

This work is licensed under a [Creative Commons “Attribution-NonCommercial-ShareAlike 4.0 International”](#) license.



Diverse Documents: Challenging Optical Character Recognition

Orville “El” Anderson

and10393@umn.edu

Division of Science and Mathematics

University of Minnesota, Morris

Morris Minnesota USA

Abstract

Optical Character Recognition (OCR) is technology used to extract text from images. OCR has a wide variety of uses, with one common application being to digitize scanned documents. OCR has three main categories of challenges that reduce accuracy when applied to scanned documents, stemming from page layouts, the writing system used, and visual noise. By increasing the number of document collections that contain these variations, we make it easier to develop OCR techniques to target them. This paper looks at document collections for evaluating the extent of these weaknesses and researching methods to address them. Additionally, this paper looks at the role these collections can play in training neural network based OCR approaches.

Keywords: optical character recognition, scanned documents, layout, languages, visual noise, datasets

1 Introduction

Physical documents are used frequently and for a variety of purposes. Some examples are pages of notes, tax forms, and recipes. A common way to save and share these documents is to photograph them. This creates a *scanned document*.

For many reasons such as research and record keeping, the contents of the scanned document may be important. Scanned documents do not contain machine-readable text. While text on a scanned document may be readable visually, no record of the contents are made when the image is made. As a consequence, it is not possible to edit or search within the text of a scanned document. One method to digitize a scanned document, is to read and re-type the contents. An alternative method is to use an *Optical Character Recognition (OCR) model*. OCR models are programs specifically made to identify text from images.

OCR began as a tool to read messages for Blind people and later to convert messages to Morse code. As the technology developed, the purposed shifted to more commercial applications, such as record keeping. The technology grew dramatically in popularity because it could reduce the human labor needed for data entry [4]. As applications for OCR have grown, the technology has been applied to increasingly diverse types of documents. Several aspects of documents play a role in the effectiveness of OCR on them. There is

an increased need for OCR that works on a wide variety of documents. One of the big challenges to OCR accuracy is the writing system of the language of the document. As the third most used writing system, this paper uses examples of Arabic documents to discuss this topic. The other two challenges, page layout and visual noise are covered, but to a lesser degree.

Section 2 of this paper, covers the stages and techniques used in OCR. The three challenges related to document variations are the page layout, the visual noise of the image, and the writing system used. Section 3, Challenges, looks at how layout, noise, and writing system specifically impact OCR. It covers how these challenges relate to the stages of OCR and covers some additional steps which can be used to minimize the effect. Section 4, goes deeper into the techniques covered in Sections 2 and Section 3, specifically into the datasets that were used to research these areas. Section 5, Conclusion, discusses the importance of specialized datasets and their role in increasing accuracy of OCR.

2 Stages and Techniques

The input and output of the OCR process depends on the model used. In general, the input is an image and the output is a text file. Some models, in addition to extracting text from the image will record the location of where the text was found to construct more complex output formats. Tesseract, an open-source model discussed in this paper, accepts a variety of file formats, mainly, PNG, JPEG, and TIFF files. Tesseract currently can format the output as plain text, HTML, as several types of PDFs, and more [5].

The process of OCR can be split into a number of stages, but for the purpose of this paper is made of three: *Document Layout Analysis (DLA)*, *Text Line Detection (TLD)*, and *Recognition*. Two common stages which are not recognized in this list are a pre-processing and post-processing stage. Because these stages are not necessary to perform OCR, they are instead mentioned in Section 3.2 as noise-reduction techniques.

Figure 1 shows a document, in Persian, going through each of the three stages of OCR, where the output of the model is the text identified from the document. The first stage, DLA, breaks the image up into regions of text. For

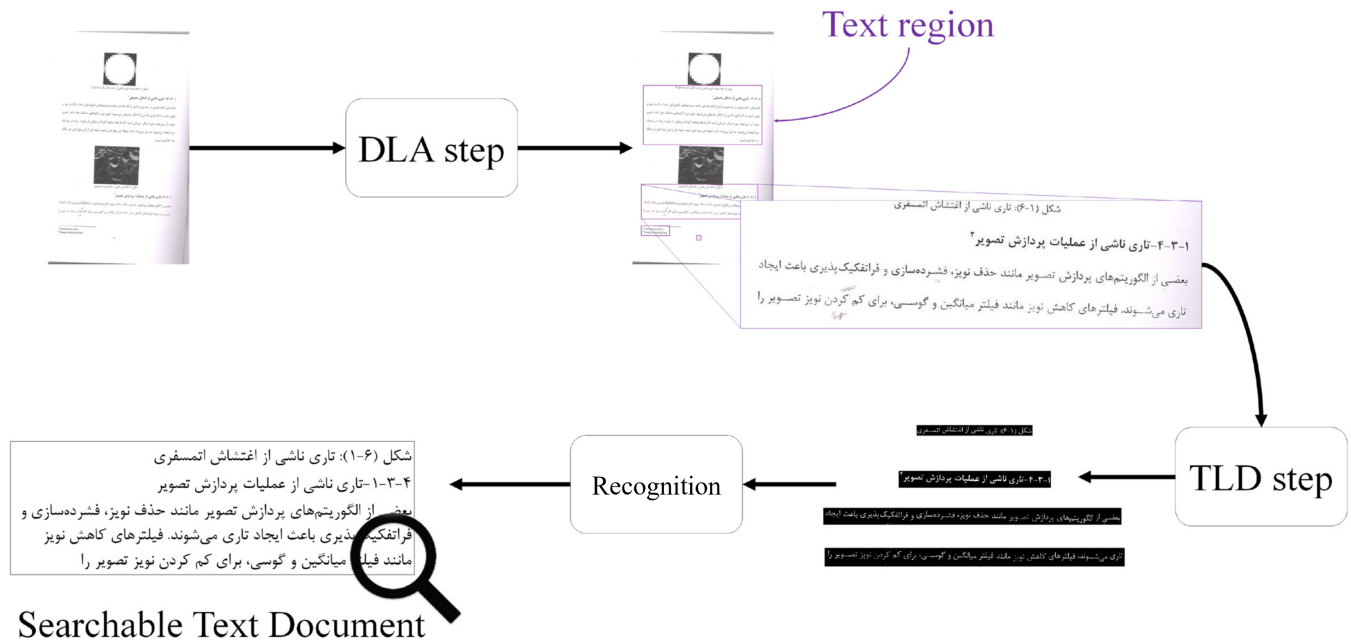


Figure 1: Stages of OCR pictured on a Persian Document. Modified from Fateh et al [2].

the example document, this is two paragraphs, a page number, and some isolated words. The second stage, TLD, breaks down the sections of text into individual lines of text. Some TLD methods will further break the lines of text down into individual words or characters. The final stage, Recognition, identifies the text in each line and combines the results into one search-able text document.

2.1 Neural Networks

One technique which can be used in each of the three OCR stages, is to use *Neural Networks*. Neural Networks are a subsection of Machine Learning meant for pattern recognition. These networks use large functions with many parameters to make predictions. In the context of OCR and scanned documents, these predictions are often if a region contains text, or are about the value of some text.

To chose the parameters, or *weights*, in the Neural Network, it must go through a training process. This process begins with a large collection of examples, a *dataset*. Datasets can be anything from full documents to individual characters, depending on which stage the Neural Network will be used for. It is important for the training process that there is an existing machine-readable copy of the contents of the examples. The dataset is divided into two or more groups. The Neural Network is exposed to the first group of examples, to get a sense of patterns within the possible outcomes. In this step the Neural Network will set initial values for the weights. After exposure, the network is then exposed to the second group of examples. The network is meant to use the weights chosen in the earlier step to predict the values

of the second set. The accuracy of the predictions are then used to adjust the weights in a way that minimizes errors. This process repeats as needed. If the dataset was divided up into more than two groups, the remaining examples can be used as a final measure for accuracy of the Neural Network's predictions.

2.2 Document Layout Analysis

The purpose of DLA is to identify what part of the input image is text and what is not. To do this, a model can use a variety of techniques to draw rectangular boxes around the text. These boxes can be a variety of sizes and can contain varying amounts of text. The areas not captured by the boxes are ignored for the rest of the stages, to reduce the workload.

An important step of DLA is preserving the reading order of the document, which can be non-obvious when it comes to documents beyond a single column of text. Features like the number of columns of text, and the presence of components like images and tables, disrupt a traditional top-down reading order.

Fateh et al [2] propose a DLA method which uses a voting system on the output of four neural networks trained to recognize regions of text. Each of the four voting neural networks, use coordinates to record where on the original image the text was found. The study used Tesseract, a modular OCR model, as the foundation. The recognition stage of Tesseract, takes full lines of text as an input and returns full lines of machine-readable text as the output. Output formats of Tesseract, such as HTML and PDFs, do not recombine text from multi-line regions, like paragraphs. Instead, the

text is left in individual text boxes. The need to preserve the document order is partially hidden when OCR is used in this way. The human labor needed to combine text boxes of individual lines of text, is less than the labor needed to combine individual characters.

2.3 Text Line Detection

TLD takes the previously identified regions of text and further breaks them down into lines, words, or individual characters. The output of TLD, is each identified unit of text in its own defined box of pixels. To best fit the text into the boxes, TLD may rotate, center or scale the text. Traditionally, these boxes have a uniform pre-set size. It is important for the boxes in question to be large enough to contain the intended character(s), but not to include characters from the neighboring lines or words. This results in a long discussion of what is the proper size box.

The size of the unit of text is determined by the technique used in the Recognition stage 2.4. Modern techniques, such as Neural Networks 2.1, can be used on full lines of text, instead of individual characters. By ending the segmentation process at lines of text, instead of further breaking the text into individual characters, we can prevent introducing errors related to character overlap. One downside of using full lines of text, is that because they cover more of a page, the line has a larger capacity to be curved.

Fateh et al [2] propose a TLD technique which uses font size to rotate and standardize full lines of text. This can be seen in Figure 1, where the text entering the TLD step is at an angle, but the output is rotated, to create a baseline. By creating this baseline, and somewhat standardizing the characters, the accuracy when defining the characters is improved. To develop this technique, they built a special dataset made to include curved lines of text and lines with very little space in-between.

2.4 Recognition

The final step in the OCR process is to attempt to recognize the text identified in the document. Because OCR was originally developed to convert printed characters to sound, many of the initial recognition techniques are no longer used. When the technology moved from noise to text output, some techniques arose which are still relevant. One such technique is known by many names, one being Matrix Matching. Matrix Matching is important to understand because it is the foundation of many modern recognition techniques and can help explain the common restrictions in recognizing text.

Matrix Matching uses templates of known characters. In this process, a single unknown character is compared to all of the templates. For each template, the number of non-matching pixels between it and the single character are recorded. The output of Matrix Matching is a list of the templates and how similar they were to the unknown character.

The identity of the template with the highest similarity, is chosen as the identity of the unknown character.

Neural Networks are a popular technique in the Recognition stage. The networks work similarly to Matrix Matching, but the weights add a level of flexibility when it comes to identifying characters. Neural Networks are more accommodating to types of fonts and variations in character rotation and placement. Because of the training process for Neural Networks, through repeated exposure, it is also possible to train the model to predict previously unknown characters.

3 Challenges

There are three main categories of issues that decrease accuracy of OCR for documents. These categories are the layout of the original document, the presence of visual noise, and the writing system used.

3.1 Layout

Documents come in many different layouts. Images, figures, number of columns, and similar aspects, add a layer of complexity to documents. In the DLA stage, and when the model outputs the final result, to be accurate, the model must have some method to record and replicate the reading order. A paper formatted with two columns, such as this, is meant to be read left column, then right column. Unless otherwise instructed, an OCR model will take the first line from each column and treat them as one line.

Some OCR models are intentionally made to only handle one layout, such as a specific tax form, or a job application for a specific company. A specialized OCR model, when applied to the layout it is made for yields higher accuracy. Layout specific-OCR strategies are not directly applicable to other layouts, and can not be readily combined with other layout-specific models.

One related feature of document text, is line spacing. If lines of text are close together, or overlapping, the boxes drawn around the lines, will overlap. This can also occur if the lines of text are at an angle, or are curved. This results in the offending overlap being taken into consideration when identifying the otherwise unrelated text.

3.2 Visual Noise

One big factor in the accuracy of OCR is the quality of the initial image. Marks on the page, or noise acquired when the image was captured, add additional complexity to the OCR process. In 2022, Thomas Hegghamer [3], a historian, performed a bench-marking experiment to better understand how OCR models of the time were impacted by the types of noise he found in historical documents. Figure 2 shows the six different types of visual noise he studied. To perform this experiment Hegghammer created a dataset based on 422 documents from two existing datasets. Hegghammer kept a color version of each of the documents and made a

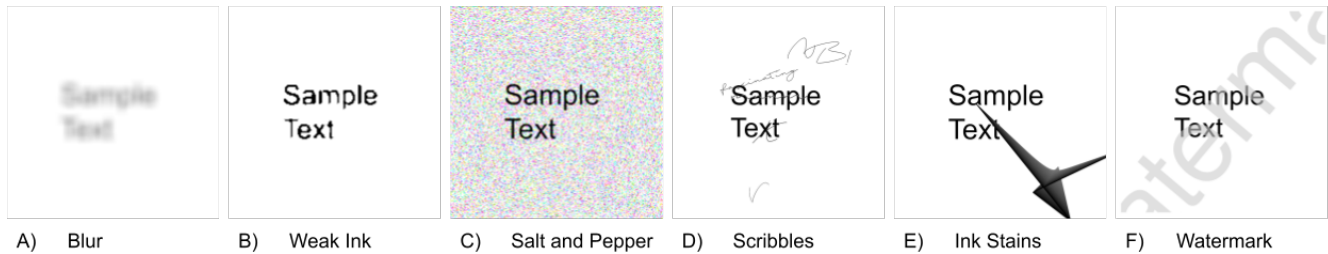


Figure 2: Examples of visual noise: Blur, Weak Ink, Salt and Pepper, Scribbles, Ink Stains, and Watermarks.

copy in black and white. To each of the versions, he then applied layers of each of the noise types, so each version had additional versions with zero, one, or two layers of noise.

Hegghammer came to two conclusions about the impact of noise on the OCR model’s accuracy. First, documents with several layers of noise, produced OCR output with higher percentages of incorrect words, than those with less layers of noise. For Tesseract, on English documents, with no applied noise, the mean word error rate was 2.4%. For the same documents with one and two layers of noise, the mean word error rates were 23.3% and 41.4% respectively.

Hegghammer’s second finding was that *integrated noise*, noise that was built into the document, had a larger impact on accuracy than *superimposed noise*. In Figure 2, Blur and Salt and Pepper are the two types of integrated noise. Scribbles, Ink Stains, and Watermarks, are considered superimposed noise. Weak ink does not neatly fit into either category. Of the English documents, with one layer of noise, the smallest mean word error rate for Tesseract came from the documents with weak ink applied. That mean value was 10.6%. The second smallest value came from the documents with the Ink Stains applied, with a mean value of 16.1%. The highest mean word error rate came from the documents with Salt and Pepper, applied, with a value of 70.2%. This means that over 70% of the words in those documents had at least one misidentified character¹.

Integrated noise impact the ability to distinguish the boundaries and lines of characters. Two additional ways that noise can reduce OCR accuracy, is by being mistaken for characters, or by covering up characters. In instances like, Scribbles and Watermark, from Figure 2, the noise is made up of text, but is frequently not an indented part of the output. Because of the placement of Watermark, the superimposed character can be identified and inserted throughout the output. Noise types like Ink Stain can partially and fully obscure text. In these cases, a human reader is left to fill in the missing text, by guessing or attempting to counteract the stain. Obscured

text in OCR can result in missing text and disrupts the reading order.

The errors resulting from noise, can be addressed in two parts of the OCR process: pre-processing and post-processing. Avyodri et al [1] performed a literature review of 29 OCR-related studies, to identify techniques used in pre-processing, post-processing and the three main OCR stages. Three out of the four papers related to pre-processing that they highlight are largely focused on addressing image rotation. Additional techniques used in this step were removing any borders, and sharpening or blurring regions of the image. Five of the papers reviewed by Avyodri et al were focused on post-processing. All of the post-processing approaches frequently, post-processing consists of removing spelling and grammatical errors. Some techniques use external spelling tools, such as Google’s online spelling suggestions.

3.3 Writing Systems

The most commonly used writing systems, by number of users worldwide: are Latin, Chinese, and then Arabic [6]. The majority of OCR models are trained to recognize characters from the Latin alphabet. As mentioned in Recognition 2.4, OCR models are generally limited in what characters they can recognize, to ones they have been exposed to previously. Additionally, the techniques used to identify Latin characters do not automatically extend to characters from other writing systems. The current limitations in identifying a variety of text types is most easily seen in non-Latin language documents, but also apply for documents with a variety of fonts, or documents with handwritten text.

Because of its popularity, and how different it is from Latin, Arabic is a relevant example to discuss this weakness in OCR. The Arabic alphabet has three main differences from Latin: the use of connected characters, the use of diacritics, and the reading order.

3.3.1 Connected Characters. Connected characters often appear in handwritten documents, regardless of writing system. Arabic is notable here because typed characters are also cursive, where typed Latin characters are not. In a study by Fateh et al [2] on improving TLD accuracy for Persian

¹Hegghammer uses the ISRI word accuracy tool, which does not count case errors or excess words.

أبجدية رومانية

Figure 3: The heading of the Latin Alphabet Wikipedia page, in Arabic.

text, they found, to better account for connected characters, the boxes drawn around each character must be larger. The study specifically put forth a TLD technique which uses box sizes determined by the font size. They tested their TLD technique on three datasets, one they created, an existing Arabic dataset, and one existing Persian dataset. They found that for each of the datasets, when the proposed technique was used by Tesseract, as opposed to base Tesseract, the total error rate for text and diacritics dropped. The total error rate went from 6.235% to 3.431%, from 13.38% to 1.52%, and to from 6.22% to 3.88% for each of the datasets, respectively. This shows that accuracy of this OCR model on Arabic and Persian documents was greatly improved with the addition of this specialized TLD technique.

3.3.2 Diacritics. A diacritic is a small graphic symbol added to a letter. In Figure 3, diacritics can be seen both above and below the main line of text. Written Arabic does not include vowels, and instead relies on the reader to use context clues to place them. Diacritics are especially important to the Arabic alphabet because they can be used to indicate the necessary vowel when the context is ambiguous [7]. Diacritics appear in other Latin-based languages such as German and Spanish, but to a lesser degree. These diacritics can be mistaken for visual noise.

3.3.3 Direction. The Arabic writing system is read from right to left, where the Latin writing system is read from left to right. During the DLA stage 2.2, and when the identified characters are assembled into the output, parts of the process must be reversed, to accommodate this. For writing systems such as those based on Chinese characters, where they are read top to bottom, the techniques used in OCR must be adjusted further, to change how the regions of text are drawn to begin with.

4 Datasets

Thomas Hegghammer’s effort to understand the effect of noise on OCR accuracy was limited by the datasets he used to compare models on, specifically in the variety of layouts present in the documents and in the types of noise. Hegghammer says, “While not an exhaustive list of possible noise types, they represent several of the most common ones found in historical document scans.” Additionally, he remarks, that if not restricted by dataset size and computation costs, the

dataset could have been expanded to cover up to 10-20 total noise types, instead of the six he used. The total dataset consists of 422 documents, with 43 variations of each, for a total of 18,568 documents. When compressed, these files are about 26 GB and are about 193 GB uncompressed. Thomas Hegghammer made contents of his dataset publicly available, along with the noise generator he made, and the output from each of the models evaluated, for each of the documents.

Fateh et al [2] look at new methods to improve accuracy of the TLD and DLA steps when applied to Persian text. Persian is derived from the Arabic script. In this paper, the authors discuss the use of separate datasets to test their proposed TLD and DLA methods. This paper highlights several DLA-specific datasets which utilize newspapers and magazine collections to provide a variety of layouts. When it came to testing their TLD method, they specifically note: “TLD in complex scripts like Persian and Arabic presents unique challenges, and the availability of suitable standard datasets is limited. Unlike English or other widely studied languages, Persian and Arabic require specialized datasets and approaches to tackle text line extraction effectively.” In total, this paper used three TLD datasets, and five DLA datasets. Of the three TLD datasets, one of them was specifically created for this study. The Official Iranian Newspapers (OIN) dataset, which is made of images of Iranian newspapers, was made to have rotated lines of text, regions with closely spaced lines, and to have a large amount of diacritics.

5 Conclusion

A 2022 literature review of OCR research papers found that several implementations and techniques for English scanned documents resulted in accuracy rates above 90%. In some cases, techniques employed resulted in accuracy rates as high as 98 and 99% [1]. While this is impressive, when taken in the context of larger bodies of work, this still leaves a large amount of errors. In projects, where truthful representation of the source documents is a high priority, these errors are often fixed using human labor. For small projects with limited resources, and large scale digitization efforts, any change in human labor required can have a significant effect.

Pre-developed specialized datasets help support research for real applications of OCR. The motivation behind Thomas Hegghammer’s study was his experience as a historian working with scanned documents [3]. His dataset built off of two existing datasets. Specialized datasets fill a niche needed for

this technology, while working within storage and processing constraints.

The increasingly specialized datasets created for recent OCR developments, serve to better understand and improve accuracy in their respective cases, but also serve a purpose in the wider OCR field. Many of the techniques developed for these cases, can also be applied to current applications of OCR. For example, the techniques designed to recognize cursive characters in Arabic, also work for many handwritten documents.

With the increase in OCR techniques that use neural networks, these datasets develop a second purpose, to train models. Due to the nature of the challenges presented in this paper, it is not realistic to create a dataset which contains every major document variation. With the development of these specialized datasets, comes tools that were designed to generate them. Between existing publicly available datasets, and these tools, it becomes incrementally easier to develop more datasets which cover the true diversity of documents. By understanding how the training of neural networks affects the capabilities of OCR models, we can understand the importance of intentionally introducing a variety of documents to the training process.

Acknowledgments

Thank you to my adviser, Elena Machkasova, and to my proofreader, Skatje Myers.

References

- [1] Ridvy Avyodri, Samuel Lukas, and Hendra Tjahyadi. 2022. Optical Character Recognition (OCR) for Text Recognition and its Post-Processing Method: A Literature Review. In *2022 1st International Conference on Technology Innovation and Its Applications (ICTIIA)*. 1–6. doi:10.1109/ICTIIA54654.2022.9935961
- [2] Amirreza Fateh, Mansoor Fateh, and Vahid Abolghasemi. 2024. Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection. *Engineering Reports* 6, 9 (2024), e12832. doi:10.1002/eng2.12832
- [3] Thomas Hegghammer. 2022. OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. *Journal of Computational Social Science* 5 (05 2022). doi:10.1007/s42001-021-00149-1
- [4] Herbert F Schantz. 1982. *The History of OCR, Optical Character Recognition*. Recognition Technologies Users Association.
- [5] tesseract ocr. 2025. tesseract-ocr/tesseract. <https://github.com/tesseract-ocr/tesseract>
- [6] Don Vaughan. 2025. The World's 5 Most Commonly Used Writing Systems. <https://www.britannica.com/list/the-worlds-5-most-commonly-used-writing-systems>
- [7] Wikipedia contributors. 2025. Arabic diacritics — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Arabic_diacritics&oldid=1317677291 [Online; accessed 26-October-2025].