Assignment 13: Text Extraction from Given Images.

Name: Gaurav Ganesh Bhogale

Registration ID: SIRSS2101

Q1. What is Text Extraction?

OPTICAL CHARACTER RECOGNITION:

OCR (optical character recognition) is the use of technology to distinguish printed or handwritten text characters inside digital images of physical documents, such as a scanned paper document. The basic process of OCR involves examining the text of a document and translating the characters into code that can be used for data processing. OCR is sometimes also referred to as text recognition.

OCR systems are made up of a combination of hardware and software that is used to convert physical documents into machine-readable text. Hardware, such as an optical scanner or specialized circuit board is used to copy or read text while software typically handles the advanced processing. Software can also take advantage of artificial intelligence (AI) to implement more advanced methods of intelligent character recognition (ICR), like identifying languages or styles of handwriting.

The process of OCR is most commonly used to turn hard copy legal or historic documents into PDFs. Once placed in this soft copy, users can edit, format and search the document as if it was created with a word processor.

Q2. How Optical Character Recognition works?

The first step of OCR is using a scanner to process the physical form of a document. Once all pages are copied, OCR software converts the document into a two-color, or black and white, version. The scanned-in image or bitmap is analyzed for light and dark areas, where the dark areas are identified as characters that need to be recognized and light areas are identified as background.

The dark areas are then processed further to find alphabetic letters or numeric digits. OCR programs can vary in their techniques, but typically involve targeting one character, word or block of text at a time. Characters are then identified using one of two algorithms:

1. Pattern recognition- OCR programs are fed examples of text in various fonts and formats which are then used to compare, and recognize, characters in the scanned document.
2. Feature detection- OCR programs apply rules regarding the features of a specific letter or number to recognize characters in the scanned document. Features could include the number of angled lines, crossed lines or curves in a character for comparison. For example, the capital letter "A" may be stored as two diagonal lines that meet with a horizontal line across the middle.

When a character is identified, it is converted into an ASCII code that can be used by computer systems to handle further manipulations. Users should correct basic errors, proofread and make sure complex layouts were handled properly before saving the document for future use.

Q3. Use cases of OCR:

OCR can be used for a variety of applications, including:

- Scanning printed documents into versions that can be edited with word processors, like Microsoft Word or Google Docs.
- Indexing print material for search engines.
- Automating data entry, extraction and processing.
- Deciphering documents into text that can be read aloud to visually-impaired or blind users.
- Archiving historic information, such as newspapers, magazines or phonebooks, into searchable formats.
- Electronically depositing checks without the need for a bank teller.
- Placing important, signed legal documents into an electronic database.
- Recognizing text, such as license plates, with a camera or software.
- Sorting letters for mail delivery.
- Translating words within an image into a specified language.

Q4. What is Text Extraction?

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.

Machine learning algorithms

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.

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.

Text extraction from images using machine learning, black, white

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!

Text extraction from images using software

Altair Monarch (according to G2.com, 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).

Use cases of text extraction from images

Every day, 2.5 quintillion bytes of data are generated by Internet users. A fascinating fact is that by 2020, each person generated 1.7 gigabytes in a single second. [7] Comments on social media, product reviews, emails, blog articles, search queries, discussions, and so on. But the question is, how might text extraction from images help especially your company in becoming more efficient and take full advantage of the potential of data?

Social Media Monitoring

Your company can use text extraction from images to follow social media conversations to better understand customers, improve products, or take quick action to avoid a PR crisis. Text extraction from images may offer specific examples of what people on social media are saying about your business. Moreover, you may discover keywords and track trends with text extraction from an image.

Client Service with Text Extraction

Quality customer service can give your company a competitive edge. After all, when it comes to buying something, 64% of customers choose the quality of customer service over the price. In other words, text extraction from images allows customer service staff to automate the process of tagging tickets, saving dozens of hours that might be spent on real-world problem-solving. So, this is the key to customer satisfaction.

Client service with text extraction

Business Intelligence and Text Extraction from Images

Text extraction from images can also be effective in business intelligence (BI) applications such as market research and competition analysis. You may also get information from a variety of sources, including product reviews and social media, and participate in discussions on topics of interest. Furthermore, you can compare your product reviews with those of your competitors using text extraction from images and other text analysis tools. This helps in getting information that will help you in making data-driven decisions to improve your product or service.