Inhaltsverzeichnis

1 What is Scene Text?

We interact with hundreds of different instances of text every day. From text messages on our phones, invoices in our mail and e-mails at work to hand-written birthday cards and wildly stylized advertisement to draw our eyes in. The pattern-desperate human brain has almost no problem recognizing text in even its most abstract form, provided it knows the right language. So it often seems hard to understand why something that comes so easily to humans could pose such a challenge to machines.

The instances of text described above can roughly be put into two categories: clear, machine-made text (e.g. typed invoices, e-mails, newspapers) and text as it appears 'in the wild' (e.g. a sign on a shop in an artsy font viewed at an angle, the text on a footballers shirt on a TV broadcast). The latter constitutes the field of Scene Text: text within natural images or videos, in often uncontrolled and complex environments.

Clear, machine-made text had been the focus of automated text recognition for a long time, such a long time in fact that it had become almost synonymous with automated text recognition. Scene text recognition was not feasibly accomplishable with the tools and computing power available in computer vision. It poses a number of challenges compared to clear black-on-white machine text in a controlled environment:

- Diversity and Variability of Text in Natural Scenes Distinctive from scripts in documents, text in natural scene exhibit much higher diversity and variability. For example, instances of scene text can be in different languages, colors, fonts, sizes, orientations, and shapes. Moreover, the aspect ratios and layouts of scene text may vary significantly. All these variations pose challenges for detection and recognition algorithms designed for text in natural scenes
- Complexity and Interference of Backgrounds 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.
- Imperfect Imaging Conditions In uncontrolled 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 noised on account of low light level, or corrupted by highlights or shadows.

2 A short history

Mainly the major changes brought on by Deep Learning, but also a short look at what methods existed before
Character vs. word based recognition
New focus on specializing models and datasets for specific tasks

3 State of the art

Graph of accuracy development over time
No in-depth explanation, for that I can link to other resources, but a quick deeper dive into what is state of the art now
Question of synthetic datasets

4 So what do you want?

What do you want to achieve? Do you want to train your own model (with your own dataset)? Do you want to evaluate a pre-trained model on a (new) dataset? Do you want to use a pre-trained model, ready to go, for text detection/recognition/spotting?
Question of scope, factor of computation time, depending on what you what (i.e. some models are slow at training, but faster at evaluating/is training time even relevant if you only want to test?)
Ease of setup, but hopefully this tutorial can help with that

4.1 Detection, Recognition, Spotting

4.2 Training

4.3 Evaluation

4.4 Testing

4.5 Other factors to consider