NAME: NISHANT KUMAR  
ID: SIRSS2272

Text extraction from given images

**What is text recognition?**

As you know, you 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.

**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.The most advanced OCR systems are focused on replicating natural human recognition. The OCR systems are based on three main rules – INTEGRITY, PURPOSEFULNESS and ADAPTABILITY.First, the observed object has always to be considered as one entity comprising many interrelated parts. In our case, the diploma is such an entity. Second, any interpretation of data must always serve some purpose. And finally, the OCR program has to be capable of self-learning.

**The usage of the OCR software: -**

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: -

• **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 scenarios/examples: -**

**Google Lens text recognition**

As you 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.

**Text extraction from images using machine learning.**

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.

**Text extraction methods:**

**1. 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.

**2. TEXTURE-BASED METHOD**

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

**3. 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.

**4. 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.

**5. MORPHOLOGICAL BASED METHOD**

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