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

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 several 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. ร ก ข ฃ ว ฅ ฆ ง จ ฉ ด ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ บ ป ผ ฝ ฟ ภ ฮ ร ร น บ ป ผ ฝ ฟ ภ ฮ ร ร น are consonant characters located on the mainline, ค ฅ ฆ ง จ ฉ ด ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ บ ป ผ ฝ ฟ ภ ฮ ร ร น 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.

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 ทร, curly loops such as ถร, or serration such as ฏร. The tail is a concatenated line above or below zone 2, such as ยร and ยร. 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 คร. The serration, or a sawtooth, occurs at the characters’ head or tail such as in ษร or ฏร. The beak is a line that looks like a bird’s horny projecting jaw in some characters such as ญร and รย. The flag is the end of the line and resembles a flying flag. Examples of this are ษร and รย. Finally, a pedestal can be considered the character’s foot, which occurs in a few characters such as ฏร and ฏร.

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 ี ไ ฤ ฤ. Back vowels are vowels that occur behind the consonant and include า า. Above vowels are vowels that occur above the consonant and include ำ ำ ำ ำ. Below vowels occur below the consonant characters and include ำ ำ ำ ำ. Surrounding vowels are vowels whose components surround consonant characters and include ำ ำ ำ ำ ำ ำ ำ ำ ำ ำ. 
Table 1. The 17 Thai vowel forms based on Thai keyboard.

| Before (5) | After (5) | Above (5) | Below (2) |
|------------|-----------|-----------|-----------|
|  бюджет     | .Direct     |  DIRECT |  DIRECT  |
| บะเร น     | บันที่      | บันที่   | บันที่    |
| ค้ามค     | ค้ามค      | ค้ามค   | ค้ามค    |
| ค้านค      | ค้านค      | ค้านค   | ค้านค    |
| ค้าสค      | ค้าสค      | ค้าสค   | ค้าสค    |

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 ภ. The head can also start from the center of the line and turn to the right such as in ง and ร or turn to the left such as in ฎ, ฏ, and ฏ. The head with serration is also a unique characteristic of Thai consonants and occurs when writing ง and ฏ. In addition, some consonants contain two round heads starting from the top line with the head facing out such as ฎ and ฏ.

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 ง, ง, ง, and ง, ง, ง, 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 ง, ง, ง, and ง, 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 ง and ง.
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.](image)

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.
We will call the rotate image function from the GraphicsService library in the skewed case, deskewing, line detection, and image segmentation.

Adding 90 to the guessedAngle (line 9). The output of the Image_deskewing algorithm is guessedAngle. The guessedAngle value is used to check for the proper angle at function from DeskewService (DeskewService().deskew) that returns deskewedimage and procedure is deskewing (straightening) text in image form (line 7) by calling the deskew 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.

7:   deskewedImage, guessedAngle = DeskewService().deskew(imageCv)
8:   if (guessedAngle < -20.0):
9:     dd = GraphicsService().rotateImage(imageCv, (guessedAngle + 90.)* 1.0)
10:   end if
11:   DatasetService().saveData(imagePath, deskewedImage)
12:   end

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 x 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_sm | #a deskewing handwritten from image |
|----------------|-------|------------------------------------|
| Output:        | stats, labels | # stat matrix and the label matrix |
| Procedure:     | DetectionBox/dd_sm, line_min_width = 5 |                                |
| 1:             | begin |                                    |
| 2:             | gray_scale = convert dd_sm 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 vertical 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 \( (DL) \), 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 \( DL \) 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:** \( DL = \{dl_1, dl_2, dl_3, \ldots, dl_{|D|}\} \) \#a set of segmented handwritten image |
| **Procedure** LineDetection(D) |
| 1: begin |
| 2: for each document \( d_m \) in \( D \) do |
| 3: \( dd_m = \text{DeskewingImage}(d_m) \) |
| 4: \( \text{stats, labels} = \text{DetectionBox}(dd_m, \text{line_min_width} = 5) \) |
| 5: for \( \text{left, top, width, height, area in stats [2]} \) do |
| 6: if \( (\text{top} > 800 \text{ and width} > 100 \text{ and height} > 100) \) |
| 7: \( dl = dd_m [\text{top: top + height, left: left + width}] \) |
| 8: append \( dl \) in a set of segmented handwritten image \( DL \) |
| 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.

| Incorrect | Unreadable | Heavily distorted | Over edge cut |
|-----------|------------|------------------|--------------|
| ![Incorrect](image) | ![Unreadable](image) | ![Heavily distorted](image) | ![Over edge cut](image) |

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.
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.](image)

### 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 |
| 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             | 17,236/570/6360                  |
| 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 | 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 | ขา | 22-183-B7-THO THAHAN |
| ข | 01-162-A2-KHO KHAI | ย | 23-184-B8-THO THONG |
| ฃ | 02-163-A3-KHO KHUAT | ฆ | 24-185-B9-NO NU |
### Table A1. Cont.

#### Consonants (44)

| ะ | 03-164-A4-KHO KHWAI | บ | 25-186-BA-BO BAIMAI |
| ต | 04-165-A5-KHO KHON | ป | 26-187-BB-PO PLA |
| ษ | 05-166-A6-KHO RAKHANG | ม | 27-188-BC-PH PHUNG |
| ง | 06-167-A7-NGO NGU | ฉ | 28-189-BD-FO FA |
| จ | 07-168-A8-CHO CHAN | ญ | 29-190-BE-PH PHAN |
| ฉ | 08-169-A9-CHO CHING | พ | 30-191-BF-FO FAN |
| ช | 09-170-AA-CHO CHANG | ฎ | 31-192-CH-PH SAMPHAO |
| ฌ | 10-171-AB-SO SO | ฏ | 32-193-C1-MO MA |
| นะ | 11-172-AC-CHO CHOE | ฏ | 33-194-C2-YO YAK |
| ญู | 12-173-AD-YO YING | ฏ | 34-195-C3-RO RUA |
| ญู | 13-174-AE-DO CHADA | ฏ | 36-197-C5-LO LING |
| ญู | 14-175-AF-TO PATAK | ฏ | 38-199-C7-WO WAEN |
| ญู | 15-176-B0-THO THAN | ฏ | 39-200-C8-SO SALA |
| ข | 16-177-B1-THO NANGMOUTH | ฏ | 40-201-C9-SO RUSI |
| ฌ | 17-178-B2-THO PHUTHAO | ฏ | 41-202-CA-SO SUA |
| ฆ | 18-179-B3-NO NEN | ฏ | 42-203-CB-HO HIP |
| ศ | 19-180-B4-DO DEK | ฏ | 43-204-CC-LO CHULA |
| ษ | 20-181-B5-TO TAO | ฏ | 44-205-CD-O ANG |
| ง | 21-182-B6-THO THUNG | ฏ | 45-206-CE-HO NOKHUK |

#### Vowels (20)

| ฯ | 46-207-CF-PAIYANNOI | อ | 56-217-D9-SARA UU |
| ะ | 47-208-D0-SARA A | อ | 57-224-E0-SARA E |
| ำ | 48-209-D1-MAI HAN-AKAT | อ | 58-225-E1-SARA AE |
| า | 49-210-D2-SARA AA | อ | 59-226-E2-SARA O |
| อ | 50-211-D3-SARA AM | อ | 60-227-E3-SARA AI MAIMUAN |
| ำ | 51-212-D4-SARA I | อ | 61-228-E4-SARA AI MAIMALAI |
| ำ | 52-213-D5-SARA II | อ | 62-231-E7-MAITAIIKHU |
| ะ | 53-214-D6-SARA UE | อ | 67-236-EC-THANTHAKHAT |
| ำ | 54-215-D7-SARA UEE | กรณี | 35-196-C4-RU |
| ะ | 55-216-D8-SARA U | กรณี | 37-198-C6-LU |

#### Tone Markers (4)

| ฯ | 63-232-E8-MAI EK | " | 65-234-EA-MAI TRI |
| ะ | 64-233-E9-MAI THO | " | 66-235-EB-MAI CHATTAWA |
Table A2. Thai digits classes considered for the dataset constructions.

| Digits (10) | Thai Digit  |
|------------|------------|
| ๐          | 68-240-F0-THAI DIGIT ZERO  |
| ๑          | 69-241-F1-THAI DIGIT ONE   |
| ๒          | 70-242-F2-THAI DIGIT TWO   |
| ๓          | 71-243-F3-THAI DIGIT THREE |
| ๔          | 72-244-F4-THAI DIGIT FOUR  |

| Digits (10) | Thai Digit  |
|------------|------------|
| ๕          | 73-245-F5-THAI DIGIT FIVE |
| ๖          | 74-246-F6-THAI DIGIT SIX  |
| ๗          | 75-247-F7-THAI DIGIT SEVEN|
| ๘          | 76-248-F8-THAI DIGIT EIGHT|
| ๙          | 77-249-F9-THAI DIGIT NINE |

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

| Syllables (320) | Thai Syllable |
|-----------------|---------------|
| เก็บ             | 000-KEP1      |
| เกิด             | 001-KOET1     |
| เกิน             | 002-KOEN0     |
| เขิด             | 003-KHEET1    |
| เขา              | 004-KHAO4     |
| เขา              | 005-KHAO2     |
| เชื้อต้อง        | 006-KHIAN4    |
| เครื่อง           | 007-KHRUEANG2 |
| เงิน             | 008-NGOEN0    |
| เจอก             | 009-JOEOE1    |
| เจ้า              | 010-JAAO2     |
| เช่น              | 011-CHEA2     |
| เชิง              | 012-CHOENG0   |
| เชือ              | 013-CHUEA2    |
| เติม              | 014-DEK1      |
| เต้น              | 015-DOEN0     |
| เติม              | 016-DOEM0     |
| เติม              | 017-DIAO0     |
| เติม              | 018-DUEAN0    |
| เท่า              | 019-THAO0     |
| เหอ              | 020-THOEOE0   |
| เป็น              | 021-PEN0      |
| เปล่า             | 022-PLAEO1    |
| เปลี่ยน           | 023-PLIAN1    |
| เปลี่ยน           | 024-PLIAN2    |
| เปลี่ยน           | 025-PLIAN0    |
| เปลี่ยน           | 026-PLIAN1    |
| เปลี่ยน           | 027-PLIAN2    |

| Syllables (320) | Thai Syllable |
|-----------------|---------------|
| เริ่ม             | 31-MUEA2      |
| เมื่อ              | 32-MUEANG0    |
| เรา              | 33-RAO0       |
| เริ่ม             | 34-ROEM2      |
| เริ่ม             | 35-RAIO2      |
| เริ่ม             | 36-RIAN0      |
| เริ่ม             | 37-RUEANG2    |
| เริ่ม             | 38-RIEAE2     |
| เริ่ม             | 39-REN2       |
| เริ่ม             | 40-REN3       |
| เริ่ม             | 41-REN4       |
| เริ่ม             | 42-REAO2      |
| เริ่ม             | 43-RIAEN2     |
| เริ่ม             | 44-RIAAT2     |
| เริ่ม             | 45-REAO3      |
| เริ่ม             | 46-REN4       |
| เริ่ม             | 47-REN5       |
| เริ่ม             | 48-REN6       |
| เริ่ม             | 49-REN7       |
| เริ่ม             | 50-REN8       |
| เริ่ม             | 51-REAO3      |
| เริ่ม             | 52-REAO4      |
| เริ่ม             | 53-REAO5      |
| เริ่ม             | 54-REAO6      |
| เริ่ม             | 55-REAO7      |
| เริ่ม             | 56-REAO8      |
| เริ่ม             | 57-REAO9      |
| เริ่ม             | 58-REAO10     |
| เริ่ม             | 59-REAO11     |
| เริ่ม             | 60-REAO12     |
| เริ่ม             | 61-REAO13     |
| เริ่ม             | 62-REAO14     |
| เริ่ม             | 63-REAO15     |
| เริ่ม             | 64-REAO16     |
| เริ่ม             | 65-REAO17     |
| เริ่ม             | 66-REAO18     |
| เริ่ม             | 67-REAO19     |
| เริ่ม             | 68-REAO20     |
| เริ่ม             | 69-REAO21     |
| เริ่ม             | 70-REAO22     |
| เริ่ม             | 71-REAO23     |
| เริ่ม             | 72-REAO24     |
| เริ่ม             | 73-REAO25     |
| เริ่ม             | 74-REAO26     |
| เริ่ม             | 75-REAO27     |
| เริ่ม             | 76-REAO28     |
| เริ่ม             | 77-REAO29     |
| เริ่ม             | 78-REAO30     |
| เริ่ม             | 79-REAO31     |
| เริ่ม             | 80-REAO32     |
| เริ่ม             | 81-REAO33     |
| เริ่ม             | 82-REAO34     |
| เริ่ม             | 83-REAO35     |
| เริ่ม             | 84-REAO36     |
| เริ่ม             | 85-REAO37     |
| เริ่ม             | 86-REAO38     |
| เริ่ม             | 87-REAO39     |
| เริ่ม             | 88-REAO40     |
| เริ่ม             | 89-REAO41     |

| Syllables (320) | Thai Syllable |
|-----------------|---------------|
| แข็ง             | 62-RAENG0     |
| แข็ง             | 63-LAEO3      |
| แข็ง             | 64-RAEO2      |
| แข็ง             | 65-HAENG1     |
| แข็ง             | 66-DIO2       |
| แข็ง             | 67-TOO2       |
| แข็ง             | 68-ROOK2      |
| แข็ง             | 69-LOOK2      |
| แข็ง             | 70-KLAI2      |
| แข็ง             | 71-KHRAI0     |
| แข็ง             | 72-JIAI0      |
| แข็ง             | 73-CHAI2      |
| แข็ง             | 74-CHAI3      |
| แข็ง             | 75-DAI0       |
| แข็ง             | 76-NAI0       |
| แข็ง             | 77-BAI0       |
| แข็ง             | 78-SA1I       |
| แข็ง             | 79-HAI2       |
| แข็ง             | 80-YAI1       |
| แข็ง             | 81-MAI1       |
| แข็ง             | 82-NGAI0      |
| แข็ง             | 83-DAI2       |
| แข็ง             | 84-THAI0      |
| แข็ง             | 85-PAI0       |
| แข็ง             | 86-MAI2       |
| แข็ง             | 87-MAI3       |
| แข็ง             | 88-WAI3       |
| แข็ง             | 89-NAI4       |
| Syllables (320) |
|----------------|
| เพียง | 28-PHIANG0 | แพทย์ | 59-PHAET2 | ไหม | 90-MAI4 |
| เพื่อ | 29-PHUEA2 | แม่ | 60-MAE2 | กี่ | 91-KOR2 |
| เพื่อน | 30-PHUEAN2 | แรก | 61-RAEK2 | กลับ | 92-KLAP1 |

| Syllables |
|------------|
| กลาง | 93-KLANG0 | จะ | 132-JA1 | ติด | 171-TIT1 |
| กล้า | 94-KLAAO1 | จัด | 133-JAT1 | ถ้า | 172-THAA2 |
| กลุ่ม | 95-KLUM1 | จับ | 134-JAP1 | ถาม | 173-THAAM4 |
| ภัณฑ์ | 96-KWAA1 | จาก | 135-JAAK1 | ถึง | 174-THUENG4 |
| ก่อน | 97-KORN1 | จิต | 136-JIT1 | ถือ | 175-THUE4 |
| กัน | 98-KAN0 | จีน | 137-JII1N0 | ถูก | 176-THUU1K |
| กับ | 99-KAP1 | ชิ้ง | 138-JUENG0 | ทรง | 177-SONG0 |
| กระทู้ | 100-KAAN0 | จุด | 139-JUT1 | ทะรมาน | 178-SAAP2 |
| กิ้น | 101-KIN0 | ผืน | 140-CHAN4 | ทั้ง | 179-THANG3 |
| ขอ | 102-KHOR4 | ช่วง | 141-CHUANG2 | ทั้งหมด | 180-THAW1 |
| ข้อ | 103-KHOR2 | ช่วย | 142-CHUA1I2 | ทาง | 181-THAANG0 |
| ของ | 104-KHORNG4 | ชอบ | 143-CHORP2 | ท่าน | 182-THAAN2 |
| ข้าง | 105-KHAAANG2 | จิ้น | 144-CHAN3 | ท่า | 183-THAM0 |
| ชาย | 106-KHAD0 | ชาติ | 145-CHAAAT2 | ที่ | 184-THII10 |
| ชายอื่น | 107-KHAAI4 | ชาย | 146-CHAAI10 | ที่ | 185-THII2 |
| ชาว | 108-KHAAO4 | ชาว | 147-CHAAAO0 | นอก | 186-NORK2 |
| ชาวอื่น | 109-KHAAO1 | ชื่อ | 148-CHHEE2 | หน้า | 187-NORNG3 |
| ชาวอื่นอื่น | 110-KHAAO2 | ชุด | 149-CHUT3 | หน้า | 188-NORN0 |
| ซื้อ | 111-KHUEN2 | ซื้อ | 150-SUENG2 | น้อย | 189-NOI13 |
| คง | 112-KHONG0 | ซื้อ | 151-SUE3 | นะ | 190-NA3 |
| ตน | 113-KHON0 | ตัว | 152-DUA1I2 | ฝาก | 191-NAK3 |
| ครั้ง | 114-KHRANG3 | ครั้ง | 153-DANG0 | นั่น | 192-NANG2 |
| ครั้งอื่น | 115-KHRAP3 | ต้าน | 154-DAAN2 | นั่น | 193-NAN3 |
| ครู่ | 116-KHRUU0 | ครู | 155-DII10 | น่าน | 194-NAAO |
| ควร | 117-KHUAN0 | ครู | 156-DUU10 | น่า | 195-NAAO |
| ความ | 118-KHWAA0M0 | ความ | 157-TON10 | น่า | 196-NAAN0 |
| คี่ | 119-KHA2 | คี่ | 158-TON2 | นาย | 197-NAAI0 |
| คำ | 120-KHAA2 | ค่ำ | 159-TRONG0 | น้า | 198-NAM0 |
| คำอื่น | 121-KHAM0 | คำว่า | 160-TRUAT1 | น้า | 199-NAAM3 |
| คิด | 122-KHIT3 | คิด | 161-TOR1 | ที่ | 200-NII2 |
| คืน | 123-KHUEN0 | คืน | 162-TORNG2 | ที่ | 201-NII3 |
| ไค่ | 124-KHUE0 | ตอน | 163-TORN0 | บท | 202-BOT1 |
|----|--------|----|-----|----|-----|
| คุณ | 125-KHUN0 | ตอน | 164-TORP0 | บท | 203-BON0 |
| คู่ | 126-KHUU2 | ติ่ง | 165-TANG2 | บкат | 204-BORK1 |
| นาง | 127-NGAAN0 | ตัว | 166-TUA0 | บาง | 205-BAANG0 |
| ง่าย | 128-NGAAI2 | คา | 167-TAA0 | บท | 206-BAAT1 |
| จน | 129-JON0 | ต่าง | 168-TAANG1 | บ้าน | 207-BAAN2 |
| จบ | 130-JOP1 | ตาม | 169-TAAM0 | บําบัด | 208-PRAPI |
| จริง | 131-JING0 | ตาม | 170-TAAI0 | ปลําท | 209-PLOI1 |

| อั้ง | 210-PLAA0 | ยิ่ง | 247-YING2 | สาย | 284-SAAA4 |
| ปาก | 211-PAAK1 | ยิ้ม | 248-YIM3 | ตาราง | 285-SAAO4 |
| ปิด | 212-PII0 | ปี | 249-YUEN0 | ฟิ่ม | 286-SII1 |
| แม่ | 213-PHOM4 | ยุค | 250-YUK3 | ฝิ่ง | 287-SING1 |
| ผล | 214-PHON4 | รบ | 251-ROT3 | ฝิม | 288-SIP1 |
| ลําหน | 215-PHAAN1 | รวม | 252-RUAM0 | ฟิ่ม | 289-SII4 |
| ผิว | 216-PHIT1 | รัง | 253-RUAM2 | ฟุ้ย | 290-SUK1 |
| ผู้ | 217-PHUU2 | ราย | 254-ROR0 | ฟื้น | 291-SUU1 |
| ผาย | 218-FAAAI1 | รอบ | 255-RORP2 | ฟุ้ง | 292-SUUNG4 |
| พา | 219-PHIP3 | รัก | 256-RAK3 | ฟื้นหน้า | 293-YING4 |
| พวค | 220-PHAK3 | ริ้ว | 257-RAT3 | หน้า | 294-NAA2 |
| พริม | 221-PHORM3 | รับ | 258-RAP3 | หนึ่ง | 295-NUEG1 |
| พระ | 222-PHRA3 | ร่าง | 259-RAANG2 | หนู | 296-NU4 |
| พา | 223-PHUK2 | ร้าน | 260-RAAN3 | หมด | 297-MOT1 |
| พ่อ | 224-PHOR0 | รับ | 261-RAAI0 | หมา | 298-MOR4 |
| พ่อ | 225-PHOR2 | รับ | 262-RIIP2 | หรือ | 299-RUE4 |
| พา | 226-PHAAC0 | รู้ | 263-RRU3 | หลัก | 300-LAK1 |
| ฟ้า | 227-PHIIP2 | รู้ | 264-RRUP2 | หลัง | 301-LANG4 |
| พุด | 228-PHUUT2 | สงคราม | 265-LONG0 | หลายคน | 302-LAA4 |
| พี่ | 229-FANG0 | ตก | 266-LOT3 | หิ้ง | 303-HORNG2 |
| พา | 230-PHAK2 | สว่าง | 267-LA3 | หัว | 304-HAN4 |
| พา | 231-PHAAP2 | สว่าง | 268-LA4 | หัว | 305-HU4 |
| มัด | 232-MORNG0 | ลำนำ | 269-LAAN3 | แท้ | 306-HAA4 |
| ผัก | 233-MAK3 | สุก | 270-LUUK2 | หา | 307-HAA1 |
| ผัก | 234-MAN0 | รัต | 271-WAT3 | หา | 308-HA4 |
| มา | 235-MAA0 | วัน | 272-WAN0 | ยาม | 309-YAA1 |
| มา | 236-MAAK2 | ว่า | 273-WAA2 | อย่าง | 310-YAAK1 |
| มี | 237-MI3 | วาง | 274-WAANG0 | อย่าง | 311-YAANG1 |
| มี | 238-MII0 | แห้ง | 275-SONG1 | อยู่ | 312-YOO1 |
Table A3. Cont.

| Syllables | 239-MUE0 | 276-SAANG2 | 313-ORK1 |
|-----------|----------|------------|----------|
| .slug     | 240-YOK3 | 277-SUAN1  | 314-AN0  |
| ลาย       | 241-YORM0| 278-SUA14  | 315-ARE0 |
| ย dl       | 242-YORM2| 279-SORNG4 | 316-AAT1 |
| ยิ้ง       | 243-YANG0| 280-SORN4  | 317-JIK1 |
| ยา        | 244-YAA0 | 281-SAK1   | 318-AAN1 |
| ผหาก      | 245-YAAK2| 282-SAT1   | 319-UEN1 |
| ผา        | 246-YAA00| 283-SAAM4  |          |

References

1. Singh, A.; Bachchuwar, 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.; Eldredfawry, 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. Uppakitsinthapras. 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).