applied sciences

MDPI

Article

Burapha-TH: A Multi-Purpose Character, Digit, and Syllable Handwriting Dataset

Athita Onuean [1], Uraiwan Buatoom [2], Thatsanee Charoenporn [3], Taehong Kim [4] and Hanmin Jung [5,*]

[1] Faculty of Informatics, Burapha University, Chonburi 20131, Thailand; athitha@go.buu.ac.th
[2] Faculty of Science and Arts, Chanthaburi Campus, Burapha University, Chanthaburi 22170, Thailand; uraiwanu@go.buu.ac.th
[3] AAII, Faculty of Data Science, Musashino University, Tokyo 135-8181, Japan; thatsanee@ds.musashino-u.ac.jp
[4] Korea Institute of Oriental Medicine, Daejeon 34054, Korea; thkim@kiom.re.kr
[5] Korea Institute of Science and Technology Information, Daejeon 34141, Korea
* Correspondence: jhm@kisti.re.kr; Tel.: +82-42-869-1772

Abstract: In handwriting recognition research, a public image dataset is necessary to evaluate algorithm correctness and runtime performance. Unfortunately, in existing Thai language script image datasets, there is a lack of variety of standard handwriting types. This paper focuses on a new offline Thai handwriting image dataset named Burapha-TH. The dataset has 68 character classes, 10 digit classes, and 320 syllable classes. For constructing the dataset, 1072 Thai native speakers wrote on collection datasheets that were then digitized using a 300 dpi scanner. De-skewing, detection box and segmentation algorithms were applied to the raw scans for image extraction. The experiment used different deep convolutional models with the proposed dataset. The result shows that the VGG-13 model (with batch normalization) achieved accuracy rates of 95.00%, 98.29%, and 96.16% on character, digit, and syllable classes, respectively. The Burapha-TH dataset, unlike all other known Thai handwriting datasets, retains existing noise, the white background, and all artifacts generated by scanning. This comprehensive, raw, and more realistic dataset will be helpful for a variety of research purposes in the future.

Keywords: Thai language; handwriting image dataset; handwriting character recognition; Thai characters; Thai digits; Thai syllables

Citation: Onuean, A.; Buatoom, U.; Charoenporn, T.; Kim, T.; Jung, H. Burapha-TH: A Multi-Purpose Character, Digit, and Syllable Handwriting Dataset. *Appl. Sci.* **2022**, *12*, 4083. https://doi.org/10.3390/app12084083

Academic Editor: Valentino Santucci

Received: 14 March 2022
Accepted: 13 April 2022
Published: 18 April 2022

1. Introduction

Artificial Intelligence (AI) and related algorithms, especially intense learning algorithms, have become much more popular in the last few years. They have been applied to several research fields. The increase in popularity is mainly attributable to three factors: (1) higher performance hardware; (2) new and better tools, techniques, and libraries to train deeper networks; and (3) availability of more published datasets. While hardware and tools are readily available, the lack of comprehensive public datasets remains a challenge for many domains.

Handwriting recognition is an exciting research topic in the AI domain [1–6]. Applications incorporating handwriting recognition are used in the legal industry for postal mail and car plate recognition, invoice imaging, and form data entry. Handwriting recognition software requires a large quantity of high-quality data to train models effectively. Therefore, large, standard datasets are essential for enabling researchers to test, tune, and evaluate each new algorithm's performance.

Some standard datasets are available, such as the MNITS dataset [7]. MNIST is a handwritten Arabic digit dataset that contains 60,000 images for training and 10,000 images for testing. The database is also widely used for training and testing in machine learning. The EMNIST [8] is an extended MNIST dataset consisting of handwritten digits and letters of the English alphabet. This dataset was introduced in 2017 and is extensively used to improve deep learning algorithms.

There are several scripts that have been proposed as contributions to an international standard handwritten character dataset. The Institute of Automation of the Chinese Academy of Sciences (CASIA) [9] released CASIA-HWDB (offline) and CASIA-OLHWDB (online) in 2011. These datasets contain offline/online handwritten characters and continuous text written by 1020 people using Anoto pens on paper. The datasets of isolated characters contain about 3.9 million samples of 7185 Chinese characters and 171 symbols, and the datasets of handwritten texts contain about 5090 pages with 1.35 million character samples. The National Institute of Japanese Literature released the Kuzushiji dataset in November 2016 [10]. They scanned 35 classical books printed in the 18th century and organized their proposed dataset into three parts: (1) Kuzushiji-MNIST, a drop-in replacement for the MNIST dataset; (2) Kuzushiji-49, a much larger, but imbalanced dataset containing 48 Hiragana characters and one Hiragana iteration mark; and (3) Kuzushiji-Kanji, an imbalanced dataset of 3832 Kanji characters, including rare characters with very few samples. The Kuzushiji dataset currently consists of 3999 character types and 403,242 characters. A public domain handwritten character image dataset for the Malayalam language script contains data provided by 77 native Malayalam writers. It includes independent vowels, consonants, half consonants, vowels, consonant modifiers, and conjunct characters [11]. The glyphs of the Malayalam script have 85 classes that contain 17,236 training images, 5706 validation images, and 6360 testing images. An active contour model-based minimization technique was applied for character segmentation. The dataset was evaluated with different feature extraction techniques. A scattering convolutional network achieves 91.05% recognition accuracy. The PE-92 database project was started in 1992. PE-92 contains 100 image sets of 2350 Korean characters, considered general and in daily use [12,13]. The handwritten Korean language is syllable-based, not alphabet-based like western European languages. One person wrote between 100 and 500 characters for each set in the first 70 images sets. This dataset tries to accumulate as many writing styles as possible. Some problems occurred while developing the database in the data collection process, even though the characters selected to be written were considered general ones. For example, misspelling of complex vowels was often found. In 1997, SERI (System Engineering Research Institute) of Korea University created SERI95, a Hangul dataset. The SERI95 merged with ETRI, which contains 520 sets, one for each of the most frequently used Hangul characters. Each set contains about a thousand samples.

Several datasets of Thai handwriting have been published. Sae-Tang and Methasate introduced a Thai online and offline handwritten character corpus in 2004 [14]. The online handwritten character corpus contains more than 44,000 handwriting samples. The images were collected using a program developed for a WACOM 6 × 8 tablet used by 63 different writers who entered Thai characters, English characters, and special symbols. The characters written include: (1) 79 patterns of Thai consonants, vowels, tones and digit characters; (2) 62 patterns of English uppercase, lowercase, and digit characters; and (3) 15 patterns of special symbols. This offline handwritten character corpus contains handwritten isolated characters, words, and sentences, and 14,000 long samples from 143 different writers. The handwritten isolated character set contains 79 patterns of consonants, vowels, tones, and digits. The word set includes names of 76 Thai provinces and 21 Thai digits. The sentence sets include 16 Thai digits and 3 Thai general articles. A new Thai handwriting dataset, ALICE-THI, was published in 2015 [15]. This dataset was collected to support research on handwritten character recognition using local gradient feature descriptors. ALICE-THI consists of 13,130 training samples and 1360 test samples, 44 consonants, 17 vowels, 4 tones, and 3 symbols. The ALICE-THI handwritten Thai digit dataset contains 8555 training samples and 1000 test samples, for 9555 samples.

It is common knowledge that a large dataset, in terms of the number of images, can help achieve better accuracy when using deep learning techniques [16–18]. The construction of a good handwriting dataset, which is comprehensive, has enough variety and is large enough for suitable training of deep learning algorithms. The lack of a diverse dataset of handwritten Thai scripts is problematic for a researcher working on Thai handwriting recognition. Another problem with existing datasets is that they cannot be easily compared. Therefore, Thai handwriting research is not progressing as quickly as it should. These issues inspired us to construct a new dataset to foster more Thai handwriting recognition research.

This paper introduces a new Thai handwriting dataset named Burapha-TH consisting of characters, digits, and syllables. Our dataset is very different from existing standard handwriting datasets. Standard datasets are generally explicit and contain preprocessed images. The preprocessing removes noise, traces image contours, performs smoothing on the images, removes the non-essential background, and performs binarization. In contrast, our new Burapha-TH dataset has only performed de-skewing and segmentation in its preprocessing steps. We did not remove the salt and pepper noise, white background, or artifacts generated by scanning. The objectives when creating this dataset were to provide good opportunities for research on a wide variety of Thai handwriting recognition tasks. The dataset is suitable for research about handwriting recognition, including feature extraction, machine learning [19], deep learning [20], and image processing of handwriting. The dataset contains raw images (without any pre-processing) of each document. We published the original data collection sheets to permit new research on glyph image processing, including alternative preprocessing approaches with more advanced skew correction, line detection, segmentation, image smoothing, and noise and white background removal. The dataset can be used to explore writing patterns related to gender. The researcher can exploit the potential of syllables handwriting style. The syllables can represent a unique style of handwriting. For each syllable, it shows a continuous handwriting style that is suitable for word segmentation or handwriting generation.

Our goal is to provide the dataset necessary to develop more robust and practical real-world applications that make dealing with Thai script images easier. Furthermore, we used a generalized unified framework for constructing the Burapha-TH datasets. It is free for other use. In this paper, we demonstrate the Burapha-TH dataset's usefulness, an experiment was performed using serval different CNN models, and the results were analyzed.

The rest of the paper is organized as follows. An overview of the Thai language script is provided in Section 2. The construction of the Burapha-TH dataset is described in Section 3. In Section 4, the results from the experiment are discussed. Finally, Section 5 concludes and suggests possible future work.

2. Overview of Thai Language Script

The first inscription of King Ramkamhaeng is historical evidence showing that the Thai script existed and has been in use since 1826. The script has been changed in both form and orthography from time to time. In 1949, Phraya Upakit Silapasarn, an expert in Thai language, Pali language, and Thai literature, presented the patterns of the characteristics of the Thai language as follows: 44 consonants, 4 tone markers, and 21 vowels (form) [21,22]. Around 1991, professionals joined together to form the Thai API Consortium (TAPIC), headed by Thaweesak Koanantakool and sponsored by National Electronics and Computer Technology Center (NECTEC), to draft a proposal for a Thai API standard called WTT 2.0 [23]. Their draft defines two eight-bit character code sets, consisting of 66 control characters (CTRL), 44 consonants (cons), 4 tone markers, 5 diacritics, 10 Thai digits, and 18 vowels (form). Their objective was to make it easier to type Thai on computers.

According to WTT 2.0, Thai words are input and stored letter-by-letter from left to right. These characters are mixed and placed on a line in four zones. All consonant characters are essentially on the baseline. At the same time, vowel characters can be positioned before, after, above or below the consonant characters or in various combinations of these positions. Moreover, tonal characters are located above consonants. If the word contains a vowel character on top of the consonant, a tonal character will be placed above that vowel character. Figure 1 illustrates a sample of word formation in the Thai language, where characters and symbols are located in different zones. ย ก ษ ข ย ว ห ญ ด ม ก are consonant characters located on the mainline, ั ์ เ ี ้ ใ ่ ุ า are vowel characters and symbols located before, after, above and below the consonant characters. Figure 2 displays a pangram of the Thai language, a verse expression that uses almost all of the characters of the Thai language.

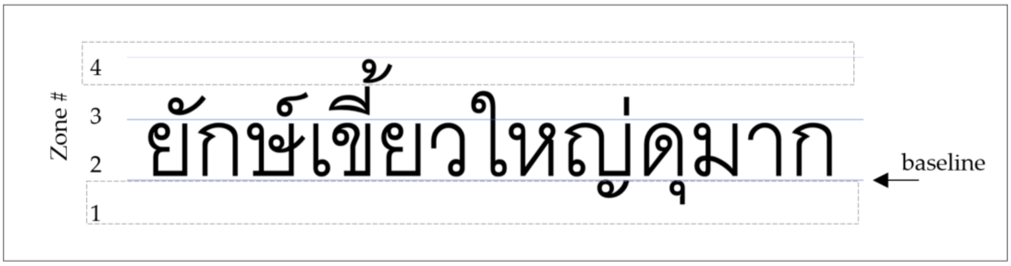
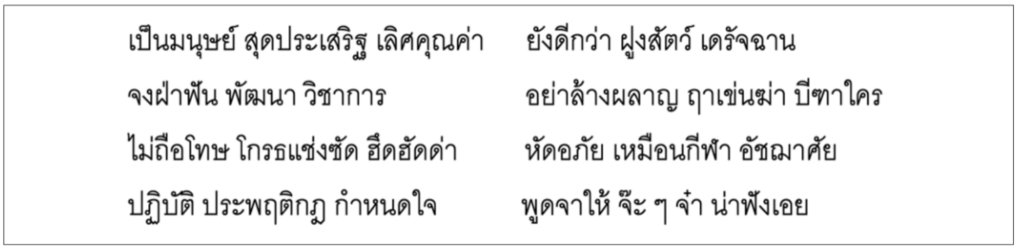


Figure 1. The Thai sentence in four-line zones style.



เป็นมนุษย์ สุดประเสริฐ เลิศคุณค่า   ยังดีกว่า ฝูงสัตว์ เดรัจฉาน
จงฝ่าฟัน พัฒนา วิชาการ   อย่าล้างผลาญ ฤาเข่นฆ่า บีฑาใคร
ไม่ถือโทษ โกรธแช่งซัด ฮึดฮัดด่า   หัดอภัย เหมือนกีฬา อัชฌาศัย
ปฏิบัติ ประพฤติกฎ กำหนดใจ   พูดจาให้ จ๊ะ ๆ จ๋า น่าฟังเอย

Figure 2. Thai Pangram.

2.1. The Characteristics of Thai Characters

2.1.1. Consonants

The structure of a Thai character has the following components: head, tail, mid loop, serration, beak, flag, and pedestal. The head or a single loop is a unique feature of Thai characters, classified into standard single loops such as บ, curly loops such as ข, or serration such as ซ. The tail is a concatenated line above or below zone 2, such as ป and ฤ. The mid loop or second loop looks like the head, but the mid loop is in the middle of the character and occurs as a line connection such as in ห ม. The serration, or a sawtooth, occurs at the characters' head or tail such as in ฅ or ฏ. The beak is a line that looks like a bird's horny projecting jaw in some characters such as ก and ถ. The flag is the end of the line and resembles a flying flag. Examples of this are ธ and ร. Finally, a pedestal can be considered the character's foot, which occurs in a few characters such as ญ and ฐ.

2.1.2. Vowels

The number of vowels in the Thai language can be counted in different ways, such as 21 forms and 32 or 36 forms and 21 sounds. This is because some vowels are formed from other vowels plus some final consonant sounds. However, the vowel and the sound are not represented differently in the overall picture. In the Thai keyboard system, there are only 17 forms of vowels, but users can still type all Thai vowels by combining them together as shown in Table 1. Vowels in words are placed in five positions: in front of a consonant character, behind a consonant character, above a consonant character, below a consonant character, and surrounding a consonant character. Front vowels are vowels that occur in front of the consonant and include เ แ ไ ใ โ. Back vowels are vowels that occur behind the consonant and include ะ า. Above vowels are vowels that occur above the consonant and include ิ ี ึ ื ั. Below vowels occur below the consonant characters and include ุ ู. Surrounding vowels are vowels whose components surround consonant characters and include เ_ะ แ_ะ โ_ะ เ_าะ เ_อะ เ ียะ เ ีย เือะ เือ ัวะ ัว.

Table 1. The 17 Thai vowel forms based on Thai keyboard.

Before (5)	After (5)	Above (5)	Below (2)
เ	ะ	ิ	ุ
แ	า	ี	ู
ไ	ำ	ึ	
ใ	ฤ	ื	
โ	ฦ	ั	

2.1.3. Tone Markers

There are four tones used in the Thai language, and they are written using tonal characters or markers, including ่ ้ ๊ and ๋. The tone markers are placed above the consonant or vowel characters.

2.1.4. Digits

The Thai language has unique digit characters that are different from Arabic digit characters. These digit characters are often used in official documents. The following characters are the ten Thai digit characters written in a sequence from 0 to 9:๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙.

2.2. *How to Write*

There are no fixed writing rules for Thai characters in the regular Thai language writing method. Nevertheless, consonant characters are written horizontally from left to right. The writing will mainly start if the character has a head. Thai consonant characters can be categorized differently depending on the criteria used, including character head, a path of the line, and character size.

2.2.1. Character Head

The head is a significant characteristic of Thai consonant characters. A Thai consonant character can be classified according to the characteristics of its head. The head can start from the top line and face out, as seen in the บ and ป characters. Alternatively, sometimes the head faces in such as in the characters ผ and ฝ. Some characters consist of a head starting from the bottom line and facing in or outwards such as ถ, ร, and ฦ. The head can also start from the center of the line and turn to the right such as in ๑ and ๑ or turn to the left such as in จ, ด, and ต. The head with serration is also a unique characteristic of Thai consonants and occurs when writing ฑ and ฆ. In addition, some consonants contain two round heads starting from the top line with the head facing out such as ข and ช.

2.2.2. Path of the Line

When considering the line path, the Thai consonant characters are constructed by four different line path types: circle, horizontal line, vertical line, and diagonal line, as displayed in Figure 1.

2.2.3. Character Size

Thai consonant characters can be separated by size, width, and height. Different characters contain small, medium, and large amounts of space between the vertical left and right sides. Examples of each type are ข, ช, ง, ก and ณ, ฒ, ฌ, respectively. When considering consonant characters on a line, they can be classified into three different groups: above, in, and below the line. Most Thai consonant characters exist in the middle or baseline such as ก, ห, ง, ป, and ฟ, which are examples of characters that have a concatenated line above the middle line. Furthermore, a few characters exist with a concatenated line below the baseline, such as ฤ and ฎ.

3. Burapha-TH Dataset Construction

Our dataset is named Burapha-TH. We created this dataset as a part of research work conducted by Burapha University, Thailand, and the Korea Institute of Science and Technology Information, Korea. In this section, we describe Thai script image data construction and processing.

3.1. Writers

The most crucial aspect is collecting data from as many different styles as possible, especially in the university student group. Figure 3 shows the sex and age distribution of writers who participated in our dataset construction. The number of writers is 1072 (721 female and 351 male). The average age ranges between 17 and 25 years old. The average age is 19.59 and the standard deviation of participants is 1.299. All of the participants were Burapha University undergraduate students.

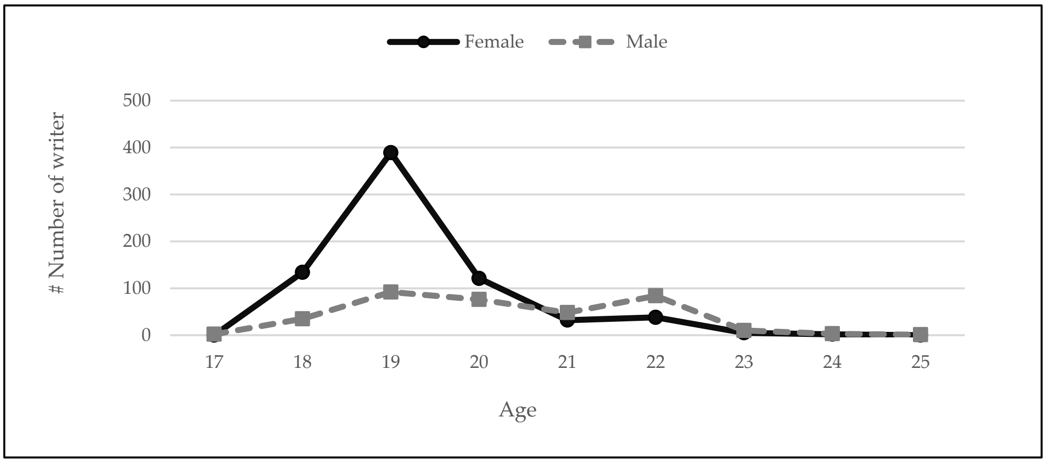


Figure 3. Statistics of age distribution about the writers.

3.2. Data Collection Sheets

We created data collection sheets, and each sheet was divided into three parts, as shown in Figure 4a. The first part records a participant's information and displays regulations for writing. The second part is the writing area, which is designed using grids. Each sheet has 100 cells, and the size of each cell is 60 × 60 pixels. We carefully designed cells with equal size, which later helped us to be able to segment each cell automatically. The top cell has a label example, which can help the participant write the blank cell character below. The third part is a blank cell for rewriting any incorrect writing, as shown in Figure 4b.

The participants wrote characters without writing style constraints, had no time limit, and could use any pen of any color. The participants were instructed to write two times on the standard collection sheets. The one regulation is that whenever participants make a writing mistake, they must cross out the wrong image and rewrite it in a blank cell in the last line of the sheet. The regulation allows the writer to write letters within the boxes, and characters must not touch the frame.

In general, Thai people usually use a pen with blue ink in everyday life, followed by black and red. In collecting the handwriting data for this research, we collected as much real data as we could. Therefore, we did not limit the ink color or size of the pens used. In terms of color, we still keep the original ink color because there is research [24] that focuses on general color tuning properties of CNNs trained for object recognition. Such research observes that color images responsible for the activation of color-sensitive kernels were more likely to be misclassified. So, this is the reason that our proposed dataset still retains the original ink color without modification.

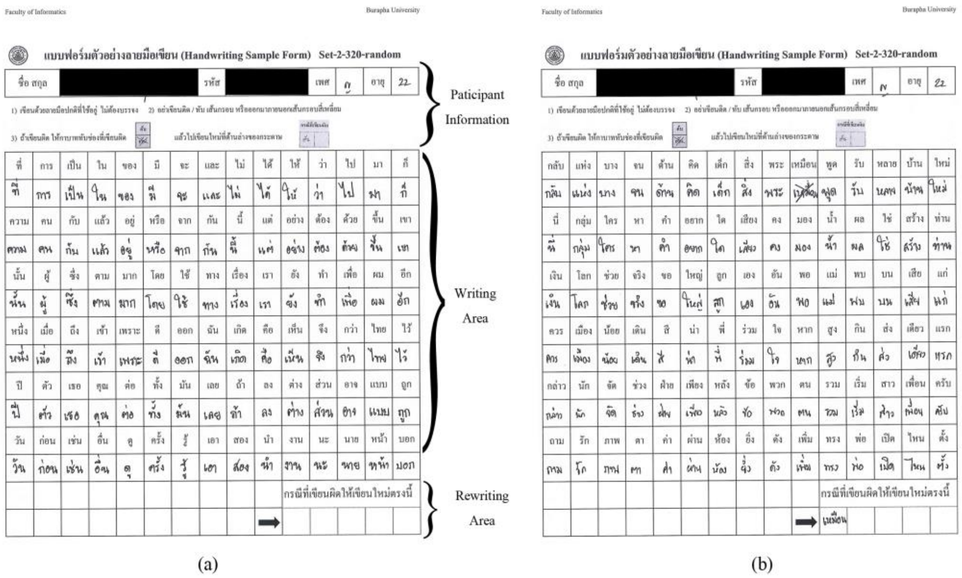

Faculty of Informatics
Burapha University
แบบฟอร์มตัวอย่างลายมือเขียน (Handwriting Sample Form) Set-2-320-random
Paticipant Information
Writing Area
Rewriting Area
(a)
(b)

Figure 4. Data Collection Sheets: (**a**) depicts three parts of collected information (**b**) depicts the case of recording the characters, vowels, and digits and the case of rewriting any incorrect writing.

3.3. Data Preparation Process

We scanned each data collection sheet using a color document scanner with 300 dpi resolution in the data preparation process. The scanner was fast for scanning the documents, but at the cost of lower resolution. Therefore, the resolution of our dataset is as general as possible. We extract information from low-quality images for handwriting recognition, and they are relatively sound. Next, we describe the algorithms for image deskewing, line detection, and image segmentation.

3.3.1. Image Deskewing

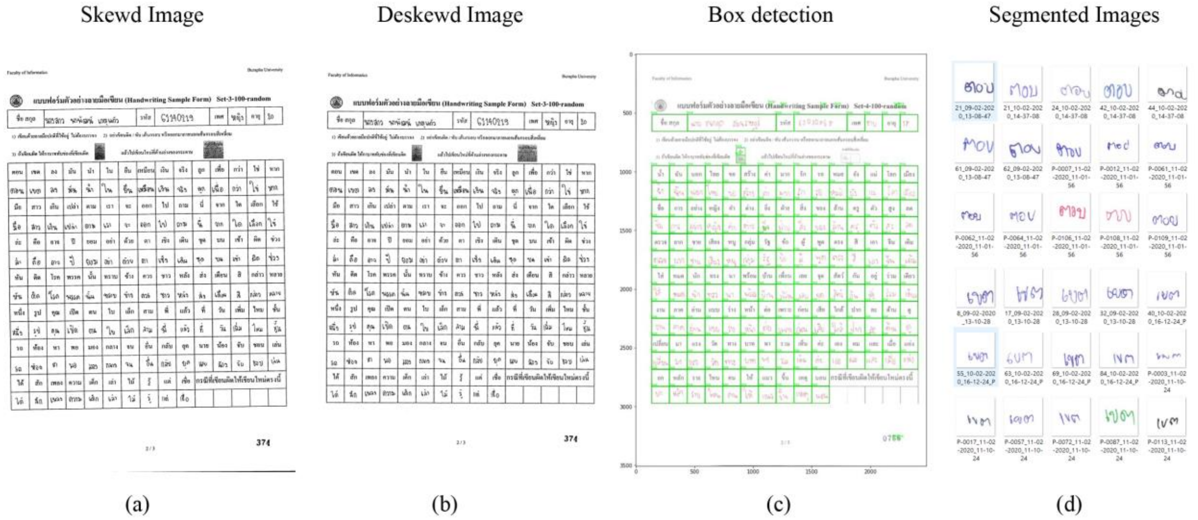
It is challenging to ensure that all paper is in the correct position for the scanning process. When scanning, some images are skewed, as shown in Figure 5a. Algorithm 1 illustrates the pseudo-code for deskewing an image to overcome this problem. We used three OpenCV libraries [25] from a GitHub webpage [26] to implement this part. The libraries are DatasetService, DeskewService, and GraphicsServices. The procedure's input is an original handwritten form image file (dm), and a deskewed handwritten glyph from an image file (ddm) is the output result of the deskewing process. In this algorithm, the main procedure is deskewing (straightening) text in image form (line 7) by calling the deskew function from DeskewService (DeskewService().deskew) that returns deskewedimage and guessedAngle. The guessedAngle value is used to check for the proper angle at −20 (line 8). We will call the rotate image function from the GraphicsService library in the skewed case, adding 90 to the guessedAngle (line 9). The output of the Image_deskewing algorithm is depicted in Figure 5b.

Algorithm 1: Pseudo-code of Image_deskewing.

```
Input:     d_m    #a original handwritten form image
Output: dd_m    #a deskewing handwritten from image
Procedure DeskewingImage(d_m)
   1:  begin
   2:        from services.DatasetService import DatasetService
   3:        from services.DeskewService import DeskewService
   4:        from services.GraphicsService import GraphicsService
   5:        imageCv = GraphicsService().openImageCv(dm)
   6:        imagePath = dm.ImagePath()
```

Algorithm 1: *Cont.*

```
        deskewedImage, guessedAngle = DeskewService().deskew(imageCv)
        **if** (guessedAngle < −20.0):
            dd = GraphicsService().rotateImage(imageCv,(guessedAngle + 90.)* 1.0)
        **end if**
        DatasetService().saveData(imagePath, deskewedImage)
    **end**
```


Skewd Image

Deskewd Image

Box detection

Segmented Images

(a)

(b)

(c)

(d)

Figure 5. Overall procedure of data preparation process: (**a**) is an original image from a document scanner, (**b**) depicts output images from Algorithm 1, (**c**) shows line detection process from Algorithm 2, and (**d**) depicts output from Algorithm 3.

3.3.2. Line Detection

Algorithm 2 is the pseudo-code for the Line detection box. The input is the deskewing handwritten result of Algorithm 1 and the output is in two forms: a matrix and a label matrix. To separate characters in size (60 × 60), we need to detect the box's size following the cell. The main procedure needs to preprocess the image to obtain a grayscale image (line 2). This grayscale image is used to find the image's horizontal and vertical scale (line 3–line 5). Then these are combined to form a big picture of an outlier (line 6) before increasing the white region in the final binary image (img_bin_final). The main procedure is used to find a stat matrix (i.e., left, top, width, height, area) with the Perform operation on (line 7).

Algorithm 2: Pseudo-code of Line_detection.

```
**Input**:   $dd_m$          #a deskewing handwritten from image
**Output**: stats, labels          # stat matrix and the label matrix
**Procedure** DetectionBox($dd_m$, line_min_width = 5)
  1:  **begin**
  2:        gray_scale = convert $dd_m$ to gray scale
  3:        img_bin_h = find the outline of the horizontal gray_scale object
  4:        img_bin_v = find the outline of the veritical gray_scale object
  5:        img_bin_final = img_bin_h | img_bin_v
  6:        img_bin_final = increases the white region in the img_bin_final image
  7:        ret, labels, stats, centroids = Perform the operation of img_bin_final
  8:  **end**
```

3.3.3. Image Segmentation

To provide a set of segmented handwritten images, Algorithm 3 is used. We applied the threshold of the cell's outlier to segmented images. The input is a set of handwritten form images (D), and the output is a set of segmented handwritten images (D_L), as shown in Figure 5d. Then we call the DeskewingImage function in algorithm 1 (line 3), and the detection box function step in algorithm 2 (line 4), respectively. The box detection is shown in Figure 5c. The stat matrix describes the optimal region to cut cells from the grid picture. Finally, each image is appended to the D_L set (line 8).

Algorithm 3: Pseudo-code of Image_segmentation.

```
Input:   $D = \{d_1, d_2, d_3, \ldots, d_{|D|}\}$   #a set of handwritten form image
Output: $D_L = \{dl_1, dl_2, dl_3, \ldots, dl_{|D|}\}$   #a set of segmented handwritten image
Procedure LineDetection($D_D$)
 1:  begin
 2:      for each document $d_m$ in $D$ do
 3:          $dd_m$ = DeskewingImage($d_m$)
 4:          $stats, labels$ = DetectionBox($dd_m$, $line\_min\_width$ = 5)
 5:          for left, top, width, height, area in stats [2:] do
 6:              if (top > 800 and width > 100 and height > 100)
 7:                  $dl = dd_m$ [top: top + height, left: left + width]
 8:                  append $dl$ in a set of segmented handwritten image $D_L$
 9:              end if
10:          end for
11:      end for
12:  end
```

3.4. Statistical Properties

The raw standard collection sheet images of characters and digits form 1156 sheets, and after segmentation yielded 107,506 images. Simultaneously, there were 1920 sheets of Thai syllables, and those sheets were segmented to produce 279,730 images. We did not resize the images, making the sizes of images non-uniform. For the next step, Thai language experts eliminated some images by using majority voting. The conditions for eliminating images were: (1) incorrect writing, (2) unreadable, (3) heavily distorted, and (4) over edge cut. Some eliminated example images are shown in Figure 6.

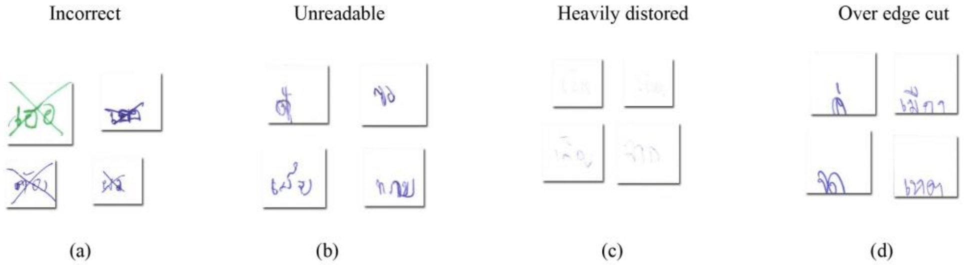


Figure 6. Example of eliminated images: (**a**) depicts incorrect writing, (**b**) depicts unreadable images, (**c**) shows examples of heavily distorted images, and (**d**) shows over edge cuts.

After unclear images were eliminated, the number of remaining characters and digits was 87,600 samples (19,906 were removed), and the number of remaining syllables was 268,056 (11,674 were removed). Thus, the total number of proper images is 355,656 samples, and 31,583, or about 8%, were discarded, as shown in Table 2.

Table 2. The statistics of raw, segmented, and current data of Burapha-TH Dataset.

Category	Collection Sheets	Segmented Images	Image Dataset
Character + Digit	1156	107,506	87,600
Syllable	1920	279,730	268,056
Total	**3076**	**387,236**	**355,656**

Currently, Burapha-TH has three categories: characters, digits, and syllables. The character dataset has 68 classes consisting of 44 Thai characters, 20 Thai vowels, and 4 Thai tone markers. We separated it into 63,327 samples of the training set and 13,600 samples of the test set. The average number of images in the training set is about 931 samples in each class, and the minimum and maximum are 790 and 995. The testing set has 200 samples in each class.

The digit dataset has 10 classes. We sequenced several Thai digits from zero to nine. The images have 10,673 samples, which we divided into training and testing sets with 8673 and 2000 samples, respectively. The average number of images in the training set is about 867 samples in each class, and the minimum and maximum are 772 and 923.

We created a new Thai syllable dataset for extending handwriting recognition research. The number of images in the training and testing sets is 236,056 and 32,000 samples. The average number of images in the training set is about 738 samples, and 503 and 905 are the minimum and maximum counts. The statistics of the proposed Burapha-TH are described in Table 3.

Table 3. Number of images in our proposed Burapha-TH datasets.

Statistic Topics	Character Dataset	Digit Dataset	Syllable Dataset
Number of Class	68	10	320
Number of Train sample	63,327	8673	236,056
Number of Test sample	13,600	2000	32,000
Average sample in Train/class	931	867	738
Min—Max sample in Train/class	790–995	772–923	503–905
Number of samples in Test/class	200	200	100
Total	**3076**	**387,236**	**355,656**

The Burapha-TH handwriting images are written in cursive forms, with or without a head, and they usually have several writing styles. We have tried to gather various collections of Thai scripts in our proposed datasets, as shown in Figure 7. The dataset has example consonants, vowels, and tone markers for characters. The digit dataset depicts Thai digits from zero to nine. Simultaneously, the syllable dataset consists of Thai syllables with different styles of consonants and vowels. The complete version of our proposed dataset includes Table A1 Thai characters with 68 classes, Table A2 Thai digits with 10 classes, and Table A3 Thai syllables with 320 classes.

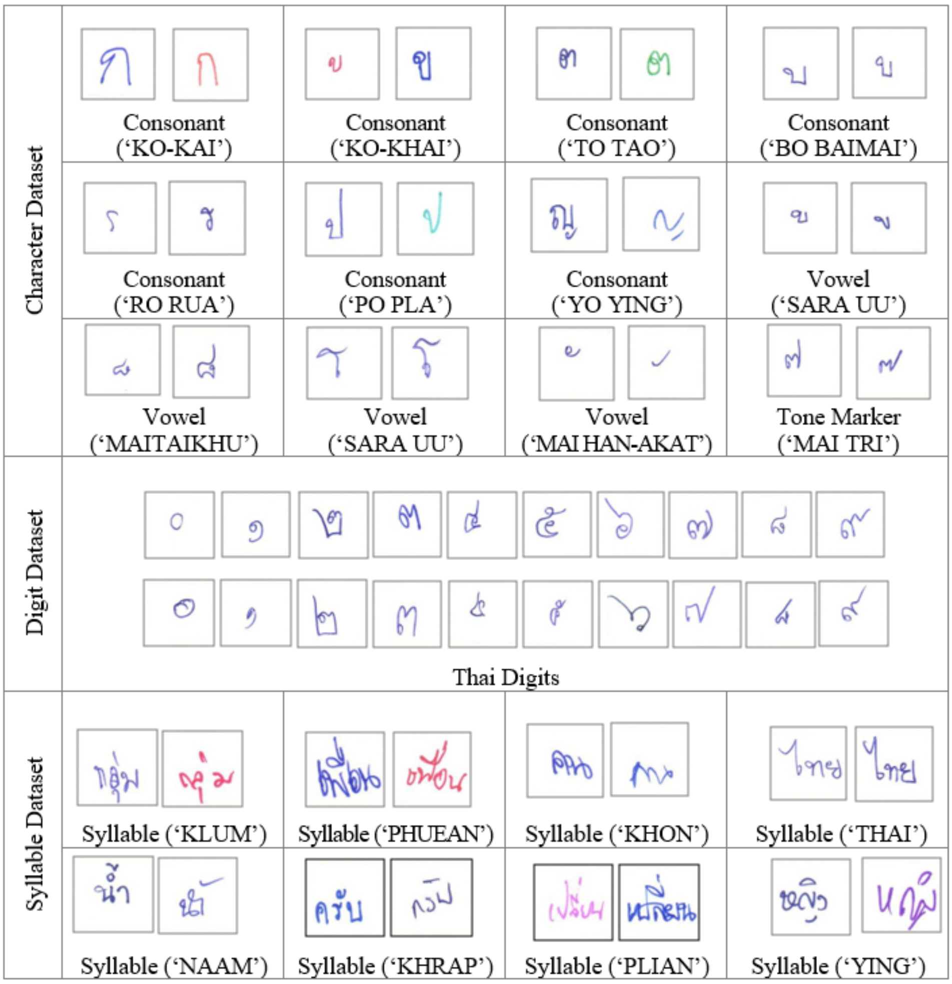


Figure 7. Examples of handwriting styles included in Burapha-TH.

4. Experiments and Discussion

We performed our study using a desktop computer with an Intel Core i7 3.6 GHz (CPU), 16 GB of memory capacity, and a Nvidia GeForce 1080Ti VGA card. We used Pytorch 1.1.0 with the Ubuntu 16.04 operating system and Python 3.7.

To show the usefulness of our proposed dataset for research in Thai handwriting recognition, we selected three popular CNN architectures: CNN with four convolutional layers, LeNet-5, and VGG-13 with batch normalization. The results are shown in Table 4 to benchmark the proposed dataset. All classifiers were repeated five times by shuffling the training set and averaging accuracy on the test set. The hyper-parameters used to train all the models were batch size 32, dropout 0.5, epoch 100, and optimizer Adam. The testing results show that VGG-13 with BN outperforms the others in terms of accuracy.

To measure the VGG-13 BN model's performance on the proposed datasets, we used k-fold cross-validation, with three, five, and ten-fold cross-validation. This technique accounts for the model's variance concerning differences in the training and test datasets and the learning algorithm's stochastic nature. A model's performance can be taken as the mean performance across k-folds, given the standard deviation, which could be used to estimate a confidence interval. We used the scikit-learn API to implement the k-fold cross-validation in our experiment. Figure 8 shows the results of percentage accuracy of all data partitions. Syllable and digit datasets show almost 99%, while the character data shows an average value of about 97%.

Table 4. An overview of the public datasets discussed in the paper and Burapha-TH dataset.

Dataset	Recognizer	Recognition Accuracy (%)		
		Training	Validation	Testing
Character	CNN (4 conv)	87.34	75.28	78.51
	LeNet-5	88.16	75.42	78.84
	VGG_13_BN	97.88	92.64	95.00
Digit	CNN (4 conv)	94.01	83.00	83.93
	LeNet-5	95.40	85.63	87.26
	VGG_13_BN	98.51	94.03	98.29
Syllable	CNN (4 conv)	83.61	75.00	78.25
	LeNet-5	76.15	71.64	74.98
	VGG_13_BN	98.41	94.40	96.16

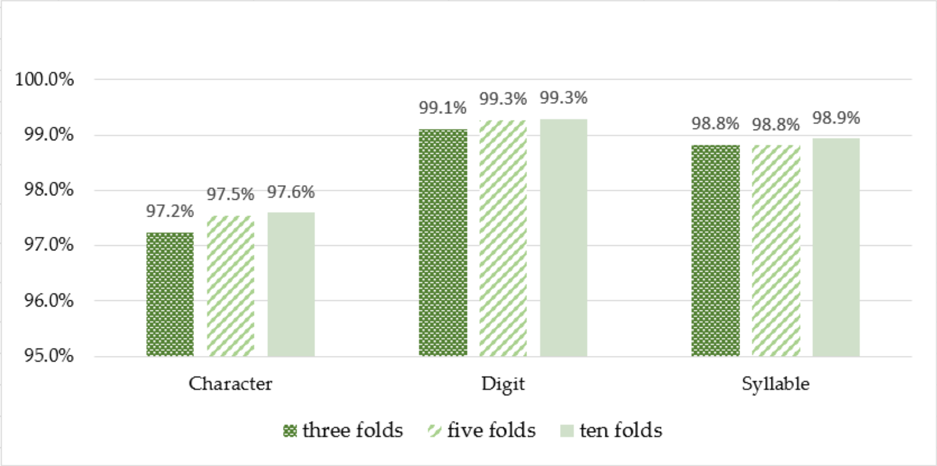


Figure 8. Result of BURAPHA-TH dataset ranking K-fold cross-validation.

A statistics overview of the public dataset discussed in the paper is compared with Burapha-TH datasets in Table 5. Our proposed datasets have a wider variety of content when compared with other public datasets. The number of writers of Burapha-TH is similar to CASIA-HWDB, which implies that our dataset has a variety of handwriting styles. The training, validation, and test samples are much more extensive than existing Thai handwriting datasets.

Table 5. A statistics comparison of Burapha-TH datasets and public datasets discussed in the paper.

Dataset	Language	Year	Content	Class	No. of Writers	Statistics (Train/Validate/Test)
MNIST	EN	1998	Digit	10	-	60,000/0/10,000
EMNIST	EN	2017	Letter	26	>500	124,800/0/20,800
CASIA-HWDB	CN	2011	Text, Character	7356	1020	3.5 M isolated character, 1.35 M characters in text
Kuzushiji-49	JP	2016	Character	49	-	232,365/0/38,547
Malayalam	IN	2019	Character	85	77	17,236/5706/6360
PE92	KR	1992	Syllable	2350		~100 per class
SERI95	KR	1997	Syllable	520		465,675/0/51,785

Table 5. *Cont.*

Dataset	Language	Year	Content	Class	No. of Writers	Statistics (Train/Validate/Test)
Thai handwritten character corpus [1]	TH	2004	Character	79	143	14,000
ALICE-THI	TH	2015	Character	68	150	13,138/0/1360
ALICE-THI	TH	2015	Digit	10	150	8555/0/1000
Burapha-TH	TH	2021	Character	68	1072	63,327/0/13,600
Burapha-TH	TH	2021	Digit	10		8673/0/2000
Burapha-TH	TH	2021	Syllable	320		236,056/0/32,000

[1] dataset not available.

5. Conclusions and Future Work

In this paper, we have presented the Burapha-TH Thai handwriting dataset. The 1072 participants wrote characters, digits, and syllables on standard collection sheets. We extracted dataset samples from 3076 sheets. These were passed through preprocessing consisting of deskewing, line detection, and image segmentation. The expert group eliminated some images.

Our proposed dataset covers 68 consonants and vowels, 10 Thai digits, and 320 Thai syllables. It contains three subsets: 76,927 images in the character dataset, 10,673 images in the digit dataset, and 268,056 images in the syllable dataset. The Burapha-TH dataset images are original, in JPG file format, true color, and without any de-noising or cleansing processing. The best performance shows 95.00%, 98.29%, and 96.16% accuracy using the VGG-13_BN model on the character, digit, and syllable data. Developing this Thai handwriting dataset is essential for improving Thai script recognition research. Our proposed dataset is available for downloading at https://services.informatics.buu.ac.th/datasets/Burapha-TH/ (accessed on 1 April 2022).

In future work, we will be publishing an extension of the present dataset, adding binarization and edge datasets, and expanding it to include more samples and syllable classes. We will also focus on model optimization for Thai handwriting recognition based on the Burapha-TH dataset.

Author Contributions: Conceptualization, H.J. and A.O.; methodology, A.O., T.K. and U.B.; software, T.K.; validation, A.O., H.J. and T.C.; formal analysis, A.O. and U.B.; investigation, A.O., H.J. and T.C.; writing—original draft preparation, A.O.; writing—review and editing, H.J., T.K., U.B. and T.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study. This data can be found here: https://services.informatics.buu.ac.th/datasets/Burapha-TH/ (accessed on 1 April 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Thai character classes considered for the dataset constructions.

Consonants (44)			
ก	00-161-A1-KO KAI	ท	22-183-B7-THO THAHAN
ข	01-162-A2-KHO KHAI	ธ	23-184-B8-THO THONG
ฃ	02-163-A3-KHO KHUAT	น	24-185-B9-NO NU

Table A1. *Cont.*

Consonants (44)			
ค	03-164-A4-KHO KHWAI	บ	25-186-BA-BO BAIMAI
ฅ	04-165-A5-KHO KHON	ป	26-187-BB-PO PLA
ฆ	05-166-A6-KHO RAKHANG	ผ	27-188-BC-PHO PHUNG
ง	06-167-A7-NGO NGU	ฝ	28-189-BD-FO FA
จ	07-168-A8-CHO CHAN	พ	29-190-BE-PHO PHAN
ฉ	08-169-A9-CHO CHING	ฟ	30-191-BF-FO FAN
ช	09-170-AA-CHO CHANG	ภ	31-192-C0-PHO SAMPHAO
ซ	10-171-AB-SO SO	ม	32-193-C1-MO MA
ฌ	11-172-AC-CHO CHOE	ย	33-194-C2-YO YAK
ญ	12-173-AD-YO YING	ร	34-195-C3-RO RUA
ฎ	13-174-AE-DO CHADA	ล	36-197-C5-LO LING
ฏ	14-175-AF-TO PATAK	ว	38-199-C7-WO WAEN
ฐ	15-176-B0-THO THAN	ศ	39-200-C8-SO SALA
ฑ	16-177-B1-THO NANGMONTHO	ษ	40-201-C9-SO RUSI
ฒ	17-178-B2-THO PHUTHAO	ส	41-202-CA-SO SUA
ณ	18-179-B3-NO NEN	ห	42-203-CB-HO HIP
ด	19-180-B4-DO DEK	ฬ	43-204-CC-LO CHULA
ต	20-181-B5-TO TAO	อ	44-205-CD-O ANG
ถ	21-182-B6-THO THUNG	ฮ	45-206-CE-HO NOKHUK
Vowels (20)			
ฯ	46-207-CF-PAIYANNOI	ู	56-217-D9-SARA UU
ะ	47-208-D0-SARA A	เ	57-224-E0-SARA E
ั	48-209-D1-MAI HAN-AKAT	แ	58-225-E1-SARA AE
า	49-210-D2-SARA AA	โ	59-226-E2-SARA O
ำ	50-211-D3-SARA AM	ใ	60-227-E3-SARA AI MAIMUAN
◎	51-212-D4-SARA I	ไ	61-228-E4-SARA AI MAIMALAI
◎	52-213-D5-SARA II	็	62-231-E7-MAITAIKHU
◎	53-214-D6-SARA UE	์	67-236-EC-THANTHAKHAT
◎	54-215-D7-SARA UEE	ฤ	35-196-C4-RU
◎	55-216-D8-SARA U	ฦ	37-198-C6-LU
Tone Markers (4)			
◎	63-232-E8-MAI EK	๊	65-234-EA-MAI TRI
◎	64-233-E9-MAI THO	๋	66-235-EB-MAI CHATTAWA

Table A2. Thai digits classes considered for the dataset constructions.

Digits (10)			
๐	68-240-F0-THAI DIGIT ZERO	๕	73-245-F5-THAI DIGIT FIVE
๑	69-241-F1-THAI DIGIT ONE	๖	74-246-F6-THAI DIGIT SIX
๒	70-242-F2-THAI DIGIT TWO	๗	75-247-F7-THAI DIGIT SEVEN
๓	71-243-F3-THAI DIGIT THREE	๘	76-248-F8-THAI DIGIT EIGHT
๔	72-244-F4-THAI DIGIT FOUR	๙	77-249-F9-THAI DIGIT NINE

Table A3. Thai Syllable classes considered for the dataset constructions.

Syllables (320)					
เก็บ	000-KEP1	เมื่อ	31-MUEA2	แรง	62-RAENG0
เกิด	001-KOET1	เมือง	32-MUEANG0	แล้ว	63-LAEO3
เกิน	002-KOEN0	เรา	33-RAO0	และ	64-LAE3
เขต	003-KHEET1	เริ่ม	34-ROEM2	แห่ง	65-HAENG1
เขา	004-KHAO4	เรียก	35-RIAK2	โดย	66-DOI0
เข้า	005-KHAO2	เรียน	36-RIAN0	โต	67-TOO0
เขียน	006-KHIAN4	เรื่อง	37-RUEANG2	โรค	68-ROOK2
เครื่อง	007-KHRUEANG2	เล็ก	38-LEK3	โลก	69-LOOK2
เงิน	008-NGOEN0	เล่น	39-LEN2	ใกล้	70-KLAI2
เจอ	009-JOOE0	เลย	40-LOEI0	ใคร	71-KHRAI0
เจ้า	010-JAAO2	เล่า	41-LAO2	ใจ	72-JAI0
เช่น	011-CHEN2	เลือก	42-LUEAK2	ใช่	73-CHAI2
เชิง	012-CHOENG0	เสีย	43-SIA4	ใช้	74-CHAI3
เชื่อ	013-CHUEA2	เสียง	44-SIANG4	ใด	75-DAI0
เด็ก	014-DEK1	เหตุ	45-HEET1	ใน	76-NAI0
เดิน	015-DOEN0	เห็น	46-HEN4	ใบ	77-BAI0
เดิม	016-DOEM0	เหมือน	47-MUEAN4	ใส่	78-SAI1
เดียว	017-DIAO0	เหลือ	48-LUEA4	ให้	79-HAI2
เดือน	018-DUEAN0	เอง	49-EENG0	ใหญ่	80-YAI1
เท่า	019-THAO0	เอา	50-AO0	ใหม่	81-MAI1
เธอ	20-THOOE0	แก	51-KAAE0	ไง	82-NGAI0
เป็น	21-PEN0	แก่	52-KAAE1	ได้	83-DAI2
เปล่า	22-PLAAO1	แก้	53-KAAE2	ไทย	84-THAI0
เปลี่ยน	23-PLIAN1	แดง	54-DAENG0	ไป	85-PAI0
เปิด	24-POOET1	แต่	55-TAAE1	ไม่	86-MAI2
เพราะ	25-PHROR3	แทน	56-THAAEN0	ไม้	87-MAI3
เพลง	26-PHLEENG0	แนว	57-NAAEO0	ไว้	88-WAI3
เพิ่ม	27-PHOEM2	แบบ	58-BAEP0	ไหน	89-NAI4

Table A3. *Cont.*

Syllables (320)					
เพียง	28-PHIANG0	แพทย์	59-PHAET2	ไหม	90-MAI4
เพื่อ	29-PHUEA2	แม่	60-MAE2	ก็	91-KOR2
เพื่อน	30-PHUEAN2	แรก	61-RAEK2	กลับ	92-KLAP1
Syllables					
กลาง	93-KLANG0	จะ	132-JA1	ติด	171-TIT1
กล่าว	94-KLAAO1	จัด	133-JAT1	ถ้า	172-THAA2
กลุ่ม	95-KLUM1	จับ	134-JAP1	ถาม	173-THAAM4
กว่า	96-KWAA1	จาก	135-JAAK1	ถึง	174-THUENG4
ก่อน	97-KORN1	จิต	136-JIT1	ถือ	175-THUE4
กัน	98-KAN0	จีน	137-JIIN0	ถูก	176-THUUK1
กับ	99-KAP1	จึง	138-JUENG0	ทรง	177-SONG0
การ	100-KAAN0	จุด	139-JUT1	ทราบ	178-SAAP2
กิน	101-KIN0	ฉัน	140-CHAN4	ทั้ง	179-THANG3
ขอ	102-KHOR4	ช่วง	141-CHUANG2	ทั่ว	180-THAW1
ข้อ	103-KHOR2	ช่วย	142-CHUAI2	ทาง	181-THAANG0
ของ	104-KHORNG4	ชอบ	143-CHORP2	ท่าน	182-THAAN2
ข้าง	105-KHAANG2	ชั้น	144-CHAN3	ทำ	183-THAM0
ขาด	106-KHAD0	ชาติ	145-CHAAT2	ที	184-THII0
ขาย	107-KHAAI4	ชาย	146-CHAAI0	ที่	185-THII2
ขาว	108-KHAAO4	ชาว	147-CHAAO0	นอก	186-NORK2
ข่าว	109-KHAAO1	ชื่อ	148-CHUE2	น้อง	187-NORNG3
ข้าว	110-KHAAO2	ชุด	149-CHUT3	นอน	188-NORN0
ขึ้น	111-KHUEN2	ซึ่ง	150-SUENG2	น้อย	189-NOI3
คง	112-KHONG0	ซื้อ	151-SUE3	นะ	190-NA3
คน	113-KHON0	ด้วย	152-DUAI2	นัก	191-NAK3
ครั้ง	114-KHRANG3	ดัง	153-DANG0	นั่ง	192-NANG2
ครับ	115-KHRAP3	ด้าน	154-DAAN2	นั้น	193-NAN3
ครู	116-KHRUU0	ดี	155-DII0	นา	194-NAA0
ควร	117-KHUAN0	ดู	156-DUU0	น่า	195-NAA2
ความ	118-KHWAAM0	ตน	157-TON0	นาน	196-NAAN0
ค่ะ	119-KHA2	ต้น	158-TON2	นาย	197-NAAI0
ค่า	120-KHAA2	ตรง	159-TRONG0	นำ	198-NAM0
คำ	121-KHAM0	ตรวจ	160-TRUAT1	น้ำ	199-NAAM3
คิด	122-KHIT3	ต่อ	161-TOR1	นี่	200-NII2
คืน	123-KHUEN0	ต้อง	162-TORNG2	นี้	201-NII3

Table A3. *Cont.*

Syllables					
คือ	124-KHUE0	ตอน	163-TORN0	บท	202-BOT1
คุณ	125-KHUN0	ตอบ	164-TORP0	บน	203-BON0
คู่	126-KHUU2	ตั้ง	165-TANG2	บอก	204-BORK1
งาน	127-NGAAN0	ตัว	166-TUA0	บาง	205-BAANG0
ง่าย	128-NGAAI2	ตา	167-TAA0	บาท	206-BAAT1
จน	129-JON0	ต่าง	168-TAANG1	บ้าน	207-BAAN2
จบ	130-JOP1	ตาม	169-TAAM0	ปรับ	208-PRAP1
จริง	131-JING0	ตาย	170-TAAI0	ปล่อย	209-PLOI1
Syllables					
ปลา	210-PLAA0	ยิ่ง	247-YING2	สาย	284-SAAI4
ปาก	211-PAAK1	ยิ้ม	248-YIM3	สาว	285-SAAO4
ปี	212-PII0	ยืน	249-YUEN0	สิ	286-SI1
ผม	213-PHOM4	ยุค	250-YUK3	สิ่ง	287-SING1
ผล	214-PHON4	รถ	251-ROT3	สิบ	288-SIP1
ผ่าน	215-PHAAN1	รวม	252-RUAM0	สี	289-SII4
ผิด	216-PHIT1	ร่วม	253-RUAM2	สุข	290-SUK1
ผู้	217-PHUU2	รอ	254-ROR0	สู่	291-SUU1
ฝ่าย	218-FAAI1	รอบ	255-RORP2	สูง	292-SUUNG4
พบ	219-PHOP3	รัก	256-RAK3	หญิง	293-YING4
พรรค	220-PHAK3	รัฐ	257-RAT3	หน้า	294-NAA2
พร้อม	221-PHRORM3	รับ	258-RAP3	หนึ่ง	295-NUENG1
พระ	222-PHRA3	ร่าง	259-RAANG2	หนู	296-NUU4
พวก	223-PHUAK2	ร้าน	260-RAAN3	หมด	297-MOT1
พอ	224-PHOR0	ราย	261-RAAI0	หมอ	298-MOR4
พ่อ	225-PHOR2	รีบ	262-RIIP2	หรือ	299-RUE4
พา	226-PHAA0	รู้	263-RUU3	หลัก	300-LAK1
พี่	227-PHII2	รูป	264-RUUP2	หลัง	301-LANG4
พูด	228-PHUUT2	ลง	265-LONG0	หลาย	302-LAAI4
ฟัง	229-FANG0	ลด	266-LOT3	ห้อง	303-HORNG2
ภาค	230-PHAAK2	ละ	267-LA3	หัน	304-HAN4
ภาพ	231-PHAAP2	ล่ะ	268-LA4	หัว	305-HUA4
มอง	232-MORNG0	ล้าน	269-LAAN3	หา	306-HAA4
มัก	233-MAK3	ลูก	270-LUUK2	หาก	307-HAAK1
มัน	234-MAN0	วัด	271-WAT3	หาย	308-HAAI4
มา	235-MAA0	วัน	272-WAN0	อย่า	309-YAA1
มาก	236-MAAK2	ว่า	273-WAA2	อยาก	310-YAAK1
มิ	237-MI3	วาง	274-WAANG0	อย่าง	311-YAANG1
มี	238-MII0	ส่ง	275-SONG1	อยู่	312-YOO1

Table A3. *Cont.*

Syllables					
มือ	239-MUE0	สร้าง	276-SAANG2	ออก	313-ORK1
ยก	240-YOK3	ส่วน	277-SUAN1	อัน	314-AN0
ยอม	241-YORM0	สวย	278-SUAI4	อา	315-ARE0
ย่อม	242-YORM2	สอง	279-SORNG4	อาจ	316-AAT1
ยัง	243-YANG0	สอน	280-SORN4	อีก	317-IIK1
ยา	244-YAA0	สัก	281-SAK1	อ่าน	318-AAN1
ยาก	245-YAAK2	สัตว์	282-SAT1	อื่น	319-UEN1
ยาว	246-YAAO0	สาม	283-SAAM4		

References

1. Singh, A.; Bacchuwar, K.; Bhasin, A. A survey of OCR applications. *Int. J. Mach. Learn. Comput.* **2012**, *2*, 314. [CrossRef]
2. Jangid, M.; Srivastava, S. Handwritten Devanagari character recognition using layer-wise training of deep convolutional neural networks and adaptive gradient methods. *J. Imaging* **2018**, *4*, 41. [CrossRef]
3. Ahlawat, S.; Choudhary, A.; Nayyar, A.; Singh, S.; Yoon, B. Improved handwritten digit recognition using convolutional neural networks (CNN). *Sensors* **2020**, *20*, 3344. [CrossRef] [PubMed]
4. Arora, S.; Bhatia, M.S. Handwriting recognition using deep learning in keras. In Proceedings of the 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 12–13 October 2018; pp. 142–145.
5. Vaidya, R.; Trivedi, D.; Satra, S.; Pimpale, P.M. Handwritten Character Recognition Using Deep-Learning. In Proceedings of the second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 20–21 April 2018; pp. 772–775. [CrossRef]
6. Eltay, M.; Zidouri, A.; Ahmad, I. Exploring deep learning approaches to recognize handwritten Arabic texts. *IEEE Access* **2020**, *8*, 89882–89898. [CrossRef]
7. LeCun, Y.A. The MNIST Database of Handwritten Digits. 1998. Available online: http://yann.lecun.com/exdb/mnist/ (accessed on 1 March 2021).
8. Cohen, G.; Afshar, S.; Tapson, J.; Van Schaik, A. EMNIST: Extending MNIST to handwritten letters. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 2921–2926. [CrossRef]
9. Liu, C.L.; Yin, F.; Wang, D.H.; Wang, Q.F. CASIA online and offline Chinese handwriting databases. In Proceedings of the International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 37–41. [CrossRef]
10. Clanuwat, T.; Bober-Irizar, M.; Kitamoto, A.; Lamb, A.; Yamamoto, K.; Ha, D. Deep learning for classical Japanese literature. *arXiv* **2018**, arXiv:1812.01718.
11. Manjusha, K.; Kumar, M.A.; Soman, K.P. On developing handwritten character image database for Malayalam language script. *Eng. Sci. Technol. Int. J.* **2019**, *22*, 637–645. [CrossRef]
12. Kim, I.J.; Xie, X. Handwritten Hangul recognition using deep convolutional neural networks. *Int. J. Doc. Anal. Recognit. (IJDAR)* **2015**, *18*, 1–13. [CrossRef]
13. KIM, D.H.; Hwang, Y.S.; Park, S.T.; Kim, E.J.; Paek, S.H.; BANG, S.Y. Handwritten Korean character image database PE92. *IEICE Trans. Inf. Syst.* **1996**, *79*, 943–950.
14. Sae-Tang, S.; Methasate, L. Thai handwritten character corpus. *IEEE Int. Symp. Commun. Inf. Technol.* **2004**, *1*, 486–491. [CrossRef]
15. Surinta, O.; Karaaba, M.F.; Schomaker, L.R.; Wiering, M.A. Recognition of handwritten characters using local gradient feature descriptors. *Eng. Appl. Artif. Intell.* **2015**, *45*, 405–414. [CrossRef]
16. Liu, C.L.; Nakagawa, M. Evaluation of prototype learning algorithms for nearest-neighbor classifier in application to handwritten character recognition. *Pattern Recognit.* **2001**, *34*, 601–615. [CrossRef]
17. Ciresan, D.C.; Meier, U.; Gambardella, L.M.; Schmidhuber, J. Convolutional neural network committees for handwritten character classification. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 1135–1139. [CrossRef]
18. Pratt, S.; Ochoa, A.; Yadav, M.; Sheta, A.; Eldefrawy, M. Handwritten digits recognition using convolution neural networks. *J. Comput. Sci. Coll.* **2019**, *34*, 40–46.
19. Michie, D.; Spiegelhalter, D.J.; Taylor, C.C. *Machine Learning, Neural and Statistical Classification*; Ellis Horwood Ltd.: Chichester, UK, 1994.
20. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
21. Uppakitsinlapasn. Principles of Thai Language—Akkhawawithi, Wachiwiphak, Wakkayasamphan, Chanthalak. Thai Wattana Phanit Publisher: Bangkok, Thailand, 1931. (In Thailand)

22. Iwasaki, S.; Ingkaphirom, P.; Horie, I.P. *A Reference Grammar of Thai*; Cambridge University Press: Cambridge, UK, 2005.
23. Koanantakool, H.T.; Karoonboonyanan, T.; Wutiwiwatchai, C. Computers and the Thai language. *IEEE Ann. Hist. Comput.* **2009**, *31*, 46–61. [CrossRef]
24. Flachot, A.; Gegenfurtner, K.R. Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks. *Vis. Res.* **2021**, *182*, 89–100. [CrossRef] [PubMed]
25. Bradski, G. The open CV library. *Dr. Dobb's J. Softw. Tools Prof. Program.* **2000**, *25*, 120–123.
26. Ertuna, L. (n.d.) Open CV Library: GitHub—JPLeoRX/Opencv-text-deskew. Available online: https://github.com/JPLeoRX/opencv-text-deskew (accessed on 1 December 2021).