Name: Diptesh Milind Chaudhari
Registration id: SIRSS02214

Text extraction involves text detection, localization of text, tracking of text, extraction of text, enhancement, and text recognition from a given image. Text detection discovers whether text is present in a given image or not. Normally text detection is applied for a sequence of images.

Text recognition with machine learning

As we know, we need to teach the computer to recognize what we know is text. The task is a bit simpler when we talk about high-quality, legible pictures, where the text is clearly visible, and so are all the letters and digits. But what about pictures or scans of more mediocre quality? This is where the challenge begins. However, let's see how exactly does machine learning text recognition work.

1.OCR – Optical Character Recognition

First, we begin with the most common text recognition technique, and this is the OCR–Optical Character Recognition. OCR yields outstanding results only in very specific use cases, but in general, it is still considered as challenging. Optical Character Recognition is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files, or images captured by a digital camera into editable and searchable data.
Let's say we have a piece of paper–a high school diploma. You can use your scanning device to put it into a computer, but it's not editable, for instance, with the MS Office tool. You need much more advanced graphics software to edit it. That takes time and requires specific skills. If you want to extract and repurpose data from this scanned document, you need an OCR software that would single out letters, put them into words, and then–words into sentences. This allows you to access and edit the document's contents at once.

The usage of the OCR software

The OCR software is by no means one, a uniform application that serves one and the same purpose. The OCR applications are used to serve lots of different intents. We can start with "reading" the printed page from a book or a random image with text (for instance, graffiti or advertisement), but we go on to reading street signs, car license plates, and even captchas. OCR software takes into consideration the following factors and attributes[1]:

Text density. On a printed page, the text is dense. However, given an image of a street with a single street sign, the text is sparse. The OCR software has to recognize both.
Text structure. Text on a page is usually structured, mostly in strict rows, while text in the wild may be scattered everywhere, in different rotations, shapes, fonts, and sizes.
Font. While computer fonts are quite easy to recognize, handwriting font is much more inconsistent and, therefore, harder to read.
Artifacts. There are almost none of them on a perfectly scanned page, but what about outdoor pictures? In short, this is a completely different story, and you have to keep that in mind when using OCR.

Real-world examples

1. Now, let's consider major examples for the real-world, outdoor conditions: House numbers and car license plates. House plates are extremely important, just to mention Google Street View and Google Maps. This is a massive source of tons of different house numbers. And as an example, Stanford University created out of them the SVHN[2] (Street View House Numbers) dataset. SVHN incorporates over 600,000 digit images and is aimed at developing machine learning and object recognition algorithms
2. One of the companies that manage private car parks is Unipark. This is a company that operates in several European countries, such as Poland, Lithuania, Latvia, Estonia, and Belarus. It uses text recognition and extraction to manage cars driving in and driving out. When a vehicle approaches the barrier, the camera (similar to speed cameras used in Poland) takes a picture of its license plate, sends it to the company's central database, and the barrier automatically opens. When the text recognition part is done, the software extracts the car's number plate and processes it into a plain, editable text, written in regular font.
3. Here's another example. As we already know, Google Lens is an app that uses some image processing techniques along with machine learning technologies to give you more information about the object you're pointing at. But what happens if a printed document is an object in question? Google Lens fires up its text recognition algorithm and allows you to directly translate the text from the original language into output one.

2.Text extraction from images using machine learning

With the text recognition part done, we can switch to text extraction. You see, at the end of the first stage, we still have an uneditable picture with text rather than the text itself. To solve this problem, the next step is based on extracting text from an image. Right after text recognition, the localization process is performed. All the related features about a particular image are gathered.

Text extraction: how does it work?

Text extraction, also known as keyword extraction, bases on machine learning to automatically scan text and extract relevant or basic words and phrases from unstructured data such as news articles, surveys, and customer support complaints.

The text extraction and enhancement methods are applied with the help of machine learning algorithms. And finally, the extracted text is collected from the image and transferred to the given application or a specific file type. There are many types of text extraction algorithms and techniques that are used for various purposes. Therefore, we can divide them into five main methods[3].

REGION-BASED METHOD

This method of text extraction uses a sliding window to detect text from any kind of image. This approach relies on several factors, such as color, edge, shape, contour, and geometry features.

TEXTURE-BASED METHOD

This method uses various kinds of texture and its properties to extract text from an image.

HYBRID TECHNIQUE

It's the combination of the previous two techniques. First, the region-based approach is used to detect a text. Then, with the usage of the texture-based method, all the features are extracted from the text region.

EDGE BASED METHOD

As its name indicates, this method is based on the detection of the edges of every letter and digit. This method is used to develop a high-level contrast between the text and the background.

MORPHOLOGICAL BASED METHOD

This method is used to extract all the text-related features from the processed image.

The text extraction from images using machine learning software

There are many programs, algorithms, and applications that make text extraction from an image accessible. In fact, the list is very long, and it comprises several dozen apps and programs. Most of them are paid, but we have two free and handy tools of text extraction from images on our list as well!

- Altair Monarch (according to G2.com[4], it is the fastest and easiest way to extract data from any source)
- Webhose.io (this app specializes in providing access to structured data from millions of web sources, even from deep and dark web)
- Import.io (it's a SaaS product that enables users to convert the mass of data on websites into structured, machine-readable data)
- DocuClipper (it's a cloud solution to extract fields and tables from scanned documents)
- Photo Scan (it is a free Windows 10 OCR app you can download from Microsoft Store. It recognizes the text from photo files but also directly from the PC's webcam)
- Microsoft OneNote (as it turns out, this Windows 10 free tool can also extract text from a multi-page printout with one click! It works both on pictures and handwriting text).