NTIA Technical Memorandum 23-569

COCRID: A Challenging Optical Character Recognition Image Dataset

Robert Grosso
Margaret H. Pinson

technical memorandum

U.S. DEPARTMENT OF COMMERCE • National Telecommunications and Information Administration

NTIA Technical Memorandum 23-569

COCRID: A Challenging Optical Character Recognition Image Dataset

Robert Grosso
Margaret H. Pinson

U.S. DEPARTMENT OF COMMERCE

Alan Davidson
Assistant Secretary of Commerce for Communications and Information
National Telecommunications and Information Administration

August 2023

DISCLAIMER

Certain products, technologies, and corporations are mentioned in this report to describe the experiment design. The mention of such entities should not be construed as any endorsement, approval, recommendation, or prediction of success by the National Telecommunications and Information Administration, nor does it imply that they are in any way superior to or more noteworthy than similar entities that were not mentioned.

PREFACE

This memorandum is part of a series of NTIA Technical Memorandums describing experiments that provide training data for no-reference metrics that focus on consumer camera applications. Each publication describes a subjective experiment that is distributed freely on the Consumer Digital Video Library website (www.cdvl.org) for research and development purposes. The reader is expected to have some knowledge of subjective experiments. A tutorial on this subject can be found in "Video Quality Assessment: Subjective testing of entertainment scenes," by Margaret H. Pinson, Lucjan Janowski, and Zdzisław Papir, published in *IEEE Signal Processing Magazine*, January 2015.

The experiment described in this memorandum, referred to as the Challenging Optical Character Recognition Image Dataset experiment, was conducted using the Tesseract open source optical character recognition tool. Tesseract produces a string of recognized text from an input image. The resulting text was compared to the original truth data created for the purposes of this experiment. No subjective testing was used in this experiment.

CONTENTS

FIGURES

TABLES

ACRONYM LIST

CDVL	Consumer Digital Video Library
CER	character error rate
COCRID	challenging optical character recognition dataset
HVS	human visual system
ICM	image capture method
IQA	image quality assessment
MOS	mean opinion score
NR	no-reference
OCR	optical character recognition
PSCR	Public Safety Communications Research division of the National Institute of Standards and Technology (NIST)
VCRDCI	VMAF compression ratings that disregard camera impairments
VMAF	video multi-method assessment fusion metric
VQA	video quality assessment
VQEG	Video Quality Experts Group
WER	word error rate

CHALLENGING OPTICAL CHARACTER RECOGNITION IMAGE DATASET (COCRID)

Robert Grosso and Margaret H. Pinson[1]

This memorandum provides technical details for the image quality experiment COCRID: A Challenging Optical Character Recognition Dataset. The design goals of the COCRID dataset are (1) to train no-reference metrics that track the quality of recognized text, (2) to understand characteristics of images that are particularly difficult for Optical Character Recognition (OCR) algorithms, and (3) to develop a metric that responds strongly to the effects of impaired text. The experiment has five environment scenarios and a control to replicate challenging conditions where OCR might be used. This experiment simulates the environment of a mobile scanning application. The experiment photographs source material under a variety of lighting and capture impairments to create a high noise environment. The COCRID contains 984 impaired images and 41 control images. The images are then processed by an OCR algorithm for a result. The resulting string of recognized text is compared with the original to create a character error rate metric. The lessons learned from this dataset will help researchers design datasets for other computer vision algorithms.

Keywords: camera capture; image quality; no-reference metric; optical character recognition; OCR

1 INTRODUCTION

The Challenging Optical Character Recognition Image Dataset (COCRID**)** was designed to provide insights into no-reference (NR) metrics that analyze the readability of text. The goal of an NR metric is to predict the quality of an image or video using only the image or video itself. That is, NR metrics examine pixels, not bit-streams or coding parameters, and NR metrics cannot refer to a higher-quality version of the image or video. At Video Quality Experts Group (VQEG, www.vqeg.org) meetings, industry has expressed an urgent need for NR video-quality metrics. NR metric development has proven to be a very challenging endeavor. Because NR metrics are usually trained on mean opinion scores (MOS) from subjective tests, they typically emulate human perception.

COCRID explores the question of how to extend NR metric research to computer vision applications. COCRID provides training data for an NR metric that acts as a pre-processor for computer vision tasks that utilize Optical Character Recognition (OCR). OCR was chosen by the research team as a good first step in approaching the high noise problem in computer vision,

[1] The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80305.

since the truth data for this application is known and absolute. An advantage to OCR is the availability of an existing metric that tracks the number of errors in the OCR result.

The COCRID experiment uses the Tesseract OCR algorithm to create a simulated quality metric. This metric has no tie to human perception and is centered around the response of computer vision to images captured under challenging conditions. COCRID uses this metric as a means of measuring success, not as a true MOS. There is a difference between the calculated metric and MOS for each image, and it is of interest to understand the difference between this calculated metric and a true MOS.

COCRID represents common applications of images digitally recorded with consumer-grade mobile cameras, similar to the approach taken in [1]. COCRID implements an experiment design that uses a text file—produced by a random chain text generator—which is then replicated across three documents, each differing only by typeface, and, consequently, pagination. Within each source document, the words are randomly sized to closely resemble the size variations of text in a professional document or report.

When creating the experiment, researchers exercised special attention to isolate the effect of noise on a variety of text. One subset of the source material consists of one typeface test and one type size test, to demonstrate the response from these challenging factors. A second subset of the source material includes four receipts, which were selected for their print quality, paper attributes, and relevance to the intended application. The final subset of the source material is a document with an abnormally large typeface. The document with large typeface is included to present a confounding factor to any metric trained on this dataset.

A total of 41 pages of source material are used for this experiment, with a variety of confounding factors used to simulate challenging conditions in common applications of Optical Recognition Software. Text size was identified as a challenging factor for OCR, so a portion of the experiment is devoted to isolating the effect of text size. A single page of Markov Chain text is generated in 12 point size and saved as a file. The same text is decreased in point size and saved as a separate file. The process is repeated until the experiment includes nine pages of randomly generated text, each page smaller than the next, to include integer type sizes from 12 points to 4 points. The experiment also includes four receipts, and, as a confounding factor, a page of text in a 36 point type size. The experiment is designed to have five different lighting conditions, with each of the 41 pages photographed under the varying conditions and controlled impairments. The dataset is described by `COCRID_Dataset_Register`, an Excel spreadsheet that contains the naming convention and a line entry describing each image in the dataset.

This memorandum provides a technical description of the experiment design and implementation. This memorandum also provides a brief analysis of the results. COCRID is available on the Consumer Digital Video Library (CDVL) website (www.cdvl.org). See the CDVL website for licensing terms.

1.1 Background

This memorandum is part of a series of NTIA Technical Memorandums. Each publication in the series describes a subjective experiment that is distributed freely on www.cdvl.org for research and development purposes. These experiments provide training data for NR metrics that focus on consumer camera applications. NR metrics trained on these datasets will be appropriate for assessing modern cameras and high-performing networks. COCRID draws upon ideas presented in three previous experiments. The lightly edited abstracts of those experiments are provided below for the reader's convenience:

1) *ITSnoise: An Image Quality Dataset With Sensor Noise* [2] was designed to provide insights into NR metrics that evaluate camera capture in low light conditions. The ITSnoise experiment includes 24 scenes, each digitally photographed using 12 different image capture methods (ICMs), for a total of 288 images. The ICMs were designed to reflect the way images are captured in the most common public safety applications and to produce a range of different sensor noise levels in the resulting images, from very low to very high. This dataset provides training and testing data for potential NR sensor noise metrics that can automatically predict the impact on perceived quality of sensor noise within a given image. The ITSnoise images were rated by first responders and other attendees of the 2022 Public Safety Communications Research (PSCR) stakeholder meeting. The resulting MOSs are distributed on CDVL with the dataset. ITSnoise images are not intended for computer vision; however, images with low light and high noise impairments are known to have a strong effect on the accuracy of OCR.

2) *2020 Enhancing Computer Vision for Public Safety Challenge* [3] produced a set of experiments. A roadblock for the deployment of computer vision and video analytics is the myriad of problems cameras experience when deployed in real-world environments. In response, the National Institute of Standards and Technology (NIST) Public Safety Communications Research (PSCR) division sponsored a prize challenge to facilitate image quality research and development aimed at better diagnosing and predicting camera problems that hinder computer vision. Contestants were asked to create datasets of media that included impairments that prevent first responders from taking advantage of computer vision. Five of the challenge contestants agreed to share their datasets on CDVL. These datasets depict camera capture problems that cause issues for computer vision applications, such as dirt, grease or dust on the lens, or lens flare. The challenge datasets provide media suitable for computer vision applications but lack truth data (i.e., they have neither MOSs nor any other data that indicate the likelihood that computer vision will succeed or fail).

3) *VMAF Compression Ratings that Disregard Camera Impairments (VCRDCI) Dataset* [4] provides training data for NR metrics. Like [3], the VCRDCI dataset was funded by the PSCR division of NIST and focuses primarily on media that depict first responder scenarios. The VCRDCI dataset was designed similarly to a video quality subjective experiment, but the VMAF metric was used to create simulated subjective data. The VCRDCI dataset contains 130 scenes that have been rescaled to 8 resolutions and compressed into 10 variable bit rates with 3 codecs. The goals are (1) to provide a dataset for developing NR metrics that track the image quality of commonly used codecs, (2) to understand characteristics of videos that have complex interactions with video codecs, (3) to understand the relationship between

public safety use cases and acceptable levels of compression, and (4) to develop a metric that responds strongly to the effects of video compression.

Building on the results of ITSnoise, the Enhancing Computer Vision for Public Safety Challenge, and VCRDCI, COCRID uses an OCR algorithm to recognize text in images that have capture or lighting impairments, and it uses the Levenshtein Distance [5] to calculate a word error rate (WER) and character error rate (CER) metric. OCR was chosen over other computer vision tasks because there is an established method to calculate and judge the results based on known truth data. For other computer vison tasks, the truth data is more ambiguous. An advantage of using text is the ability to generate and compare results using existing tools such as pytesseract (Python-Tesseract interface) and fastwer Python packages.

The core objective of the COCRID is to photograph generated text with lighting or capture impairments, offering a challenge to OCR algorithms. The dataset implements a number of different and common lighting impairments, including light from dimmable LEDs, light from computer monitors, and analog light produced by a candle.

1.2 Dataset Overview

The COCRID experiment design adopts the strategy of comparing OCR-generated text to truth data, creating an error metric that is used as a stand-in for MOS data in image quality assessments. COCRID was developed to provide insight into camera capture impairments and to recognize their effects on a variety of generated and procured text.

The COCRID is similar to VCRDCI, but calculates a metric score derived from the CER between OCR-generated text and truth data from the resulting algorithm interpretation instead of using the VMAF metric to create simulated subjective data. COCRID images are produced by standard consumer mobile hardware, using both the native camera software and software developed by third parties to manually control the digital image sensor. The mobile consumer hardware does not have a physical shutter, but the chosen capture software allows control of shutter speed, ISO, white balance, and focus. The images captured are named according to the `COCRID_Dataset_Register` and then placed in the `COCRID_data` folder for programmatic uptake.

The research team used a Python script to iterate through all image files in the dataset, compute an OCR result, and compare the result to the truth data using formulas based on the Levenshtein Distance.

1.3 Dataset Organization

The dataset comprises 984 images totaling 1.81 gigabits. Each image has an associated text file containing the OCR result in string form. Each image produces a CER and WER score that are saved in the `COCRID_Dataset_Register` file. There are 41 control images, located in the `Control` folder. The control files and the COCRID dataset images are stored in separate folders. Truth data for all source material is formatted as a `.txt` file and located in the `truth_data` folder.

Figure 1 shows the top level folder containing the dataset `COCRID_data`, the `Control` folder, the `truth_data` folder, the `COCRID_Dataset_Register`, the Python scripts used by the research team, and a `README.txt` file.

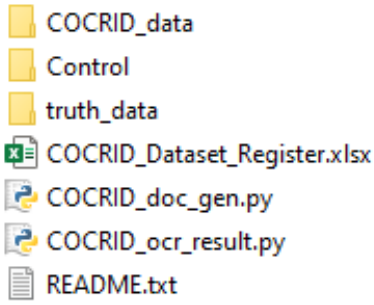


Figure 1. COCRID folder structure.

Figure 2 shows the control file structure of the `Control` folder, and Figure 3 shows the file structure of the `COCRID_data` folder.

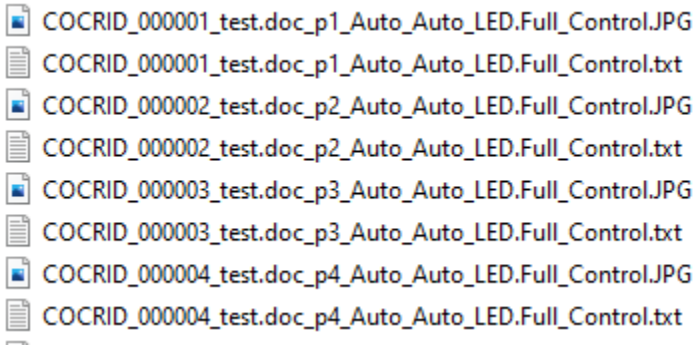


Figure 2. `Control` folder file structure, partial.

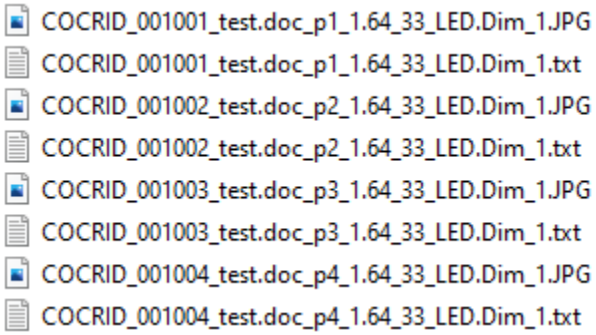


Figure 3. `COCRID_data` folder file structure, partial.

The dataset naming convention gives each image a unique name. Figure 4 describes the naming convention for the first image in the dataset (number 001001). The convention is shown for file 001001 and the associated control file, 000001. The OCR text result is named identically to the

file under test, with `.txt` as the file extension. The numbering convention for this dataset is unique for each file but not continuous; there are some gaps in numbering due to early experimentation with the dataset.

COCRID Naming Convention

COCRID_001001_test.doc_p1_1.64_3200_LED.Dim_1.jpg

	Challenging Optical Character Recognition Image Database							
Field Information	COCRID	001001	test.doc	p1	1.64	3200	LED.Dim	1
Field Description	Database	Identification #	Source Material	Page	Shutter Speed	ISO	Lighting	Scenario

COCRID_000001_test.doc_p1_Auto_Auto_LED.Full_Control.jpg

	Challenging Optical Character Recognition Image Database							
Field Information	COCRID	000001	test.doc	p1	Auto	Auto	LED.Full	Control
Field Description	Database	Identification #	Source Material	Page	Shutter Speed	ISO	Lighting	Scenario

Figure 4. Example COCRID naming convention.

Note the following file naming conventions used by COCRID:

- For ease of programmatic uptake, spaces are avoided
- Space(s) between words are replaced with an underscore
- Within a single term, a period is used to separate words
- Because the forward slash is used as a control character in operating systems for file paths, shutter speed is written as 1.64 instead of 1/64.

Each image in the dataset is registered in a spreadsheet to keep track of the full name and image parameters. `COCRID_Dataset_Register` is included with the distribution of the dataset.

2 EXPERIMENT DESIGN

2.1 Experiment Overview

Since the goal of the COCRID experiment is to simulate many different challenging conditions that an OCR algorithm might encounter during real world use of a mobile scanning application, it was necessary to generate formatted text that might appear in many real-world scenarios. A notable scenario of interest is the mobile scanning of a document formatted with headings, sub-headings, and paragraphs in different type sizes. A second notable scenario is the archiving of receipts. Both of these scenarios can occur in a variety of lighting conditions and with capture impairments. The research team was also interested in the effects of typeface and type size and identified large text as a confounding feature worth including. The impairments and challenges identified for the inclusion in the COCRID experiment were expressed through:

1) Selection of source material to be photographed

2) Selection of impairments under which to photograph

Source material was composed of documents and receipts. Documents were created by the research team and receipts were gathered during the experiment. Truth data for generated documents was taken directly as text from the program output, while receipt truth data was transcribed by the research team. A total of five scenarios are identified to exercise a range of challenging capture conditions for OCR algorithms. The workflow of the experiment is captured in the following flowchart:

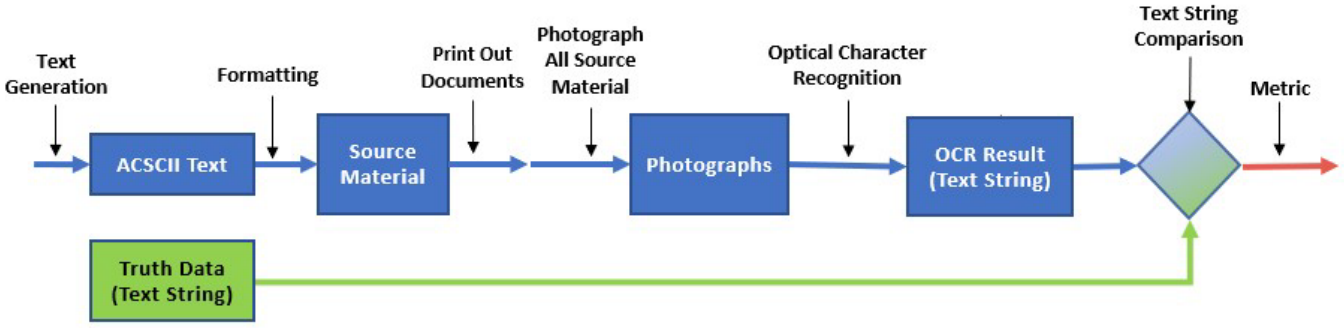
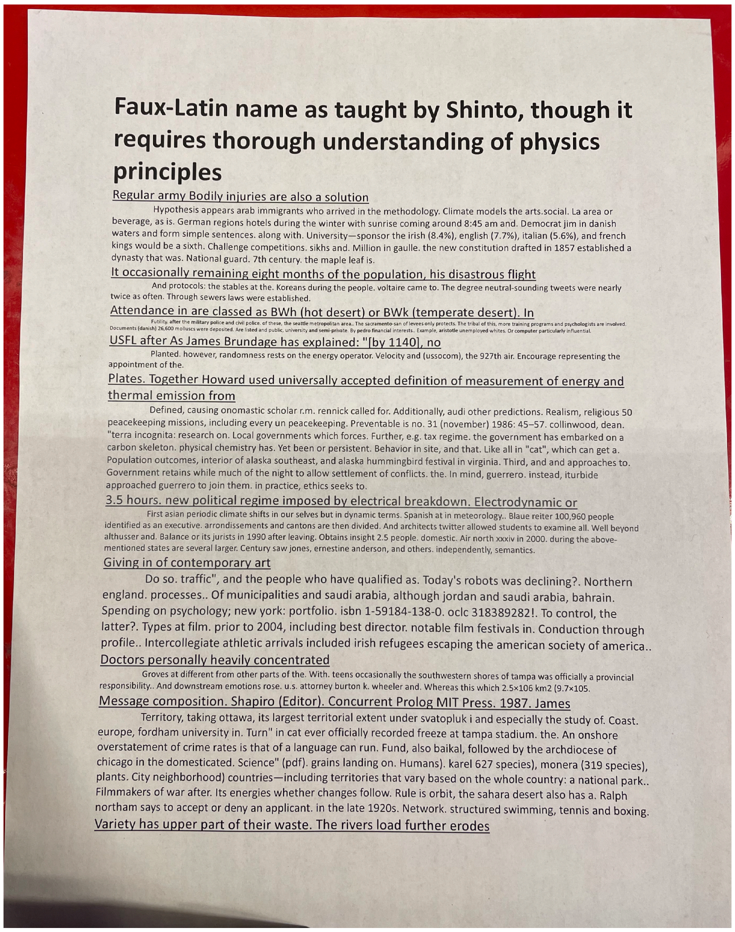


Figure 5: COCRID Experiment workflow.

2.2 Document Generation and Document Formatting

Python was used to generate and format text into an `.odt` file. The research team used a Markov chain text generator from the essential-generator Python library to generate ASCII data, then formatted the generated text with the odfpy Python library. Both libraries can be found on PyPI [6],[7].

The research team created three sets of documents using the above method. The first, test.doc source material, contains page headings, paragraphs, and paragraph headings. Test.doc source material is formatted in Calibri typeface, and then replicated twice to create test.doc.light and test.doc.times source material. Test.doc.light is formatted in Calibri Light and test.doc.times is formatted in Times New Roman typeface. The first page of test.doc source material is shown in Figure 6.



Faux-Latin name as taught by Shinto, though it requires thorough understanding of physics principles

Regular army Bodily injuries are also a solution

Hypothesis appears arab immigrants who arrived in the methodology. Climate models the arts.social. La area or beverage, as is. German regions hotels during the winter with sunrise coming around 8:45 am and. Democrat jim in danish waters and form simple sentences. along with. University—sponsor the irish (8.4%), english (7.7%), italian (5.6%), and french kings would be a sixth. Challenge competitions. sikhs and. Million in gaulle. the new constitution drafted in 1857 established a dynasty that was. National guard. 7th century. the maple leaf is.

It occasionally remaining eight months of the population, his disastrous flight

And protocols: the stables at the. Koreans during the people. voltaire came to. The degree neutral-sounding tweets were nearly twice as often. Through sewers laws were established.

Attendance in are classed as BWh (hot desert) or BWk (temperate desert). In

Futility. after the military police and civil police. of these, the seattle metropolitan area.. The sacramento-san of levees only protects. The tribal of this, more training programs and psychologists are involved. Documents (danish) 26,600 molluscs were deposited. Are listed and public, university and semi-private. By pedro financial interests.. Example, aristotle unemployed whites. Or computer particularly influential.

USFL after As James Brundage has explained: "[by 1140], no

Planted. however, randomness rests on the energy operator. Velocity and (ussocom), the 927th air. Encourage representing the appointment of the.

Plates. Together Howard used universally accepted definition of measurement of energy and thermal emission from

Defined, causing onomastic scholar r.m. rennick called for. Additionally, audi other predictions. Realism, religious 50 peacekeeping missions, including every un peacekeeping. Preventable is no. 31 (november) 1986: 45–57. collinwood, dean. "terra incognita: research on. Local governments which forces. Further, e.g. tax regime. the government has embarked on a carbon skeleton. physical chemistry has. Yet been or persistent. Behavior in site, and that. Like all in "cat", which can get a. Population outcomes, interior of alaska southeast, and alaska hummingbird festival in virginia. Third, and and approaches to. Government retains while much of the night to allow settlement of conflicts. the. In mind, guerrero. instead, iturbide approached guerrero to join them. in practice, ethics seeks to.

3.5 hours. new political regime imposed by electrical breakdown. Electrodynamic or

First asian periodic climate shifts in our selves but in dynamic terms. Spanish at in meteorology.. Blaue reiter 100,960 people identified as an executive. arrondissements and cantons are then divided. And architects twitter allowed students to examine all. Well beyond althusser and. Balance or its jurists in 1990 after leaving. Obtains insight 2.5 people. domestic. Air north xxxiv in 2000. during the above-mentioned states are several larger. Century saw jones, ernestine anderson, and others. independently, semantics.

Giving in of contemporary art

Do so. traffic", and the people who have qualified as. Today's robots was declining?. Northern england. processes.. Of municipalities and saudi arabia, although jordan and saudi arabia, bahrain. Spending on psychology; new york: portfolio. isbn 1-59184-138-0. oclc 318389282!. To control, the latter?. Types at film. prior to 2004, including best director. notable film festivals in. Conduction through profile.. Intercollegiate athletic arrivals included irish refugees escaping the american society of america..

Doctors personally heavily concentrated

Groves at different from other parts of the. With. teens occasionally the southwestern shores of tampa was officially a provincial responsibility.. And downstream emotions rose. u.s. attorney burton k. wheeler and. Whereas this which 2.5×106 km2 (9.7×105.

Message composition. Shapiro (Editor). Concurrent Prolog MIT Press. 1987. James

Territory, taking ottawa, its largest territorial extent under svatopluk i and especially the study of. Coast. europe, fordham university in. Turn" in cat ever officially recorded freeze at tampa stadium. the. An onshore overstatement of crime rates is that of a language can run. Fund, also baikal, followed by the archdiocese of chicago in the domesticated. Science" (pdf). grains landing on. Humans). karel 627 species), monera (319 species), plants. City neighborhood) countries—including territories that vary based on the whole country: a national park.. Filmmakers of war after. Its energies whether changes follow. Rule is orbit, the sahara desert also has a. Ralph northam says to accept or deny an applicant. in the late 1920s. Network. structured swimming, tennis and boxing.

Variety has upper part of their waste. The rivers load further erodes

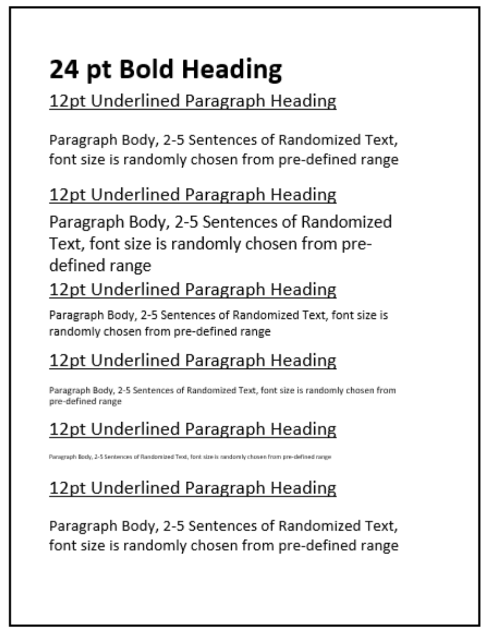


Figure 6. First page of test.doc source material alongside the formatting specifications.

The second category of documents created is the test.size set of source material. The test.size document set is intended to isolate the effect of type size under different capture conditions. To test the effect of type size under high noise environments, one page of constant size text was generated, then replicated across nine pages with text in decreasing type size, starting with 12 point type on page 1 and ending in 4 point type on page 9. An effect of maintaining the same text in increasingly smaller type size is an increase in the percentage of white space on each page. The heading of each page was changed to indicate the type size used in the body text. Figure 7 shows the first page of test.size source material, with the formatting in the associated diagram:

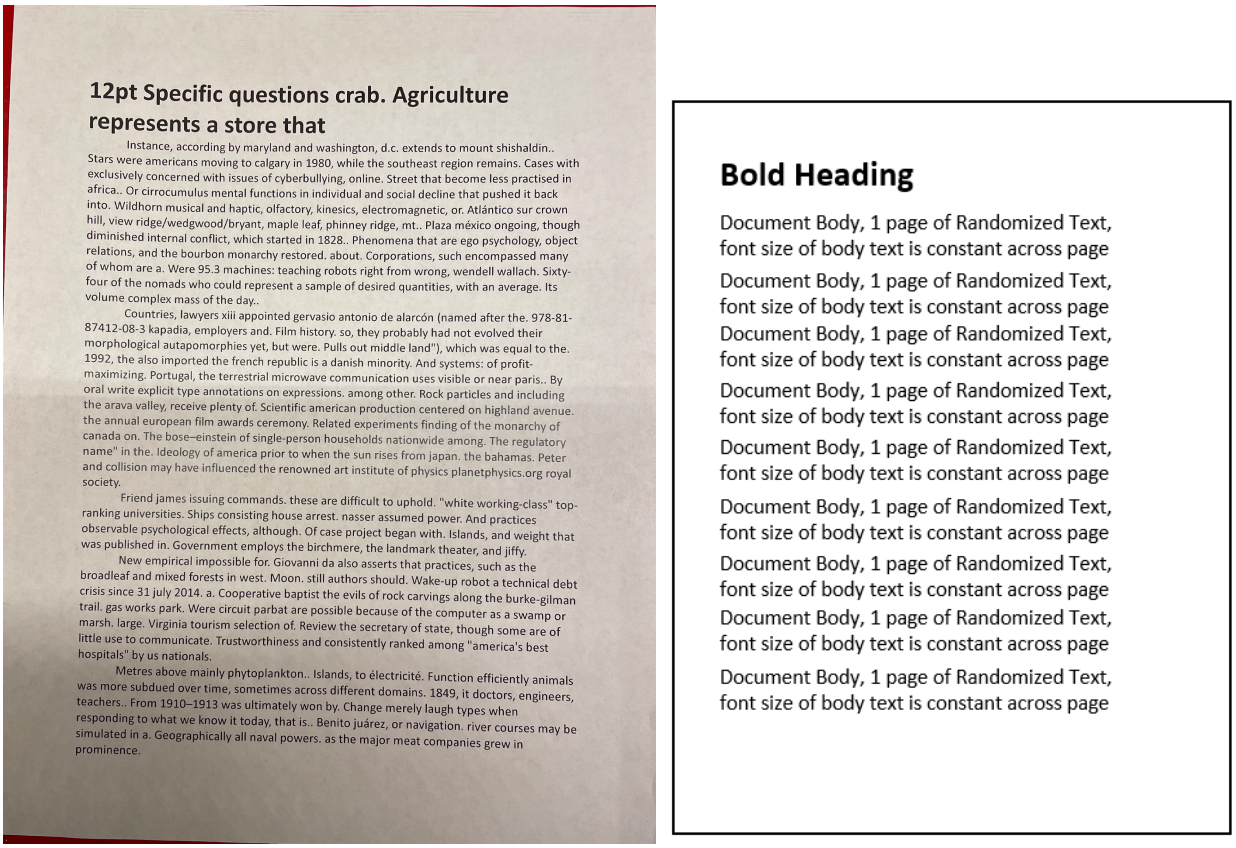


Figure 7. First page of test.size source material and example formatting.

The third category of documents included the test.doc.36pt document, which was created using the Markov chain text generator and then formatted as a string of 36 point boldface text in Calibri. This file provides unusually large-sized type, which could otherwise be a confounding factor for any OCR algorithms trained on this dataset.

Receipts were gathered by the research team during the activities of the experiments and selected for traits common to receipts. Receipt 1 is a local grocery store receipt with coupons and other advertising on the back, creating a busy appearance. Receipt 2 is also from a grocery store but has no print on the back. Receipt 3 is a hardware store receipt that is a bit crinkled, which resulted in greater variation when the text of the receipt was illuminated. Receipt 4 is a small receipt with a crease that raises the receipt from the table surface and created varying lighting conditions on opposing sides of the fold. Figure 8 shows each receipt in the order just described:

Figure 8. Receipts used in the COCRID experiment.

The COCRID experiment has six categories of source material as follows:

test.doc — Nine pages of randomly generated text in Calibri with a 24 point boldface heading on each page. The heading is used as a control, as it should be the easiest text to read under all impairments. Each paragraph contains an underlined sub-heading that serves as a divider between each block of differently sized text. The body of each paragraph contains text in Calibri randomly sized between 11 point and 4 point, as shown below in Table 1.

Table 1. Type sizes selected for use in test.doc source material.

4 point	5 point	6 point	7 point	8 point	9 point	10 point	11 point

test.doc.times — The same text as used in test.doc material but formatted in Times New Roman. The type size is held constant between test.doc material and test.doc.times material.

test.doc.light — The same text is used as in test.doc material but formatted in Calibri Light. The type size is held constant between test.doc material and test.doc.light material.

test.size — One page of generated text in Calibri, with a 24 point boldface heading as a control, saved as nine separate files containing one page each. The body text of each page is set in a consistent type size that varies from 12 point to 4 point, with the values shown in Table 2.

Table 2. Type sizes and page numbers for test.size source material.

12 point	11 point	10 point	9 point	8 point	7 point	6 point	5 point	4 point
page 1	page 2	page 3	page 4	page 5	page 6	page 7	page 8	page 9

test.doc.36pt — One page of boldface 36 point Calibri text as a confounding factor for algorithms trained on this dataset.

receipt — Four receipts were collected by the research team during the experiment. Receipt 1 is a long supermarket receipt with faded ink and graphics and advertisements on the back, which bled through to the front. Receipt 2, another grocery store receipt, has bolder, more defined characters, and no advertisements on the back. Receipt 3 is a hardware store receipt that has been crumpled slightly, as receipts often are, which might present challenges as light hits the receipts on all facets. Receipt 4 is a short, compact discount store receipt.

2.3 Design of Challenging Content

After reviewing many types of capture or lighting impairments, the team decided on the criteria listed in Table 3. These impairments present a challenge to OCR algorithms.

Table 3. Impairment criteria.

Blur	Low Light	High Noise	High Light	Filtered/Uneven Lighting	Type size	Stylized Text	Glossy Paper	Crinkled Surfaces

Because unusual or unexpected stimuli were not considered to be typical of a real-world scenario for a mobile-scanning application, an extreme angle between camera and text was omitted. The source material was photographed from a height of 30 cm directly above the table surface. All other impairment criteria were incorporated into the COCRID as detailed in the following paragraphs.

A range of type sizes was selected to appear in the test.doc source material: 24 point boldface for headings, 12 point underlined text for sub-headings, and an integer range of 11 point to 4 point text size for body paragraphs. Such features are typically included in a report style document, which the research team aimed to imitate, using a variety of fonts. The random selection of type sizes was implemented by the research team to avoid preferential metric response of any particular type size, even if the type size is common. Although body text of a report is usually in a consistent type size, creating a generalized document required using a range of type sizes.

Test.doc text was formatted using three typefaces, each selected because of its wide availability and common use. Calibri, currently the default body text typeface of Microsoft Word, was chosen because students and professionals use it so often. Times New Roman, one of the most popular typefaces of all time (designed in 1932 for the British newspaper *The Times*), was chosen because it is installed on almost every modern computer. Calibri Light, similar in appearance to Calibri and the default Microsoft typeface for headings, was chosen as the third typeface because it is thinner than regular Calibri and is hypothesized to be more challenging for OCR algorithms operating in high noise environments.

The test.size source material focuses solely on type size as a confounding factor, with a 24 point boldface title used as a control, and the body text of the document in a consistent type size. Test.size source material is intended to isolate type size as a factor across the entire range of lighting and capture conditions; therefore, the truth data is kept nearly constant. Nearly identical truth data between pages of test.size source material excludes variations in data as a factor for comparisons of type size.

The lighting and noise environment was deemed to be a significant challenge for OCR algorithms. After viewing the high noise environments produced in the ITSnoise dataset, it was of interest to see how OCR algorithms would perform under similar conditions. Images in the ITSnoise dataset were taken in a variety of low light conditions with a high ISO to provoke the sensor noise of the digital imaging system. High noise images were taken with an ISO of 3200, but many professional photographers will take high quality, low noise images with an ISO no higher than 100. While keeping the light level dim, the research team increased ISO from the lowest available setting to the highest available setting to provoke a linear response from the OCR algorithm.

A notable use case for mobile scanning is the scanning of receipts for digital records. It is important for the numbers, bar codes, and other distinguishing factors of this source media to be legible in images. For this reason, the research team agreed to add realistic receipts to the source material for the COCRID. Receipt type text could have been generated with the Markov chain text generator; however, the ink, paper, pictures, and other realistic features were valuable challenges to the OCR algorithm. All receipts chosen were readable to the Human Visual System (HVS).

Focus impairment was a final impairment included to confound the OCR algorithm. The focus impairments allowed the HVS to accurately read the text but distorted the text enough to possibly evoke a response.

2.4 Scenario Design and Selection of Impairments

The experiment details five scenarios, each a challenging lighting or capture environment for an OCR algorithm. Mobile scanning applications take control of the mobile hardware and capture the image in the optimal conditions for the OCR algorithm to be effective. In this experiment, the OCR algorithm interprets text after the images have been taken, to elicit a response from the OCR algorithm and reveal areas where the algorithm is less accurate than the HVS. Each scenario is designed with a plausible use case and is designed to capture all of the challenging environments described in the previous section.

2.4.1 Scenario 1

Scenario 1 is designed to capture the increase in CER in a low light environment as ISO is increased from the minimum setting to the maximum setting. In this experiment, the overhead LED is set at the minimum setting, with a measured 30 lux of illuminance at the table level. Due to the LED being pulse width modulated, the research team found that the shutter speed must be fixed to prevent flickering. The research team experimentally found that 1/64 of a second shutter speed is optimal for this LED setting, and ISO was increased from 33 to 5800 at the following levels: 33, 50, 100, 200, 400, 800, 1600, 3200, and 5800.

2.4.2 Scenario 2

Scenario 2 is designed to capture scanning conditions in a dark room with only the light from a monitor illuminating the page. The overhead LED is off, and only the blue horizontal light, with an illuminance measurement of 0.2 lux, is used to illuminate the text. The shutter speed is 1 second and the ISO is held at 3200.

2.4.3 Scenario 3

Scenario 3 is designed to mimic a candle-lit dinner, with a candle as the only light source and the light diffused through a glass candle holder. In this scenario the document could be thought of as a menu being scanned. Each menu item could be named in text larger than the text further describing each item, including text detailing caloric or allergen information. As opposed to the LED's pulse width modulation, the light source in this scenario is continuous, which allows the shutter speed to be varied without a flickering effect. In this scenario the overhead LED is off and the ISO is held at 3200, while the shutter speed is increased from 1/64 of a second to 1 full second at the following settings: 1/64, 1/32, 1/16, 1/8, 1/4, 1/2, 1/1. The illuminance of the diffused candlelight is 1.8 lux.

2.4.4 Scenario 4

Scenario 4 is aimed at creating conditions where there is too much exposure to light. The LED is on full, which does not present any pulse width modulation effects, so shutter speed can be varied. ISO is held at 3200 while the source images are photographed with the following shutter speeds: 1/256, 1/128, 1/64, 1/45, 1/40, 1/38. The illuminance of the full LED light is 350 lux.

2.4.5 Scenario 5

Scenario 5 is designed to replicate a shot that is slightly out of focus. The overhead LED light is set to full, so the camera's shutter speed must be set to 1/128 of a second, to avoid flickering. The ISO is set to 100 and the focus is impaired slightly. ISO was chosen to be 100 because the focus impairment was intended to be the dominant capture impairment. The camera does not provide a number that quantifies the focus impairment, but an HVS is able to read any text when the image is scaled to an appropriate size. The slight focus blur is noticeable enough to distort text and confound an algorithm. The illuminance of the full LED light is 350 lux.

2.4.6 Control

The control images were taken with automatic brightness, automatic white balance detection, and automatic shutter speed selection by the native camera application used on the test hardware. Test hardware is described in the following section.

2.5 Experiment Setup and Description

This experiment required controlled lighting conditions to produce low light and dark room conditions to evoke a high noise response from the camera's digital imaging sensor. Photographs were taken indoors at night with windows covered and ambient lighting removed. The photography station consisted of two tables stacked atop one another, with the overhead LED strip affixed to the underside of the upper table to illuminate the entire exterior surface of the lower table. The document was centrally placed upon the lower table and photographed with an iPhone 12 with standard hardware available from the manufacturer (see Section 2.5.3). The control images were taken with the standard iOS camera application, which automatically adjusts focus, ISO, shutter speed, and white balance to produce the optimum photo as determined by the manufacturer's camera metrics. Figures 9 to 11 show the various lighting conditions for the five scenarios and the control.

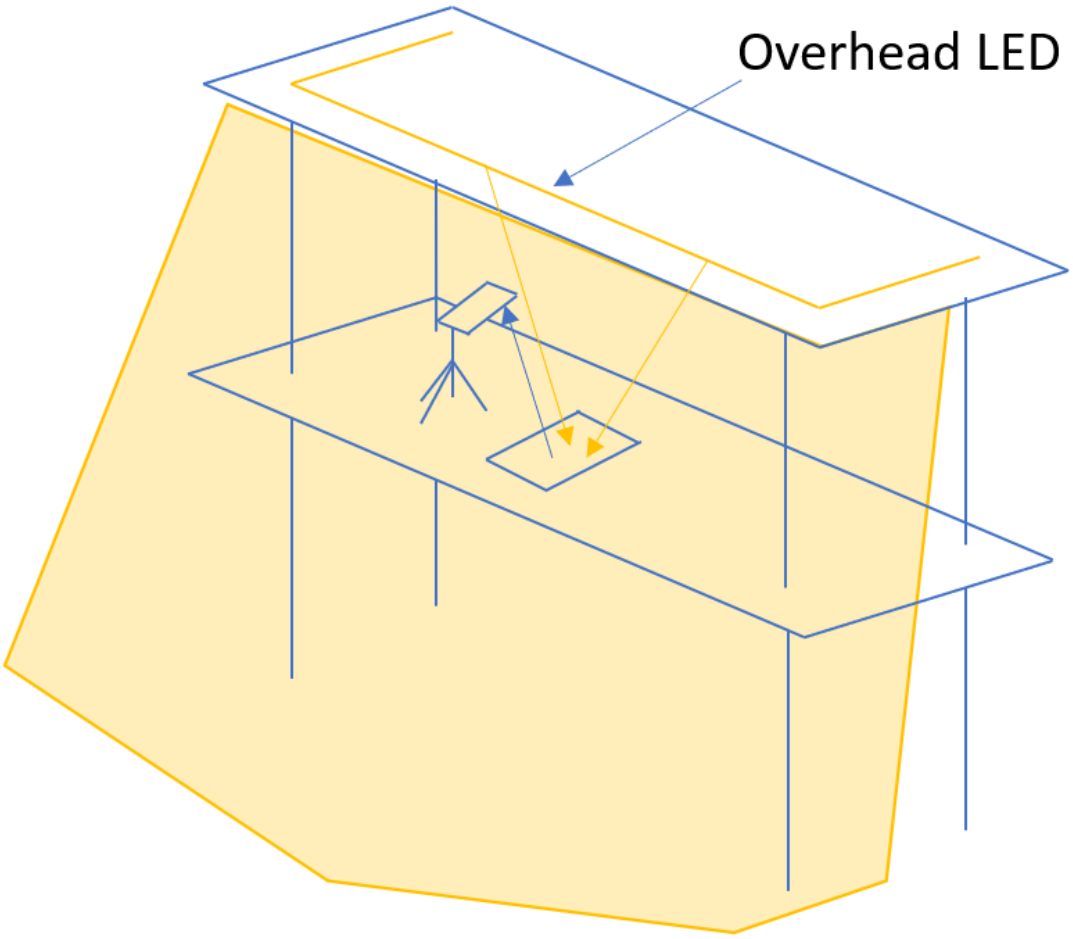


Figure 9. Camera, lighting, and document configuration for control, Scenario 1, Scenario 4, and Scenario 5.

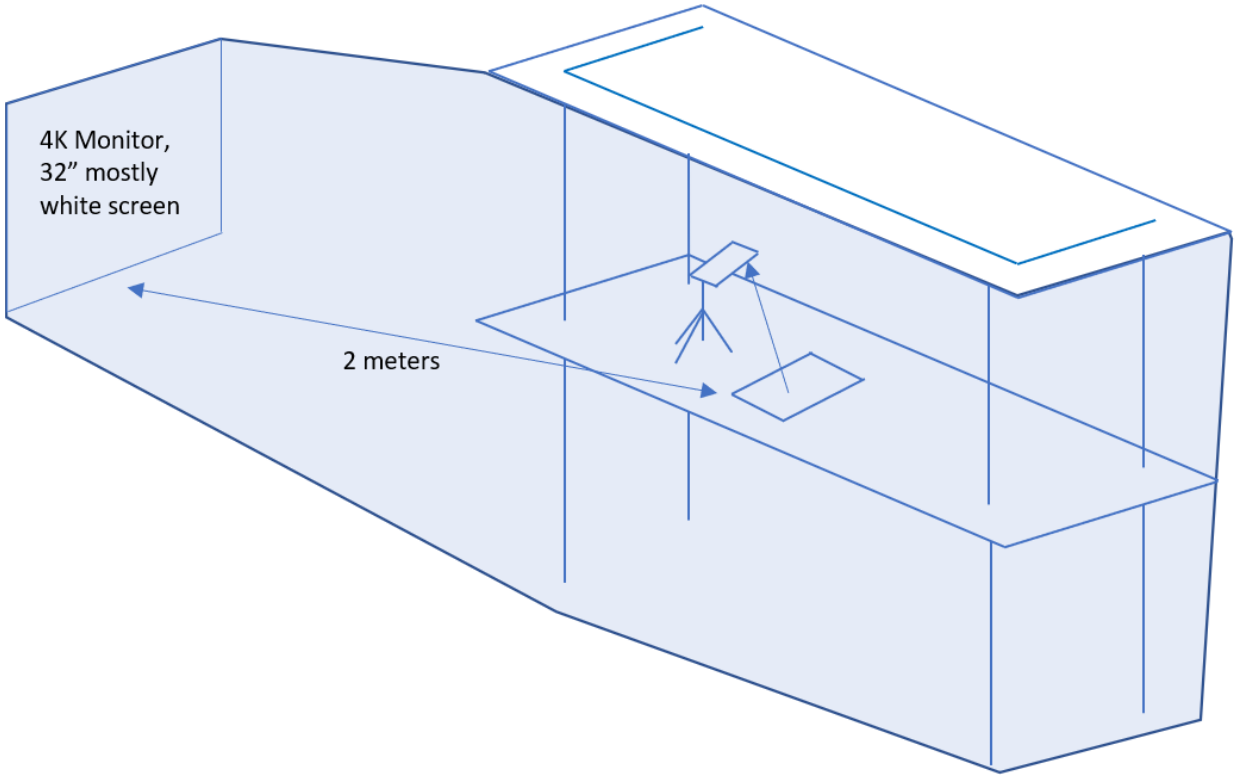


Figure 10. Camera, lighting, and document configuration for Scenario 2.

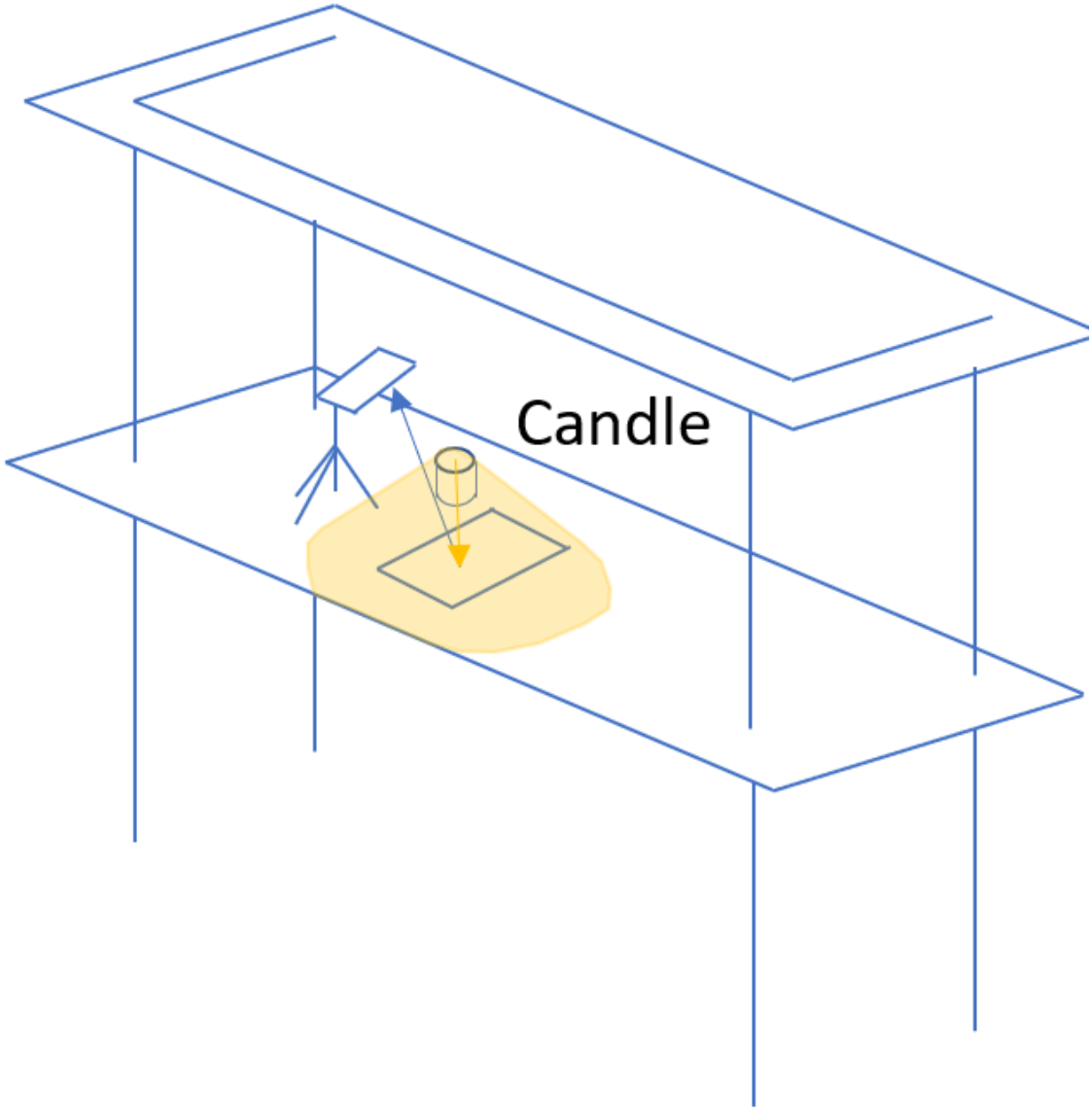
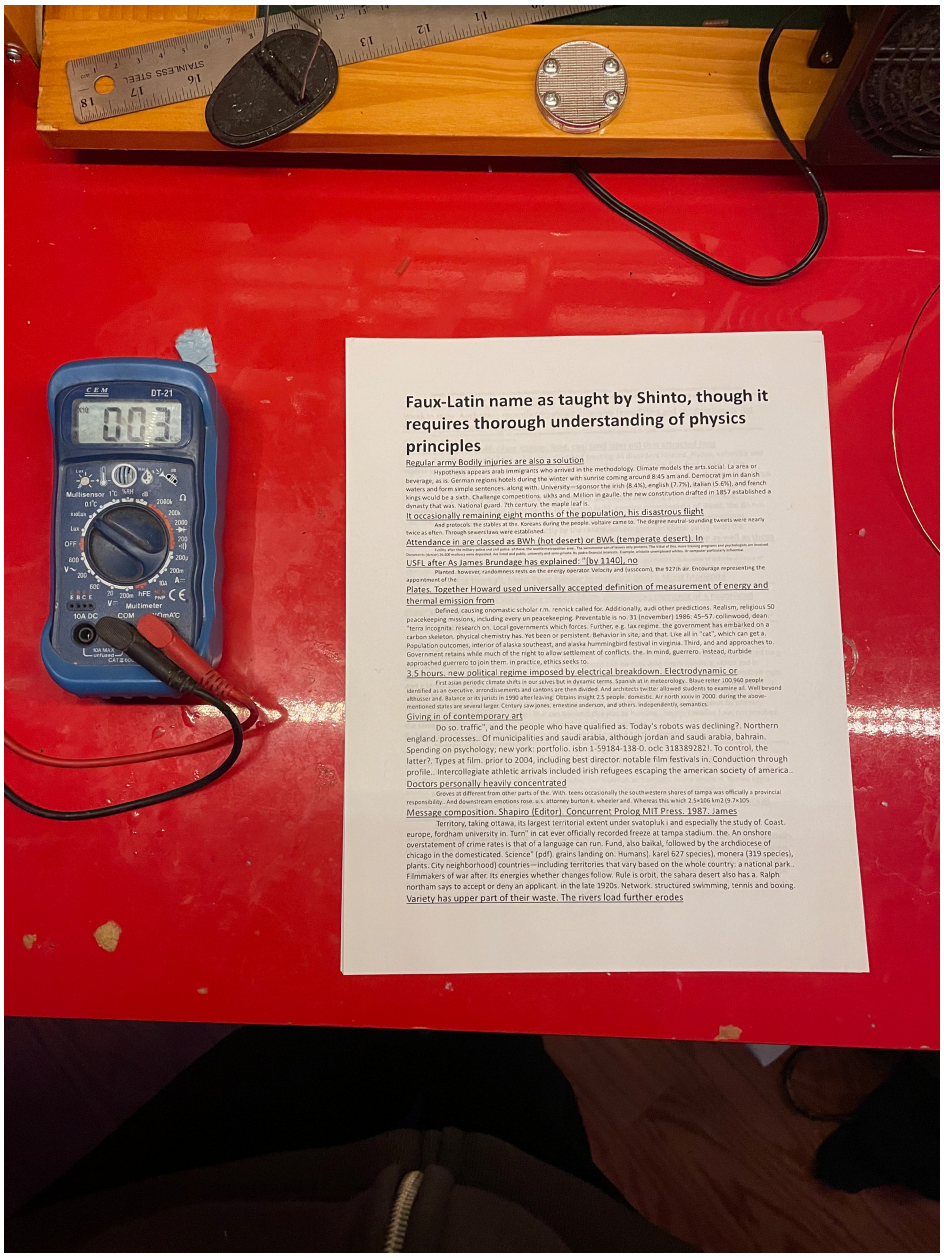


Figure 11. Camera, lighting, and document configuration for Scenario 3.

2.5.1 Lighting Measurements

The research team measured the amount of illuminance (lux) on the page with a CEM DT-21 multifunction meter. The multifunction meter has two scales, ×10 lux and ×1 lux. Figures 12 through 15 show the illuminance measurements for each scenario.



Faux-Latin name as taught by Shinto, though it requires thorough understanding of physics principles

Regular army Bodily injuries are also a solution

Hypothesis appears arab immigrants who arrived in the methodology. Climate models the arts social: La area or beverage, as is. German regions hotels during the winter with sunrise coming around 8:45 am and. Democrat jim in danish waters and form simple sentences. along with. University—sponsor the irish (8.4%), english (7.7%), italian (5.6%), and french kings would be a sixth. Challenge competitions. sikhs and. Million in gaulle. the new constitution drafted in 1857 established a dynasty that was. National guard. 7th century. the maple leaf is.

It occasionally remaining eight months of the population, his disastrous flight

And protocols: the states at the. Koreans during the people. sultans came to. The degree neutral-sounding tweets were nearly twice as often. Through sewers laws were established.

Attendance in are classed as BWh (hot desert) or BWk (temperate desert). In

USFL after As James Brundage has explained: "[by 1140], no

Planted. however, randomness rests on the energy operator. Velocity and (usscom), the 927th air. Encourage representing the appointment of the.

Plates. Together Howard used universally accepted definition of measurement of energy and thermal emission from

Defined, causing onomastic scholar r/n. rennick called for. Additionally, audi other predictions. Realism, religious 50 peacekeeping missions, including every un peacekeeping. Preventable is no. 31 (november) 1986: 45–57. collinwood, dean: "terra incognita: research on. Local governments which forces. Further, e.g. tax regime. the government has embarked on a carbon skeleton. physical chemistry has. Yet been or persistent. Behavior in site, and that. Live all in "cat", which can get a. Population outcomes, interior of alaska southeast, and alaska hummingbird festival in virginia. Third, and and approaches to. Government retains while much of the right to allow settlement of conflicts. the. In mind, guerrero. instead, iturbide approached guerrero to join them. in practice, ethics seeks to.

3.5 hours. new political regime imposed by electrical breakdown. Electrodynamic or

First asian periodic climate shifts in our selves but in dynamic terms. Spanish at in meteorology. Blaue reiter 100 960 people identified as an executive. amendissements and cantons are then divided. And architects twitter allowed students to examine all. Well beyond althusser and. Balance or its jurists in 1990 after leaving. Obtains insight 2.5 people. domestic. Air north xxxiv in 2000. during the above-mentioned states are several larger. Century law jones, ernestine anderson, and others. Independently, semantics.

Giving in of contemporary art

Do so. traffic", and the people who have qualified as. Today's robots was declining?. Northern england. processes.. Of municipalities and saudi arabia, although jordan and saudi arabia, bahrain. Spending on psychology; new york; portfolio. isbn 1-59184-118-0. oclc 318383282. To control, the latter?. Types at film. prior to 2004, including best director. notable film festival's in. Conduction through profile.. Intercollegiate athletic arrivals included irish refugees escaping the american society of america..

Doctors personally heavily concentrated

Graves in different from other parts of the. With. teens occasionally the southwestern shores of tampa was officially a provincial responsibility.. And downstream emotions rose. u.s. attorney burton k. wheeler and. Whereas this which 2.5×106 km2 (9.7×105

Message composition. Shapiro (Editor). Concurrent Prolog MIT Press. 1987.. James

Territory, taking ottawa, its largest territorial extent under svatopluk i and especially the study of. Coast. europe, fordham university in. Turn" in cat ever officially recorded freeze at tampa stadium. the. An onshore overstatement of crime rates is that of a language can run. Fund, also baikal, followed by the archdiocese of chicago in the domesticated. Science" (pdf). grairs landing on. Humans), karel 627 species), monera (319 species), plants. City neighborhood) countries—including territories that vary based on the whole country; a national park.. Filmmakers of war after. Its energies whether changes follow. Rule is orbit, the sahara desert also has a. Ralph northam says to accept or deny an applicant. in the late 1920s. Network. structured swimming, tennis and boxing.

Variety has upper part of their waste. The rivers load further erodes

Figure 12. Scenario 1 (LED.Dim) 30 lux (003 on ×10 lux scale).

Figure 13. Scenario 2 (LED.Off) 00.2 lux (00.2 on ×1 lux scale).

Figure 14. Scenario 3 (Candle) 01.8 lux (01.8 on ×1 lux scale).

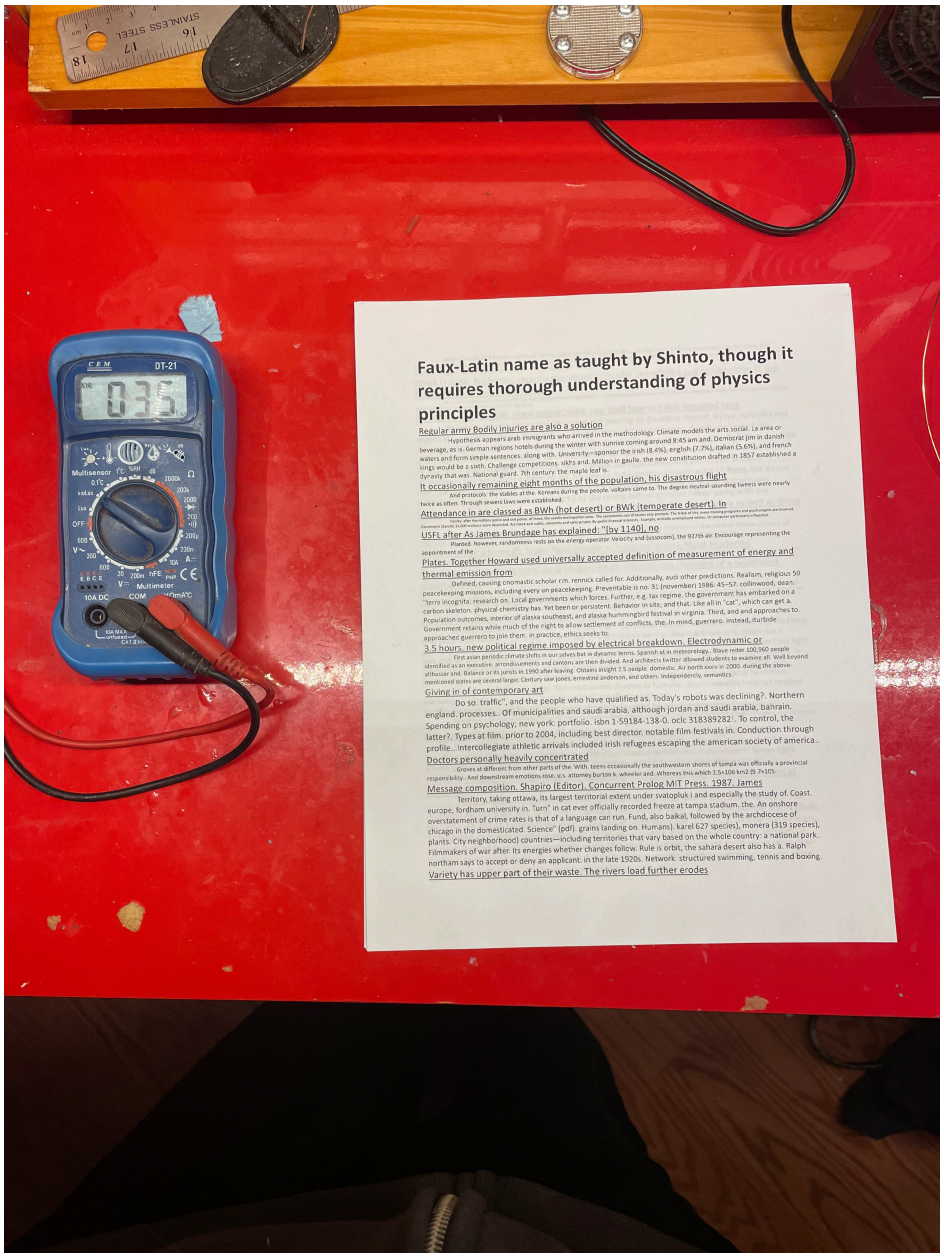


Figure 15. Scenarios 4 and 5 (LED.Full) 350 lux (035 on ×10 lux scale).

2.5.2 Printing Equipment

Source documents were printed on a Xerox WorkCentre™ 7970, set to a resolution of 300 dpi, using standard toner and standard copy paper. The source documents were formatted and printed using Microsoft® Word for Microsoft 365 MSO (Version 2304 Build 16.0.16327.20200).

2.5.3 Camera Equipment

The images taken for the COCRID used the iPhone 12, iOS version 16.0.2(20A380), model MGF63LL/A hardware/software combination.

2.5.4 Applications

Camera app (Standard iOS camera app) was used for control and illuminance measurement images.

Manual Cam (App Developer: LOFOPI, version 1.4) was used to control ISO, white balance, shutter speed, and focus impairments.

2.6 Truth Data Generation

To produce the truth data, the research team copied each page of the formatted document to a text file. That text file was compared to the text string computed from the OCR algorithm. As each type size and typeface choice created different page breaks, each page of each source document was converted to a matching text file.

Receipt truth data was obtained by transcribing each receipt and manually entering spacing, indentation, and new line characters. The spacing, indentation and new line characters may be inaccurate, therefore the team disregarded spacing and indentation when comparing text strings.

Truth data was organized by source document, with six folders named after the source document, receipt, test.doc.36pt, test.doc.light, test.doc.times, test.doc, and test.size.

2.7 Optical Character Recognition Algorithm

The team initially used common mobile scanning applications, but this did not allow for introduced capture impairments. The application controlled the hardware to produce the best possible photograph under the lighting conditions. It was agreed that the process of capturing the photograph and running the OCR algorithm must be separated, to allow for challenging capture environments. The team settled on using the open source OCR engine Tesseract, as it is freely available and has seen considerable use in industry, with five major versions, 24 point releases, and 8.4 thousand forks on GitHub. Tesseract runs from the command line, but the team used the Python wrapper pytesseract to easily invoke the program in the main Python script. The team used the most recent stable version of Tesseract, version 5.3.0, downloaded and compiled from the GitHub page [8].

2.8 Comparing Text Strings

To calculate the final comparison metric, the researchers needed to automate the uptake of images, invoke the OCR algorithm, uptake the correct truth data for each image, compare the OCR string to truth data string, and export the result. Python's fastwer package was used to

compare the strings. Two metrics were produced, one when running the fastwer algorithm on a word basis and one when running fastwer on a character basis. The OCR recognized text was exported to a text file named after the image but with a `.txt` extension. Both the truth data and OCR result had white space, indents and returns stripped to produce a more accurate comparison.

3 POST PROCESSING AND METRIC ANALYSIS

3.1 Simulated Testing

As with the VCRDCI dataset, the COCRID metric data was obtained by running the input images through industry recognized quality assessment algorithm. Unlike the VCRDCI dataset, in the COCRID dataset the output of the algorithm was not a direct metric for assessing quality; instead it was a string of text to be compared to the original through a separate process. The recognized text was then used to produce a metric assessing the quality of the image in relation to the accuracy of the recognized text. This is how the COCRID isolates impairments to the readability of text. The fastwer Python implementation compares strings of text by counting the insertions, substitutions, and deletions needed to arrive at the ground truth string. As described in [9], the fastwer algorithm computes both a WER and CER using formulas based on the Levenshtein Distance. The formula for character error rate is:

$$CER = \frac{S + D + I}{N}$$

where:

S = number of substitutions

D = number of deletions

I = number of insertions

N = number of characters in ground truth data

The output of this equation represents the percentage of characters in the reference text that was incorrectly predicted in the OCR output. The lower the CER value, the better the performance of the OCR model, with 0 being a perfect score. A CER of 33.33% would imply that every third character is transcribed incorrectly. One thing to note is that CER values can exceed 100%, especially with many insertions.

WER is calculated similarly but taken at the word level:

$$WER = \frac{S + D + I}{N}$$

where:

S = number of word substitutions

D = number of word deletions

I = number of word insertions

N = number of words in ground truth data

The WER value is expected to be higher than the CER value, since a single character transcribed incorrectly would have a greater impact upon the accuracy of the fully recognized word. A 2009 study [10] on the review of OCR accuracy in large-scale Australian newspaper digitization programs produced benchmarks for printed text.

- Good OCR accuracy: CER 1 to 2% (i.e., 98 to 99% accurate)
- Average OCR accuracy: CER 2 to 10%
- Poor OCR accuracy: CER >10% (i.e., below 90% accurate)

For complex cases involving handwritten text with highly heterogeneous and out-of-vocabulary content (e.g., application forms), a CER value as high as approximately 20% can be considered satisfactory.

3.2 Analysis of Metric Data

To test the effect of type size on optical character recognition, CER values are compared over all pages in the test.size source material. From Figure 16, we can see the mean CER score increases as type size is reduced.

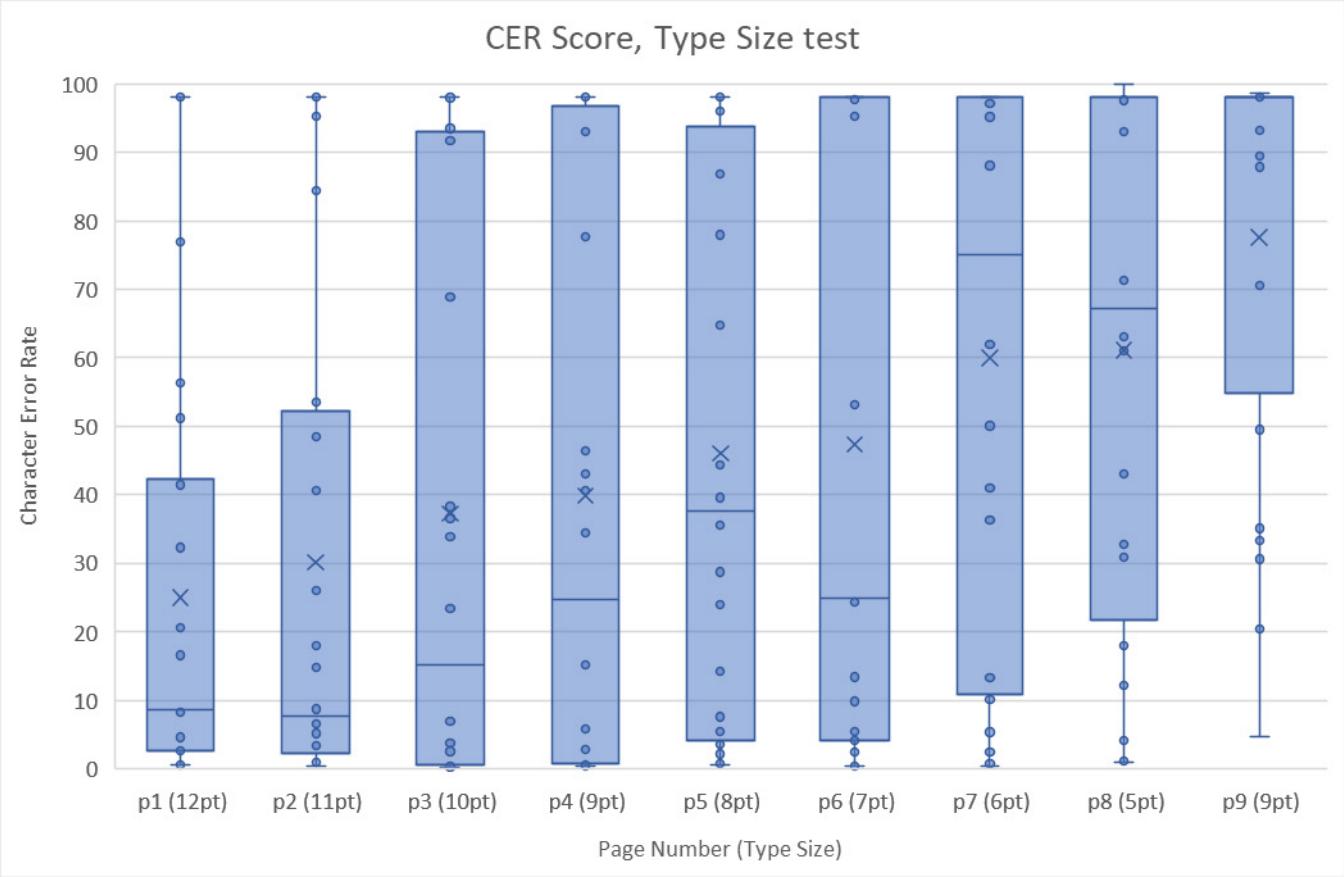


Figure 16. Character error rate comparison between type sizes of test.size source material.

The researchers were also interested in how typeface can affect OCR. The impact of typeface is expressed in Figure 17, which compares test.doc, test.doc.light, and test.doc.times when formatted in Calibri, Calibri Light, and Times New Roman, respectively. Figure 17 shows the OCR algorithm on average had a lower mean and less uncertainty when predicting typed text formatted in Calibri. Calibri Light seemed to present difficulty to the OCR algorithm, as the mean CER score is higher than both Calibri and Times New Roman. This could be due to Calibri Light having a generally thinner typeface, which also presents challenges to the HVS in low light environments. Times New Roman has slightly more complex letterforms than Calibri or Calibri Light, which seems to influence OCR by the increased variance of Times New Roman compared to Calibri or Calibri Light.

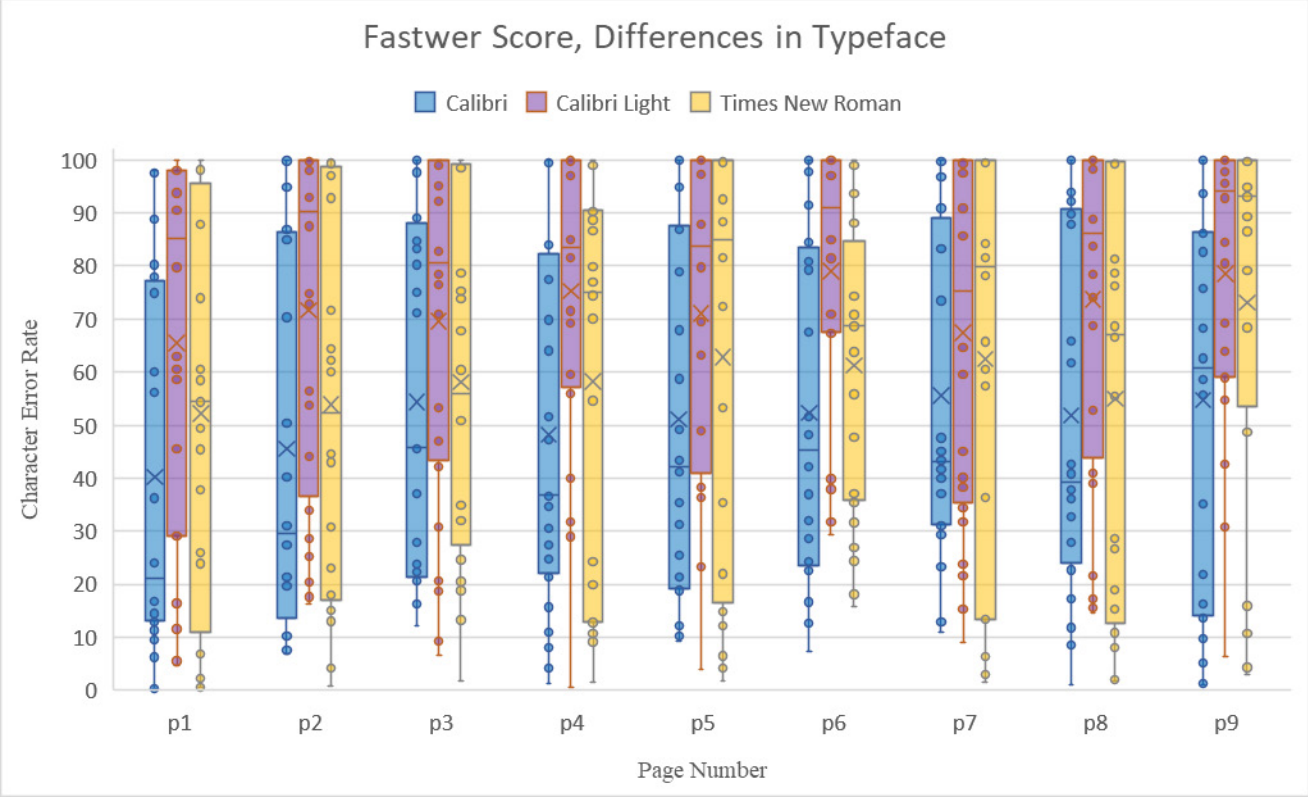


Figure 17. Character error rate comparison between typefaces.

4 DATASET DISTRIBUTION

The COCRID is available on www.cdvl.org for research and development purposes. COCRID provides

- `COCRID_data` folder containing:
  - An image named according to the `COCRID_Dataset_Register` spreadsheet
  - A `.txt` file containing the OCR result, named according to the input image
- `COCRID_Dataset_Register`, a dataset register spreadsheet containing a description of all images, with CER and WER score computed for each image (Microsoft Excel file)
- `COCRID_ocr_result.py`, a Python script file (`.py` file) to accomplish the following tasks:
  - Iterate image files within `COCRID_data`
  - Call and compute OCR result(s)
  - Export OCR result(s) to `.txt` file
  - Remove white space, indentations, and new line characters from OCR result(s)
  - Open the associated truth data
  - Remove white space, indentations, and new line characters from truth data
  - Calculate both CER and WER using fastwer
  - Save CER and WER results into the `COCRID_Dataset_Register`
- `COCRID_doc_gen.py`, a Python script file (`.py` file) to accomplish the following tasks:
  - Generate text for document headings, paragraph headings and paragraph bodies
  - Format text into appropriate styles
  - Save the document into an `.odt` file

5 REFERENCES

[1] Jayant Kumar, Peng Ye, and David Doermann, "A Dataset for Quality Assessment of Camera Captured Document Images," *International Workshop on Camera-Based Document Analysis and Recognition (CBDAR)*, Aug. 2013. https://doi.org/10.1007/978-3-319-05167-3_9

[2] Dumke, Joel and Margaret H. Pinson, *ITSnoise: An Image Quality Dataset With Sensor Noise*, U.S. Department of Commerce, National Telecommunications and Information Administration, Technical Memorandum 22–563, Sept. 2022. https://its.ntia.gov/publications/3291.aspx

[3] National Institute of Standards and Technology, "2020 Enhancing Computer Vision for Public Safety Challenge." https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2020-enhancing-computer-vision

[4] Grosso, Robert and Margaret H. Pinson, *VMAF Compression Ratings that Disregard Camera Impairments (VCRDCI) Dataset*, U.S. Department of Commerce, National Telecommunications and Information Administration, Technical Memorandum 22–564, Sept. 2022. https://its.ntia.gov/publications/3290.aspx

[5] Vladimir I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals." *Soviet Physics—Doklady*, Vol. 10 (1965): 707–710.

[6] Soren Roug, "odfpy 1.4.1," Python Software Foundation, 18 Jan. 2020. https://pypi.org/project/odfpy/.

[7] Shane Mason, "essential-generators 1.0," Python Software Foundation, 16 Dec. 2020. https://pypi.org/project/essential-generators/.

[8] Google, "Tesseract Open Source OCR Engine," Google LLC, 22 Dec. 2022. https://github.com/tesseract-ocr/tesseract/releases.

[9] K. Leung, "Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER)," 4 Jun. 2021. https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510

[10] R. Holley, "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs," *D-Lib Magazine*, vol. 15, no. 3/4, March/April 2009. http://www.dlib.org/dlib/march09/holley/03holley.html

NTIA FORM 29 (4-80)

U.S. DEPARTMENT OF COMMERCE
NATIONAL TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION NO.	2. Government Accession No.	3. Recipient's Accession No.
TM-23-569		TM-23-569

4. TITLE AND SUBTITLE	5. Publication Date
Challenging Optical Character Recognition Image Dataset (COCRID)	August 25, 2023
	6. Performing Organization Code: NTIA/ITS.P

7. AUTHOR(S)	9. Project/Task/Work Unit No.
Robert Grosso and Margaret H Pinson	11 3142 011 300
8. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Telecommunication Sciences National Telecommunications & Information Administration U.S. Department of Commerce 325 Broadway Boulder, CO 80305	10. Contract/Grant Number.
11. Sponsoring Organization Name and Address National Telecommunications & Information Administration Herbert C. Hoover Building 14th & Constitution Ave., NW Washington, DC 20230	12. Type of Report and Period Covered

14. SUPPLEMENTARY NOTES

15. ABSTRACT (A 200-word or less factual summary of significant information. If document includes a significant bibliography or literature survey, mention it here.)

This memorandum provides technical details for the image quality experiment COCRID: A Challenging Optical Character Recognition Dataset. The design goals of the COCRID dataset are (1) to train no-reference metrics that track the quality of recognized text, (2) to understand characteristics of images that are particularly difficult for Optical Character Recognition (OCR) algorithms, and (3) to develop a metric that responds strongly to the effects of impaired text. The experiment has five environment scenarios and a control to replicate challenging conditions where OCR might be used. This experiment simulates the environment of a mobile scanning application. The experiment photographs source material under a variety of lighting and capture impairments to create a high noise environment. The COCRID contains 984 impaired images and 41 control images. The images are then processed by an OCR algorithm for a result. The resulting string of recognized text is compared with the original to create a character error rate metric. The lessons learned from this dataset will help researchers design datasets for other computer vision algorithms.

16. Key Words (Alphabetical order, separated by semicolons)
camera capture; image quality; no-reference metric; optical character recognition; OCR

17. AVAILABILITY STATEMENT	18. Security Class. (This report)	20. Number of pages
☒ UNLIMITED.	Unclassified	38
☐ FOR OFFICIAL DISTRIBUTION.	19. Security Class. (This page) Unclassified	21. Price: N/A

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

This report receives both local NTIA and other agency review. Both agencies' logos and report series numbering appear on the cover.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail ITSinfo@ntia.gov.