**GREGG SHORTHAND TRANSLATOR USING OPTICAL CHARACTER RECOGNITION**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_

A Thesis Presented to the Faculty of the

College of Engineering and Architecture

**BOHOL ISLAND STATE UNIVERSITY**

Main Campus, Tagbilaran City

\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In Partial Fulfillment of the Requirements of the Degree

**Bachelor of Science in Computer Engineering**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Jovihanni Caseñas

Ronnie Ibale

Christian Paul Vertulfo

Jazmin Joy Fullante

July 2022

**TABLE OF CONTENTS**

**TITLE PAGE** ……………………………………………………………………..….. i

**APPROVAL SHEET** ...…………………………………………………………..….. ii

**TABLE OF CONTENTS** …………………………………………………………..... iii

**LIST OF FIGURES AND TABLES** ………………………………………………… iv

**Chapter**

1. **INTRODUCTION**

1.1 Background of the Study …………………………………….. 1

1.2 Statement of the Problem ……………………………………. 3

1.3 Objectives of the Study ………………………………………. 4

1.4 Scope and Limitation …………………………………………. 5

1.5 Significance of the Study …………………………………….. 7

1.6 Theoretical Background ……………………………………… 8

1.7 Operational Definition of Terms ……………………………. 14

**2 REVIEWS OF RELATED LITERATURE**

1.1 Introduction ……………. ……………………………………. 15

1.2 Related Literature ……… …………………………………… 17

1.3 Related Studies ……….……………………………………... 19

**3 METHODOLOGY**

3.1 Introduction ……………. ……………………………………. 24

3.2 Experimental Design …… ………………………………….. 24

3.3 Research Procedure ……….……………………………….. 37

3.4 Site Variables ………. ………………………………………. 38

3.5 Environment and Participants………………………………. 38

3.6 Researcher’s Instruments…………………………………… 38

3.7 Data Gathering ….……………………………………........... 39

3.8 Statistical Treatment ……….………………………….......... 42

3.9 Proposed Gantt Chart of Design Process ……….……….. 43

**REFERENCE LIST** …………………………………………………………………. 44

**APPENDIX A**

Letter of Permission for Fare Survey……………………………...52

**APPENDIX B**

Technical Design ……………………………………………………53

Estimated Bill of Materials …….……………………………………55

**Chapter 1**

**INTRODUCTION**

* 1. **Background of the Study**

Shorthand is a system for rapid writing that uses symbols or abbreviations for letters, words, or phrases (Russon, 2017). The process of writing in shorthand is called stenography, which derives from two Greek words, stenos means “narrow”, and graphein means “to write”. It was also called brachygraphy and tachygraphy, which respectively mean short writing and swift writing. Throughout the centuries, shorthand has been written in systems based on orthography, phonetics, and arbitrary symbols while improving the speed and brevity of writing compared to the conventional method of writing. This will allow a well-trained person in the system to write approximately as fast as someone speaks.

As the world evolves, the shorthand system is also evolving. There were several shorthand systems introduced, but the majority of them were only temporary. Among the modern systems, Gregg shorthand is the most popular and efficient shorthand system. It was invented by John Robert Gregg, who originally called it Light-Line Phonography, and published under that name in pamphlet form in 1888 in England (The Editors of Encyclopaedia Britannica, 2016). Gregg’s system used the curvilinear motion of longhand writing while employing phonetic rather than alphabetic spelling (Norman, 2022). Pen strokes of Gregg shorthand are formed as straight lines, ellipses, or curved lines in varying sizes; each shape is assigned to a specific letter sound. These shapes are then joined together to form whole words according to the same basic principle of writing in cursive longhand (Farrer, 2022). Since its first publication, it has been adapted into many languages including Afrikaans, Chinese, Esperanto, French, German, Hebrew, Irish, Italian, Japanese, Polish, Portuguese, Russian, Spanish, Thai, and Tagalog.

The benefits of shorthand writing have enhanced its general acceptability and recognition in the world of business. Knowledge of shorthand can contribute to the development of administration and office skills by speeding up transcription, improving accuracy, and enhancing listening skills. The skill is essential in business offices, courtrooms, government offices, committee meetings, or for those in office-based roles where recording and note-taking are required. Normal human speech is too fast for the average person to write comprehensive transcriptions in longhand cursive without missing significant information. Thus, shorthand remains in use to this day. People with several years of Gregg shorthand experience are often able to transcribe at rates of over 200 words per minute (Farrer, 2022). As a result, professionals who are experts in Gregg shorthand can record speech presentations with much greater accuracy and completeness.

Despite the advancement and modification of Gregg shorthand or other shorthand systems, students still have difficulty learning shorthand. Student experience tension, inability of students to retain what they have learnt, poor English language background, low vocabulary knowledge, lack of career guidance, students attitude to shorthand, large class etc. (Afribary, 2018). In addition to this concern, learning and comprehending shorthand will be more challenging for stenography newbies or even non-shorthand writers.

To address this problem, the researchers will propose a Gregg Shorthand Translator using Optical Character Recognition. Optical Character Recognition (OCR) is the electronic conversion of handwritten content, printed text, or image-only digital documents into a machine-readable and searchable digital data format (Callaghan, 2021). The application of OCR in this study will translate Gregg shorthand to longhand, wherein each Gregg shorthand stroke has a corresponding English word. The system is mainly used for image processing and recognition of characters.

* 1. **Statement of the Problem**

Learning shorthand is like learning a new foreign language. This means that it requires extensive learning and practice. The art of shorthand has long been a skill that can be quite tedious. Gregg shorthand usually takes several months or more than a year to master. Therefore, many individuals are still unable to read and write this shorthand writing. For this reason, the researchers will be creating a system that would address this problem.

Particularly, the researchers aim to answer these questions:

1. How to design a system for Gregg Shorthand Translator using Optical Character Recognition?

2. How accurate would the system be in translating Gregg Shorthand stroke into its corresponding English word?

3. How fast would the system translate a Gregg Shorthand stroke into its corresponding English word?

* 1. **Objectives of the Study**

This study aims to create a Gregg shorthand translation system by using optical character recognition.

At the culmination of the study, the researchers aimed to achieve the following:

1. To be able to design and develop a Gregg shorthand translator using optical character recognition.

2. To be able to recognize Gregg shorthand strokes.

3. To be able to translate Gregg shorthand stroke into its corresponding English word.

4. To be able to create a system that will help stenography newbies and non-shorthand writers.

**1.4 Scope and Limitation**

The research scope and limitation are enumerated as follows:

**1.4.1 Scope:**

A Gregg shorthand anniversary dictionary shall be used to create the train data of the proposed system. Shorthand characters shall be cropped out from the scanned pages which will then be labelled with its corresponding English word. The dictionary contains a total of 15280 words which equates to the number of the train data. For the test data, handwritten samples shall be gathered from BS Office Administration students. The images shall undergo pre-processing to be ready for training and testing. OpenCV and Python shall be used to perform both the image pre-processing and development of the KNN classifier. A desktop user interface (UI) shall be created as an interface between the user and the OCR backend. A new set of test data shall be gathered for the testing of the proposed system’s accuracy.

* + 1. **Limitations:**
  1. **Significance of the Study**

The study will be beneficial to the following:

**Court Stenographers:** The court stenographers will benefit from this study, with the assistance of the system, they can easily compile ready-to-print legal documentation.

**Journalists:** The journalists will also benefit from this study, with the system’s assistance, they can easily provide important and accurate transcribed news documentation.

**Gregg Shorthand Learners:** The Gregg shorthand learners will be able to practice reading and improve their rapid writing skills with this Gregg shorthand translation system.

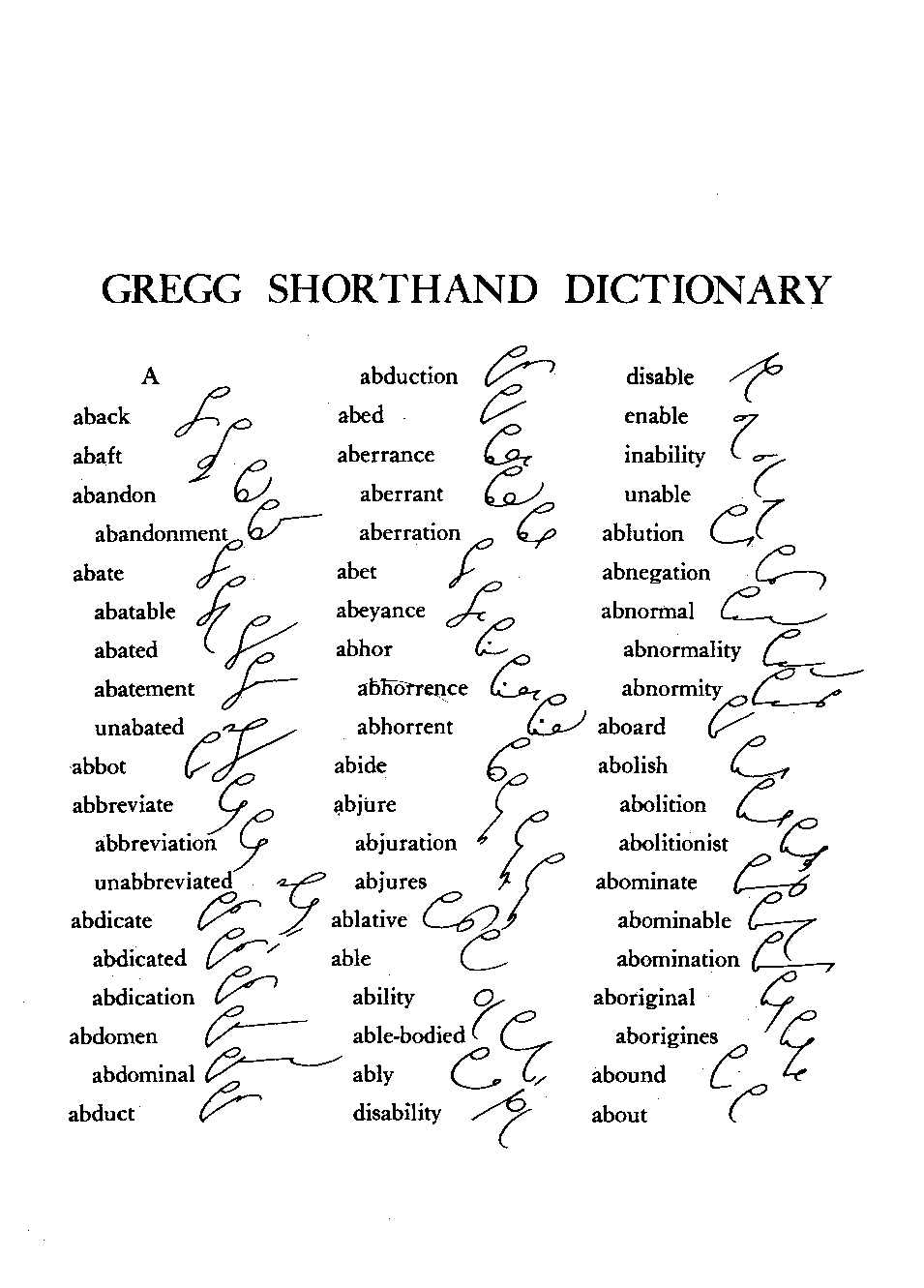
**Non-Shorthand Writers:** The non-shorthand writers can utilize this system as well; they will be able to read or even write without much prior knowledge of Gregg shorthand.

**Researchers:** The researchers will be able to learn new things and apply their knowledge to the development of this study.

**Future Researchers:** Future researchers may use this study as a source of reference data when performing their own studies or validating this project.

* 1. **Theoretical Background**

**1.6.1 Gregg Shorthand Stenography**

Gregg Shorthand, invented by John Robert Gregg on the late 19th century, is a writing system which utilizes curves, lines, hooks, and loops to deliver a handwriting speed of up to more than 200 words per minute. As technology engulfs the field of speedwriting, handwritten Gregg shorthand has slowly been replaced by stenotypes. Also known as a shorthand machine or steno writer, this device requires simultaneously pressing a combination of keys to spell out words or phrases. Although commonly used in the modern courtrooms, stenotypes are not widely accessible in the country due to its price and absence of local manufacturers.

**Figure N First Page of the Gregg Shorthand Anniversary Dictionary**

In the Philippines, the Gregg shorthand is still widely used in court rooms and journalists. What makes it preferable than other existing shorthand writing systems is its phonetic system and intuitive nature which mimics human’s natural language processing. Phonetically superfluous or redundant letters from the alphabet are deleted such as “C” and “Q” which is mostly used with the “K” and “S” sound. Phonemes with combination of several letters such as “th” is represented by one character. Even silent letters such as “gh” in “thorough” is omitted, as well as unstressed vowels.

In the present, there are no existing methods for non-Gregg shorthand writers to translate the said writing system to English longhand. As much as the Gregg shorthand aids in speedwriting, transcribing it requires time and expertise. The proposed Gregg Shorthand Translator shall address the problem by implementing OCR with a KNN classifier.

**1.6.2 Optical Character Recognition**

In 1974, the first Optical Character Recognition (OCR) product was developed by Ray Kurzweil. Later on, OCR was used to create a text-to-speech machine to help blind people. Since then, OCR has become popular as is now applied to different fields especially in the business sector wherein automation of complex document-processing workflows is important.

OCR takes the image of a scanned physical document and converts it to a black and white format wherein the characters are in white and the black is the background. Depending on the implementation, this algorithm may include identification of block or lines of text before moving onto the classification of the characters. There are two ways to do so: pattern recognition and feature recognition. For this study, the feature recognition algorithm shall be used wherein rules are applied regarding specific features of a character. For example, the word “patch” is comprised of three features: a curve that looks like the letter C, a loop, and a diagonal line. The combination of these features in this order is unique to the aforementioned word. This help the OCR program differentiate characters.

**1.6.3 K-Nearest Neighbors**

There are different ways to implement OCR. For this study, the researchers shall develop a K-Nearest Neighbors (KNN) model to classify the characters into its corresponding English longhand translation. KNN is one of the many algorithms that utilizes Supervised Machine Learning. It uses labelled data – called training data – to predict labels of unlabelled data.

The KNN algorithm works by calculating the distance between the unlabelled data and each training data. The calculated distance shall then be sorted in ascending order wherein k number of data shall be selected. The label that appears the most in the said selection is the predicted label of the unlabelled data.

This machine learning algorithm is using a lazy learning method as it stores all training data and knows that test data shall be classified to establish the classification. Contrary to other algorithms which requires a learning phase and needs the construction of a general model before accepting test data to be classified, KNN omits the learning phase. Although faster in learning, KNN is slower when it comes to classifying test data as it computes the distance of every sample to all training data every time a new sample needs to be classified.

**1.7 Legal Basis**

There are laws considered as legal basis and justification for the conduct of the study.

A section of the Philippine Supreme Court, Administrative Circular No. 24-90 states:

*“[2] (a) All stenographers are required to transcribe all stenographic notes and to attach the transcripts to the record of the case not later than twenty (20) days from the time the notes are taken. The attaching may be done by putting all said transcripts in a separate folder or envelope which will then be joined to the record of the case.”*

Jotting down stenographic notes is fast, but transcribing it to English longhand requires an ample amount of time. With the limited period that a stenographer has to submit transcripts, a system that can speed up this process can be helpful. Such system may also address the underlying problem which caused the release of this administrative circular – the absence of a stenographer who could translate shorthand transcripts of inherited cases that were passed from a Judge to another.

Another law of the constitution supports the conduct of this study. The 1987 Constitution of The Republic of The Philippines, Article XIV Section 10 states:

*“Science and technology are essential for national development and progress. The State shall give priority to research and development, invention, innovation, and their utilization; and to science and technology education, training, and services. It shall support indigenous, appropriate, and self-reliant scientific and technological capabilities, and their application to the country’s productive systems and national life. ”*

The state stresses the importance of science and technology towards the development of the country. In line with this, citizens are encouraged to partake in activities that contributes to research and development, invention, and innovation such as the conduct of this study. With the support of the state, research studies are further strengthened and empowered for the improvement of the lives of the Filipino people.

**Gregg Shorthand Stenography** (Dionis A. Padilla, Nicole Kim U. Vitug, Julius Benito S. Marquez, 2020)

“…Transcribing shorthand writing is time-consuming and sometimes confusing because of a lot of characters or words to be transcribed…”

**Optical Character Recognition** (Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, Soumya K. Ghosh, 2017)

“…process of classification of optical patterns contained in a digital image…”

**K-Nearest Neighbors** (Lishan Wang, 2019)

“…has been applied to text categorization in early research strategies and is one of the highly operational methods…”

**PHILIPPINE SUPREME COURT, ADMINISTRATIVE CIRCULAR NO. 24-90**

“…All stenographers are required to transcribe all stenographic notes and to attach the transcripts to the record of the case not later than twenty (20) days from the time the notes are taken …”

**THE 1987 CONSTITUTION OF THE REPUBLIC OF THE PHILIPPINES, ARTICLE XIV Section 10**

“…The State shall give priority to research and development, invention, innovation, and their utilization…”

Gather training data from Gregg shorthand dictionary

Develop and test K-Nearest Neighbor model for Optical Character Recognition

Design of Desktop App for Gregg Shorthand Translator

Implementation of Gregg Shorthand Translator

**Figure N Theoretical and Conceptual Framework**

**1.8 Operational Definition of Terms**

The following terms are defined to fit the needs of the researchers to conduct the study.

**Classifier.** This refers to the KNN model that shall be trained to classify the Gregg shorthand characters into English longhand.

**Gregg shorthand.** This is the type of stenography or shorthand writing that shall be translated by the proposed system.

**Longhand.** This refers to the English longhand output of the system which is comprised of words from the English language.

**Optical Character Recognition.** This refers to the algorithm to be used by the proposed system to perform shorthand translation to longhand.

**Shorthand.** This refers to the Gregg shorthand writing system which will be the input of the proposed system.

**Test data.** This refers to the set of shorthand images gathered from shorthand writers to test the system.

**Train data.** This refers to the set of shorthand images derived from a Gregg shorthand dictionary which shall be used to train the KNN model.

**Translation.** This refers to the process of converting the input image of a Gregg shorthand character into longhand text output.

**REFERENCE LIST**

Russon, A. (2017). Shorthand. Retrieved July 10, 2022, from<https://www.britannica.com/topic/shorthand>

The Editors of Encyclopaedia Britannica (2016). Gregg shorthand. Retrieved July 11, 2022, from<https://www.britannica.com/topic/Gregg-shorthand>

Norman, J.M. (2022). Gregg Shorthand is Introduced. Retrieved July 11, 2022, from<https://www.historyofinformation.com/detail.php?entryid=2416>

Farrer, A. (2022). What Is Gregg Shorthand. Retrieved July 11, 2022, from<https://www.languagehumanities.org/what-is-gregg-shorthand.htm>

Afribary (2018). PROBLEMS OF SHORTHAND TO SECRETARIAL STUDENTS IN INSTITUTIONS OF HIGHER LEARNING (A CASE STUDY OF INSTITUTE OF MANAGEMENT AND TECHNOLOGY AND FEDERAL POLYTECHNIC IDAH). Retrieved July 12, 2022, from<https://afribary.com/works/problems-of-shorthand-to-secretarial-students-in-institutions-of-higher-learning-a-case-study-of-institute-of-management-and-technology-and-federal-polytechnic-idah>

Callaghan, P. (2021). What Is Optical Character Recognition (OCR)?. Retrieved July 12, 2022, from <https://blog.pagefreezer.com/what-is-optical-character-recognition-ocr>