that can detect and recognize a text from a natural scene image and then can be translated to another language that the end user can understand.
2. The scope of this project is limited to only one language for detecting text and then converting it to another language after recognition.
3. For this project, I've chosen the ICDAR 2015 dataset which contains images for english word-level text.
    a) The ICDAR15 dataset contains 1,500 images: 1,000 for training and 500 for testing. Specifically, it contains 2,077 cropped text instances, including more than 200 irregular text samples.
    b) As text images were taken by Google Glasses without ensuring the image quality, most of the text is very small, blurred, and multi-oriented.
    c) No lexicon is provided.
4. The whole dataset can be downloaded from here.
5. As ICDAR15 images are blurred, noisy, and in low quality, so before detecting the text from those regions there are several steps needed for processing the image to deblur and de-noising it.
6. Also this dataset is multi-oriented, so there are few instances of images that are rotated or curved. So, we have to deal with this problem also.

7. To overcome this problem, there is something called Spatial transformer networks (STN).
8. Spatial transformer networks (STN) allow a neural network to learn how to perform spatial transformations on the input image in order to enhance the geometric invariance of the model.
   For example, it can crop a region of interest, scale and correct the orientation of an image.
9. It can be a useful mechanism because CNNs are not invariant to rotation and scale and more general affine transformations.
10. Overall architecture of this system can be as follows:
    1] Text detection: Detect the text region or say getting coords of text region in an image,
    2] Text recognition(Recognizing the text from the retrieved bounding box images using OCR),
    3] Text correction: There might be a possibility that 1 or 2 char of text is recognized incorrectly, which will make the whole word incorrectly recognized for word-level recognition. So this can be avoided by implementing a text correction model which can correct the spelling of recognized words to improve the model,
    4] Text language translation: Recognized and corrected text can be translated into any language that the end user can understand.
