applied sciences

MDPI

Article

Image-Based Radical Identification in Chinese Characters

Yu Tzu Wu *, Eric Fujiwara and Carlos Kenichi Suzuki

School of Mechanical Engineering, University of Campinas, Campinas 13083-860, Brazil
* Correspondence: yutzu@fem.unicamp.br

Featured Application: Auxiliary tool for Chinese language learning applications and contents classification of print material.

Abstract: The Chinese writing system, known as hanzi or Han character, is fundamentally pictographic, composed of clusters of strokes. Nowadays, there are over 85,000 individual characters, making it difficult even for a native speaker to recognize the precise meaning of everything one reads. However, specific clusters of strokes known as indexing radicals provide the semantic information of the whole character or even of an entire family of characters, are golden features in entry indexing in dictionaries and are essential in learning the Chinese language as a first or second idiom. Therefore, this work aims to identify the indexing radical of a hanzi from a picture through a convolutional neural network model with two layers and 15 classes. The model was validated for three calligraphy styles and presented an average F-score of ∼95.7% to classify 15 radicals within the known styles. For unknown fonts, the F-score varied according to the overall calligraphy size, thickness, and stroke nature and reached ∼83.0% for the best scenario. Subsequently, the model was evaluated on five ancient Chinese poems with a random set of hanzi, resulting in average F-scores of ∼86.0% and ∼61.4% disregarding and regarding the unknown indexing radicals, respectively.

Keywords: Han character; radical identification; image classification; Chinese character; logogram

Citation: Wu, Y.T.; Fujiwara, E.; Suzuki, C.K. Image-Based Radical Identification in Chinese Characters. *Appl. Sci.* **2023**, *13*, 2163. https://doi.org/10.3390/app13042163

Academic Editor: Zhengjun Liu

Received: 20 November 2022
Revised: 5 February 2023
Accepted: 6 February 2023
Published: 8 February 2023

1. Introduction

The Chinese writing system, known as hanzi or Han character, is fundamentally pictographic: each hanzi is composed of clusters of strokes [1,2]. Legend tells us that, in the beginning, a man called Cangjie observed the forms and phenomena in the surroundings to create the first hanzi. Later on, these symbols evolved into novel characters throughout Chinese language history, summing to over 85,000 individual hanzi across six distinctive classes [3–5]. This large amount of lexicon makes it difficult even for a native speaker to recognize the precise meaning of everything one reads but specific clusters of strokes, known as indexing radicals (Chinese: 部首; pinyin: bùshǒu; literal translation: "section header"), provide the semantic information of the whole character or even of an entire family of characters, and hence allow the meaning inference [6–8].

For instance, Figure 1a shows how the indexing radical "water" composes five characters with different meanings and pronunciations, but all describe water-related elements. Similarly, indexing radicals "grass" and "wood" usually compose hanzi for plants, and "fire" commonly describes fire or heat-related terms. This behavior in Chinese script determines the appropriate hanzi for a given context, e.g., both characters in Figure 1b have the same pinyin pronunciation and are morphologically alike, but their meanings are different, and the correct usage is defined by their indexing radical. The hanzi for place or territory describes a location and hence is earth-related; meanwhile, ancient mirrors were polished metal or copper surfaces, so the hanzi for mirror has "metal" rather than "earth" as bùshǒu.

From these examples, one realizes the importance of radical consciousness in the perception of the Chinese script. Different indexing radicals drastically change the semantics of a hanzi even when the pronunciation is the same. On the other hand, radical sensitivity is

helpful for someone learning Chinese either as a first or a second language, especially when the student faces two characters with close resemblances, such as the ones in Figure 1b, and contributes to the meaning inference of unknown characters [9–11]. It should be noted here that some of these works consider as "radicals" only the indexing radicals while others consider all sub-lexical and indivisible units with semantic-phonetic-indicative information (Chinese: 部件; pinyin: bùjiàn; literal translation: "parts" or "components"). Whichever the case, they demonstrate the importance of the radical information in language learning and meaning or pronunciation inference since over 80% of the hanzi are of phono-semantic type, i.e., while the indexing radical provides the meaning information, the remaining blocks are clues to the pronunciation [3,12,13]. Due to their importance, multimedia applications involving text, sound, and pictographic information have been explored in the past years to create language learning environments for Chinese and idioms alike, with radical awareness being one of the adopted approaches for hanzi fixation [14–16].

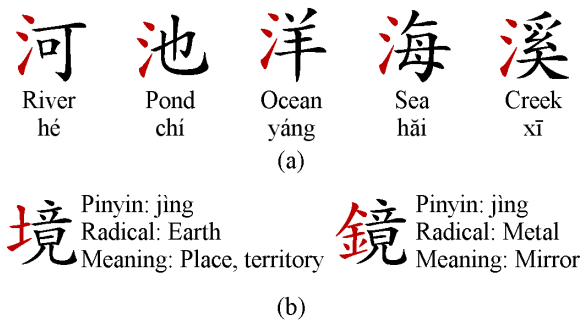

Figure 1. Examples of hanzi: (**a**) from indexing radical "water" with respective meaning and pinyin pronunciation; (**b**) with the same pronunciation, similar script, and different meanings due to the indexing radical. The radical block is marked in red.

Besides the crucial role in defining the meaning of a hanzi, the indexing radicals are also golden features in character organization since the entries in dictionaries are grouped by bùshǒu radicals and, then, by the number of strokes [17]. Alternatively, the indexing radical and secondary sub-lexical components have been an active field of research in natural language processing applications, treating Chinese characters as text or image information. For instance, Ref. [18] proposed an automatic text processing framework that combines indexing radical features and pre-trained language models to identify and classify medical terms and diagnoses from raw text, i.e., the input words are already digitized as Unicode, and the bùshǒu information is directly recovered by checking the unified Chinese-Japanese-Korean (CJK) Unicode table [19].

Similarly, Ref. [17] implemented a digital text classification routine based on the radical information. On the other hand, Ref. [20] considered the Chinese and Japanese characters as images instead of Unicode characters and proposed the extraction of the sentiment from Chinese and Japanese texts by evaluating all sub-lexical elements they contain. Refs. [21,22] played with bùjiàn radical blocks across the existing hanzi and proposed a natural calligraphy creator, while Refs. [23,24] explored sub-lexical information to identify handwriting and natural scene characters, respectively. However, most of these works, especially those taking hanzi as images, rely on all sub-lexical units instead of the indexing radical.

In this work, the authors propose an automatic image-based indexing radical identifier. In the following sections, the term "radical", without a qualifier, refers to the bùshǒu, i.e., one of the 214 indexing radicals used in dictionaries [17], instead of the 485 basic sub-lexical units with semantic-phonetic-indicative information from the GB18030 standard bùjiàn [24]. The main purpose of this research is to investigate the possibility of efficiently classifying images of hanzi according to their bùshǒu without a complete, laborious decomposition into a string of basic sub-lexical units, i.e., without a prior character recognition task. Since the Han characters are fundamentally pictographic, the hanzi are processed as images

rather than text information; thus, the component strokes and indexing radical blocks are considered visual features of the character.

The information of interest here remains the low-level indexing radical and not the high-level "which is the character?" A straightforward method of accomplishing this task would be a pipeline of digitizing the hanzi as machine Unicode and then checking its indexing radical in an electronic dictionary or even in the CJK Unicode table itself. This solution is valid, but it requires a human to manually input the hanzi into the machine first, either by transcribing it as raw text or adopting some sort of image-to-text technology, such as optical character recognition (OCR). Although OCR has reached a reliable performance for western languages, it alone is poor for Asian languages due to the close similarity of the Asian characters. Moreover, not all OCR software responds well to vertical text flows and special fonts [25,26]. Even if the end-user has access to potent software that overlooks all these limitations, the authors want to avoid the extra effort to digitize the image file as Unicode text, especially because the main concern of this work is not high-level character recognition. Although the indexing radical alone is low-level information, it is very important for hanzi acquisition during the learning process [9,10] and allows content inference without precise character recognition [17,18]. This work aims to investigate the possibility of recovering the indexing radical as image features from a picture of a hanzi for different font styles in the hope of contributing to language learning applications or content classification in the future.

2. Materials and Methods

2.1. Indexing Radical Identifier

Hanzi is a composition of a single or several smaller sub-lexical units with fairly regular morphology across different characters. Even though some bùshǒu radicals have different shapes according to their location in the hanzi (as exemplified in Figure 2), the morphological features of the indexing radicals are regular and individual. Thus, if one considers the hanzi simply as images, the characters of the same radical should all have shared morphological features, namely the strokes that compose the bùshǒu block.

In this project, an indexing radical identifier based on convolutional neural networks (CNN) was implemented with Keras and TensorFlow backend [27] to learn and recognize common features of the radical block for several indexing radicals. As represented in Figure 3, the identifier model has two layers with Leaky Rectifier Linear Unit (ReLU) activation function, and the input data is a 1-D grayscale image of 28 × 28 pixels that contains a single hanzi. The training sessions have a dropout of 50%, and 15% of the training samples are separated to validate the model parameters in each epoch. The prediction loss over these validation samples is defined as the early stopping parameter to avoid overfitting. Lastly, the output layer has 15 neurons with Softmax activation; each neuron represents a class (radical).

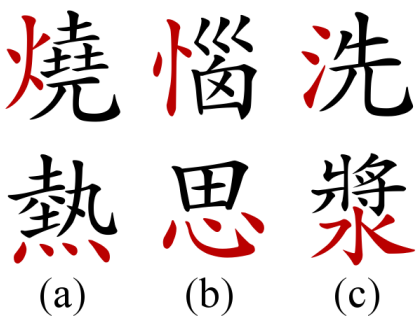

Figure 2. Different shapes of the indexing radicals (**a**) fire, (**b**) heart, and (**c**) water are on the left side or at the bottom of the hanzi. The indexing radical block is red.

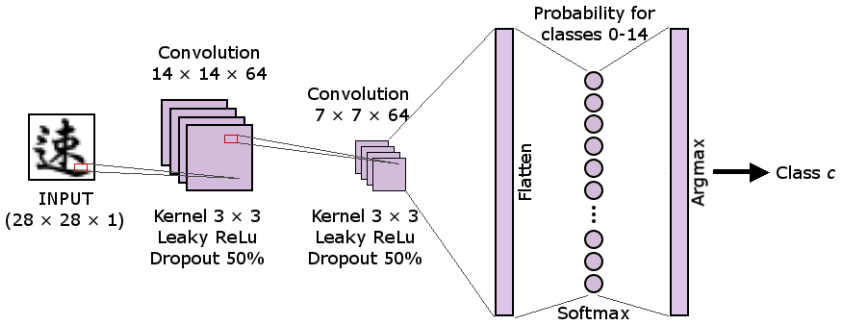

Figure 3. Diagram of the CNN model for the indexing radical identifier.

Given the low complexity of the input data and the reduced number of output classes (more details in Section 2.2), a simple CNN with two hidden layers was preferred over a model with a lot of hidden units, which shall be required once the input data grows in complexity or new functions are added to the proposed framework, increasing the number and type of output information.

2.2. Training Dataset

The dataset samples are 1-D grayscale images of 28 × 28 pixels, each holding a traditional Chinese character, where 0 corresponds to the background pixel and 255 to the actual strokes. The samples are machine-generated via a Python algorithm inspired by the zi2zi project [28].

In this work, the 15 indexing radicals with the most entries in the dictionary are studied. Table 1 lists these radicals, ordered according to the number of entries, with the respective English translation and class label. Some of them have different looks depending on their location inside the hanzi (Figure 4a–c); others assume different proportions and aspect ratios to keep the visual balance of the character even though they keep the same strokes in different positions (Figure 4d–f). The training dataset comprises all possible cases and groups them accordingly under their indexing radical, e.g., both "three-dot water" (Figure 4a, hanzi on the left) and "conventional water" (Figure 4a, hanzi on the right) have ground-truth "0". The basic dataset comprises all individual entries, of the traditional Chinese variant in the CJK table, with all possible morphology, position, and aspect ratio of the indexing radical block, i.e., the sum of the column "number of hanzi" in Table 1, totaling 9577 characters. Given the effect of the font style, these samples are augmented by rendering them in three font categories, namely regular, brush, and handwriting styles, all of the HuaKang font family (Figure 5). These styles are selected due to the calligraphy diversity, the popularity among the print texts, and the number of available fonts inside the HuaKang package. Designer fonts, such as those found in posters and marketing flyers, are usual in print documents, but there are only a few representatives within the HuaKang family, so it was not a choice. Table 2 lists the number of samples for each font style, which varies because of the available font types in each category and because some fonts cannot render every character from the basic set. The ground truth informs the actual radical of the samples as numeric labels (Table 1).

The proposed bùshǒu radical identifier is evaluated separately for each subset to establish the effects of the calligraphy over the radical identification process. For all training sessions, the corresponding dataset samples are split into training and validation datasets with an 85–15% ratio.

Table 1. List of radicals with corresponding meaning, number of hanzi in the dictionary, and class label in the dataset.

Radical	English Translation	Number of Hanzi	Class Label
水	Water	1079	0
木	Wood	1016	1
艸	Grass	981	2
口	Mouth	756	3
手	Hand	740	4
金	Gold, metal	692	5
人	Person	645	6
心	Heart	581	7
女	Girl	477	8
虫	Insect	469	9
土	Earth	460	10
火	Fire	447	11
糸	Thread	423	12
言	Word, to talk	416	13
魚	Fish	395	14

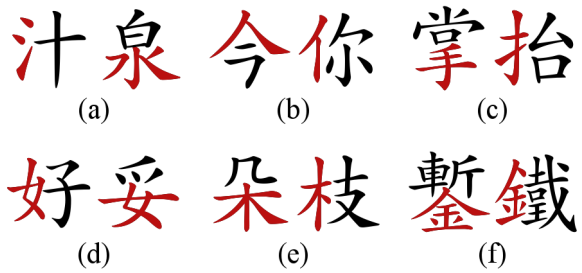

Figure 4. Examples of indexing radical variations in the forms of (**a**) water, (**b**) person, and (**c**) hand, and proportion for (**d**) girl, (**e**) wood, and (**f**) gold. The indexing radical block is red.

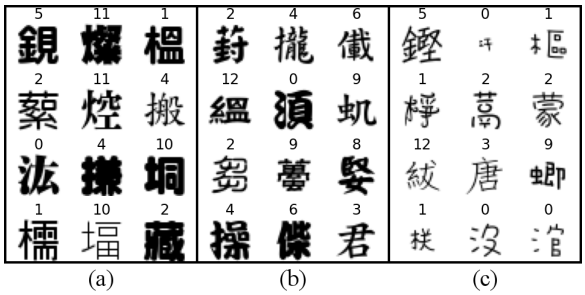

Figure 5. Samples from the training dataset for (**a**) regular, (**b**) brush, and (**c**) handwriting font styles. The number above the hanzi represents its class (radical).

Table 2. Subsets of the dataset based on the font category.

Font Category	Number of Samples
Regular	140,423
Brush	220,642
Handwriting	93,738

2.3. General Test

Whilst one can evaluate the performance of the indexing radical identifier through the validation dataset, it is not a realistic scenario since the samples are generated in the same batch with controlled parameters of size and resolution. Therefore, besides the final validation, the model is yet submitted to a general test with five different ancient Chinese

poems from the elementary school repertoire, rendered in the PMingLiU font from the MingLiU family. The choice of the font comes from the popularity of the PMingLiU font in online and offline traditional Chinese documents, allowing one to evaluate the proposed model to an unknown font in potential practical applications. The full text of the poems is reproduced in Figure 6 and contains any number of hanzi and bùshǒu radicals.

A	煮豆持作羹　漉豉以為汁 萁在釜下燃　豆在釜中泣 本是同根生　相煎何太急
B	江南可採蓮　蓮葉何田田 魚戲蓮葉間 魚戲蓮葉東　魚戲蓮葉西 魚戲蓮葉南　魚戲蓮葉北
C	皚如山上雪　皎若雲間月 聞君有兩意　故來相決絕 今日斗酒會　明旦溝水頭 躞蹀御溝上　溝水東西流 悽悽復悽悽　嫁娶不須啼 願得一心人　白頭不相離 竹竿何嫋嫋　魚尾何簁簁 男兒重意氣　何用錢刀為
D	茅檐長掃淨無苔　花木成畦手自栽 一水護田將綠繞　兩山排闥送青來
E	月落烏啼霜滿天　江楓漁火對愁眠 姑蘇城外寒山寺　夜半鐘聲到客船

Figure 6. Selected ancient poems for the general test. The translated titles and authors are (**A**) "The Quatrain of Seven Steps" by Cao Zhi; (**B**) "South of the River" from the Music Bureau of Han; (**C**) "To a Faithless Husband" by Zhuo Wenjun; (**D**) "A Poem Written on the Wall of Mr. Huyin" by Wang Anshi; and (**E**) "A Night Mooring by the Maple Bridge" by Zhang Ji.

The poems are saved as a single JPEG image file each, and the contents are extracted automatically via an image processing routine. The algorithm flattens the RGB image to a 1-D intensity channel and then segments the image into grids. Unlike alphabet-based words, machine hanzi are square-shaped and measure a fixed full-character size regardless of the visual complexity or number of strokes [29]. Therefore, the algorithm scans the whole image top-to-bottom and left-to-right to identify connected rows and columns of background pixels. These blank rows and columns are roughly the grids separating the individual characters. Subsequently, the algorithm refines the segmentation by verifying the loci individually. A locus containing a valid hanzi must satisfy two conditions:

1. It must contain foreground pixels;
2. Its total area A_i is within the 30% range of the average size of all loci (A_{avg}).

If the first condition is satisfied but the second fails, the algorithm verifies if multiple hanzi are inside A_i by dividing the locus into two or three subregions. Then, both conditions above are applied over the subregions to validate their contents. Even so, if neither is satisfied, the locus is disregarded. This verification is required to avoid missing content due to poorly connected adjacent text rows or columns. Although machine hanzi are always enclosed by fixed-width squares, depending on the font type, size, and the number of strokes, adjacent characters might have connected pixels, and translations from western languages sometimes have punctuations that spam across two full-character sized loci,

such as em dash and ellipsis. Both situations lead to a missing row or column in the initial grid, so the locus contains multiple hanzi.

Lastly, the routine verifies and resizes the extracted individual characters to 28×28 pixels to match the input data size of the classifier model. The poems are, thus, converted from a single image file to an image array of $28 \times 28 \times n$ samples, where n equals the total characters in the poem. Considering how the individual characters are extracted from the single JPEG file, there is no guarantee that their image will have the best resolution or be centered in the 28×28 window, although the final size matches the input size for the classifier.

3. Results

3.1. Indexing Radical Identification Models

An indexing radical identifier was trained especially for each font category in the training dataset and, for simplicity, henceforth, will be referred to as regular-based, brush-based, and handwriting-based identifiers according to the font category in the training dataset.

3.2. Model Validation for the 15 Indexing Radicals

The identifiers were evaluated first for the validation samples that share the same font style as the training dataset. The confusion matrices with the validation results by all three models are presented in Figure 7. The values in the main diagonal show the percentage of true positive (*TP*) detections per class, i.e., how many of the predictions of a given class belong to that class. Besides the true positive detections, the false positive (*FP*) and the false negative (*FN*) predictions were also relevant because both compute the number of misclassifications. Given *TP*, *FP*, and *FN*, the overall performance of the classifier model was evaluated in terms of the F-score,

$$\text{F-score} = \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{1}$$

where $Precision = TP/(TP + FP)$ and $Recall = TP/(TP + FN)$ [30].

The regular-based and brush-based models could identify the radicals with average F-scores of ∼98.6% and ∼96.8%, respectively, while the handwriting-based model yielded an average F-score of ∼91.8%. These values are consistent since the identifiers know both the radicals and the font style inside the validation set.

Comparing the confusion matrices in Figure 7, one concludes that the regular-based and brush-based models presented the best performance for the validation samples of the same font style as the one found in the training dataset. This behavior comes from the nature of the font style in each batch. Despite the particularities among different fonts, the regular category is upright and boxy, with clean strokes. The brush calligraphy, in turn, has an artistic flavor. Although some brush fonts are rigid with clean strokes, others are fluid and tend to join adjacent strokes, comparable to western cursive calligraphy. The handwriting fonts have similar behavior, aside from which the HuaKang handwriting has diverse scale factors within identical font size properties, leading to the (*TP*) percentage drop compared to the other two models.

The wrong detection percentages are not explicit in Figure 7, but one concludes from the scale bar that the models did not concentrate the (*FP*) and (*FN*) detections in any particular class, i.e., they did not misattribute the main visual features of an indexing radical to another. Indeed, a closer look showed that the false detections are evenly distributed outside the main diagonal, with values below 5%, and they are more noticeable in the handwriting model due to the calligraphy style and the inner scaling properties of the font itself. Even though the wrong classifications did not concentrate on a particular pair of actual-predicted classes, the classes “grass”, “wood”, “hand”, and “fire” were the most common outputs of the *FP* predictions. On the other hand, “fire”, “mouth”, and “earth” are the ones with more *FN* predictions. Both *FP* and *FN* misclassifications are reasonable by analyzing the sub-lexical elements of the hanzi. For example, a significant number of

characters have the block "wood" (木) as bùjiàn, although the indexing radical is another (to mention a few, e.g., 沐休呆慄操條茶). Another contributing factor for the higher *FP* and *FN* rates in the handwriting-based identifier is the individual styles the handwriting fonts simulate, creating sub-variations for the default ones of the indexing radical block.

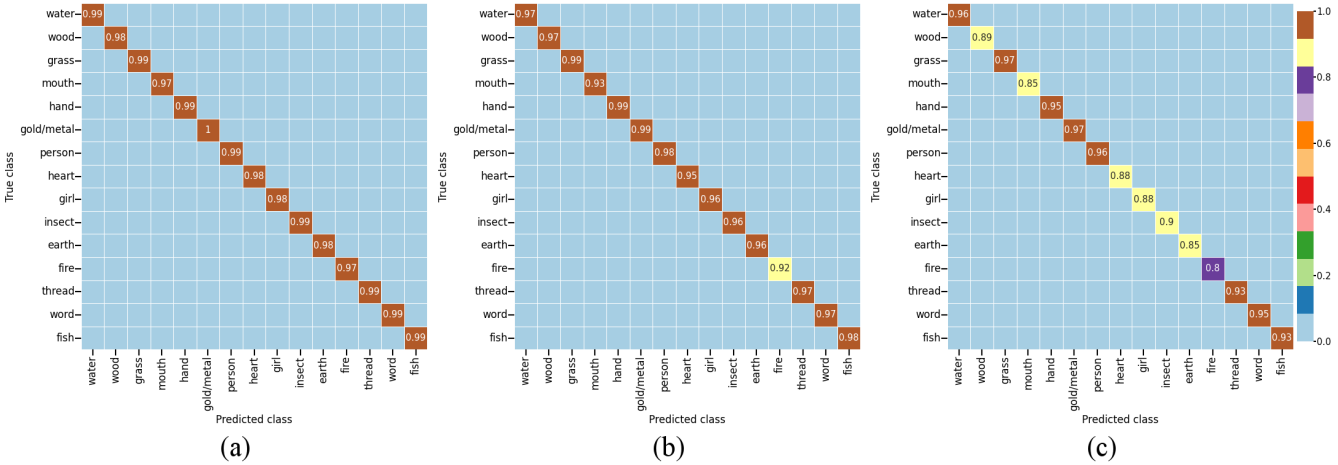

Figure 7. Confusion matrix for the radical identifier when both training and validation datasets are of (**a**) regular, (**b**) brush, and (**c**) handwriting font styles.

The model was then validated for a dataset with a font style other than the one in the training dataset to evaluate its robustness to different calligraphy styles. For instance, the regular-based model was validated with the brush and handwriting datasets, and so forth. The average F-score for all combinations of model and validation datasets is summarized in Table 3.

Table 3. Average F-score of the three indexing radical identifiers for different font styles.

Model	Validation Dataset		
	Regular	Brush	Handwriting
Regular-Based	∼98.6%	∼69.4%	∼30.9%
Brush-Based	∼83.0%	∼96.8%	∼60.4%
Handwriting-Based	∼41.8%	∼71.7%	∼91.8%

Compared with the first validation step, there is a drop in the classification performance when the validation and training datasets have different fonts, even though the radicals remained the same across the datasets. The loss, however, was the steepest for the regular-based model classifying handwriting samples, with an average F-score of ∼30.9%. This behavior is mainly due to the diverse natures of both font styles: while the regular font is upright and boxy, the handwriting is artistic and flourished. Indeed, this diversity caused a similar drop for the handwriting-based model on the regular dataset, with an F-score of ∼41.8%.

The same drop, though subtler, is observed for the brush-based model on handwriting samples, with an F-score of ∼60.4%. The better performance of the brush-based model comes from the fact that the brush calligraphy is not so sharply diverse from the handwriting as the regular font is. While the regular font is always upright and boxy, the brush writing can be fluid and flourish sometimes; hence, in a sense, it is a combination of the regular and handwriting styles. Therefore, the handwriting-based model could classify the brush samples relatively well, with an F-score of ∼71.7%.

Finally, the intermediate aspect of the brush calligraphy also explains why the regular-based model had better performance on brush samples than handwriting ones, reaching an average F-score of ∼69.4%. On the other hand, the brush-based model was somewhat robust to the regular font, with an average F-score of ∼83.0%.

3.3. Performance in General Test

In the general test, the regular-based model processed the poems A–E, rendered in the PMingLiU font from the MingLiU family (Figure 6). The font choice allows one to evaluate the proposed indexing radical identifier for a popular and unknown font style, even though the previous validation step already analyzed the three models for different calligraphy groups from the one in the training dataset. Another variable comes from the image processing operations to extract the individual characters of the poems, which add noise and shift to the input data. Moreover, since the PMingLiU font is regular and square, the test was run with the regular-based identifier alone.

The results for poem A ("The Quatrain of Seven Steps" by Cao Zhi) are shown in Figure 8. Among the 30 hanzi in the poem, 20 are from a radical among the 15 labeled classes. The correct classifications of these 20 characters are in blue, and the misclassifications are in red. As observed, the model predicted correctly 17 out of the 20 known cases, representing an F-score of 85.0%. This behavior agrees with the values obtained in the validation stage, though there was a drop in the performance. This loss is probably due to the hanzi extraction routine. While the validation samples in the previous step are individually generated with the same parameters of the training dataset, the hanzi in the poems are extracted as contents of a bigger image, considering that any character will be inside a fixed-size square. Therefore, they are bound to suffer from offset, resolution, and cropping to fit them into the expected input size of the indexing radical identifier.

Aside from the characters with known indexing radicals, there are still ten hanzi in the poem whose indexing radicals are unknown to the classifier (in black). The actual class label of these characters is indicated with a question mark in Figure 8. Since the model does not know these indexing radicals, the prediction would not match the actual class, and the F-score dropped to 68.0%. It is nevertheless interesting to analyze the results of this scenario with some case studies, as shown in Figure 9.

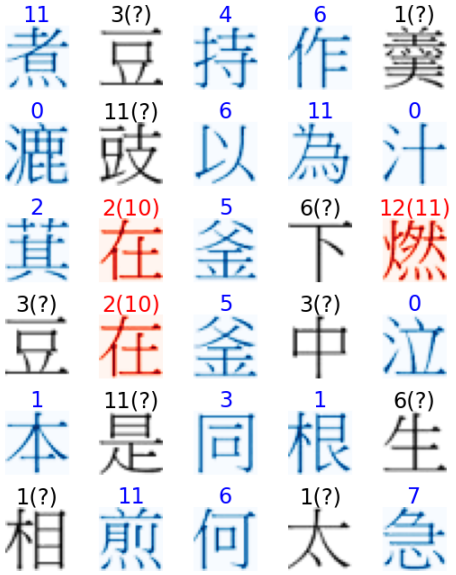

Figure 8. Classification results of the regular-based model over poem A with the correct predictions (blue), the misclassifications (red), and the unknown radicals (black). The numbers above the hanzi represent "predicted class (actual class)". The actual class is indicated with a question mark when it is unknown to the model.

The leftmost column shows the image input as the model sees it, and the second column indicates the actual radical (highlighted in red). The examples in Figure 9a,b coincide with the base form of their radical, "bean" and "center", respectively; therefore, the whole hanzi is marked in red, and the examples in Figure 9c,d are a combination of their radical, "day" and "eye" respectively, with other strokes. Since none of these

four radicals are in the training set, the model could not identify them correctly. Instead, the identification was based on other morphological traits of the character, as indicated in red in the third column. First, both hanzi in Figure 9a,b were pointed to the radical "mouth" because they both have a square, which matches the base form of "mouth" as indicated in the last column of Figure 9 in blue. Likewise, the example in Figure 9c was classified based on the second form of the radical "fire" (shown in blue). Lastly, the hanzi in Figure 9d was misclassified as the radical "wood" due to its composition. The block on the left side is "wood", while the block on the right side is "eye". Since the model does not recognize the radical "eye", it concluded naturally that the character belonged to the radical "wood" instead.

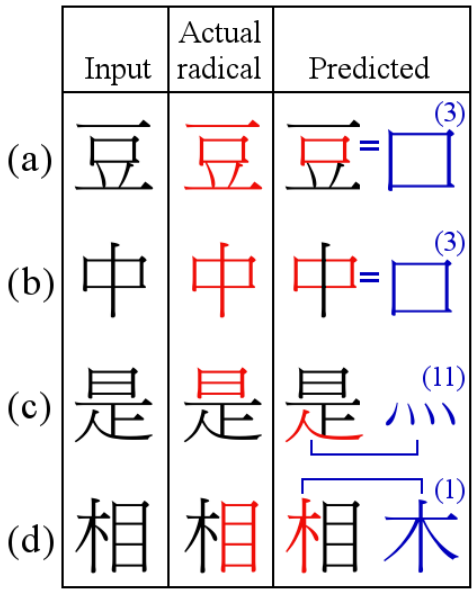

Figure 9. Examples of radical identification for unknown actual class (extracted from poem A). The original input, the actual radical (highlighted in red, 2nd column), the morphological component considered in the prediction (highlighted in red, 3rd column) as well as the base form of the predicted class (in blue, 3rd column) are presented: (**a**,**b**) actual indexing radicals with parts matching the radical block of "mouth" (class 3); (**c**) part of the hanzi is similar to the radical block of "fire" (class 11); (**d**) part of the hanzi matches the radical block of "tree" (class 1).

Finally, the same analysis is repeated for poems B–E. Table 4 summarizes the total of characters and the number of known radical hanzi in the poems. For each poem, the F-score was calculated considering only the known radicals first and then all the characters, as shown in Table 5.

Disregarding the unknown radicals, the model had an average F-score of 86.0% across the five poems, which agrees with the ∼98.6% presented during the validation step. The drop from ∼98.6% to 86.0% can be explained by some reasons. For one, the indexing radical identifier has to deal with the change in the font family. Although the PMingLiU font is still regular and is similar to one of the HuaKang fonts in the training dataset, it comes from a different family and has other features, such as different lengths, thicknesses, and angles for the strokes. For another, while the images in the training dataset have high resolution, the poems had an average image resolution to evaluate the model's robustness to noises since the latter is unavoidable in practical applications. Lastly, the training and validation samples are individually generated with well-controlled parameters, which is something the general test ruled out when preparing the input data precisely to evaluate the indexing radical identifier under more natural circumstances.

Analyzing the identifier performance over all the characters inside the poems, the model reached an average F-score of 61.4%. The performance loss is expected due to the unknown radicals. A solution for this problem is to expand the training dataset to comprise

more radicals or create an extra class (e.g., "others") to encompass the cases wherein the prediction probability for the radicals with known samples is low. Thus, instead of correlating the hanzi to a known class, the model would mark it as "others".

Table 4. Number of hanzi in the poems for the general test.

Poem	Total of Hanzi	Known Radical Hanzi
A	30	20
B	35	23
C	80	36
D	28	16
E	28	13

Table 5. F-score of the regular-based model on identifying the radicals in the poems.

Poem	Only for Known Radical Hanzi	For All Hanzi
A	85.0%	68.0%
B	100.0%	79.3%
C	83.3%	51.7%
D	75.0%	54.5%
E	84.6%	53.7%

4. Discussion

4.1. Performance of the Proposed Model

The model we proposed in this work successfully identifies the indexing radical (bùshǒu) from hanzi images rendered with different font styles as long as it knows it.

The experiments only adopted the 15 indexing radicals with the most entries in the dictionaries instead of the 214 bùshǒu in use today to first validate the viability of classifying images of hanzi according to their indexing radical. From the general test, one notices that the reduced selection of radicals is unrealistic for real-life application, so we intend to expand the known elements for the identifier. The big challenge here is the unbalanced number of entries across the indexing radicals. If the list in Table 1 continues, one realizes that only 46 indexing radicals have more than 100 characters, and 162 radicals have at least 10 characters. Therefore, we need to create a valid data augmentation routine for a fair amount of training data for each indexing radical. Moreover, we intend to extend the work to tag the indexing radical block, which requires a more elaborate training dataset [31]. Both problems are more complex and should demand a more elaborate structure for the indexing radical identifier. During this work, however, we chose a simple CNN architecture with only two hidden layers because it solves the problem at hand. Considering the unbalanced number of entries for each indexing radical and how similar some of them are [32], this simple architecture will fail in more complex situations. For simpler applications, e.g., an auxiliary learning tool for a beginner student of Chinese as a foreign language or a content classification tool to check whether a print text is plant-related or not, a reduced number of known indexing radicals is enough, and so is a simpler identifier architecture.

The proposed model can classify the images with different calligraphy styles, although the performance will still be better when the images share the characteristics of the training dataset. Comparing the performance of all three models for the three selected calligraphy groups in this work, one concludes that the identifier can deal with different calligraphy traits better than varying font sizes. Looking back at the training samples in Figure 5, it is clear that, although all of them are generated under equal window and font parameters, the actual character size varies, especially for the handwriting style. Therefore, if the regular-based model learns that the three-dot water occupies the left portion of a hanzi, from top to bottom, it might get confused in front of a three-dot water that occupies only half of the window height, even though the radical block remains at the left portion of the

character. In comparison, the brush style has more variety regarding font size and style, leading to better robustness to font variation, as shown in Table 3.

Besides the model validation with the likewise specially generated samples under well-controlled conditions, we wanted to check the identifier performance by introducing two extra variables in the general test with five ancient poems. First, we prepared the samples with a commonly found print font (PMingLiU) that was not in the training dataset. Second, each poem was saved and processed as a single JPEG image file, which required the extraction of their contents as an array of smaller images containing a single character each. Although this task was successfully implemented and returned an image array matching the input data size of the indexing radical identifier, the individual input images are not as well-controlled as the training or validation samples. The inevitable resizing operation to match the input size creates additional blurring to the hanzi. Moreover, separating the adjacent characters according to the in-line spacing property does not guarantee that the extracted hanzi will be centered in the square they belong to nor that they will not be cropped at the sides. Therefore, the general test introduced font variation, offset, and blurring to the input dataset. Under these circumstances, the indexing radical identifier showed a performance drop from ∼98.6% to 86.0% for the selected samples within the known indexing radicals, and from ∼98.6% to 61.4% if one also considers the unknown indexing radicals. The latter depends on the samples at hand, e.g., in the worst scenario of all input data belonging to unknown indexing radicals, all classifications will be wrong. Therefore, we focus on the first scenario of known cases, for which the model showed an acceptable hit rate.

The misclassifications are individually verified since there are no means of generalizing the errors. In the previous section, a case study of some misclassifications in poem A is demonstrated as a means of qualifying them. Comparing them with the results from the other four poems, we can list some of the main reasons for a misclassification:

1. The bùjiàn component of the hanzi is an indexing radical itself. For instance, “沐” is formed by the indexing radical “water” (on the left side) and the bùjiàn “wood” (on the right side), which is also an existing indexing radical.
2. The indexing radical has multiple forms, but one is much rarer than the others among the entry set. For example, “person” (人) is more common in the standing form (e.g., 何) than the regular form (e.g., 企).
3. The same form of the indexing radical assumes different positions inside a hanzi, but one is more common than the other. For example, “fire” in its traditional form (火) is more common at the left side (e.g., 燒) than at the bottom of a hanzi (e.g., 災).

In some cases, the above reasons combine to explain a single misclassification, i.e., not only is the hanzi composed of multiple blocks that are indexing radicals themselves, but the one corresponding to the bùshǒu has the rarest position or form while the other block is usual among the entries of that radical. In such cases, even the unaware human tends to mistake which is the right indexing radical.

Observing the classification results for all five poems, we verified that the three reasons above happen to known and unknown indexing radicals. Although the result is wrong, it nevertheless shows that the identifier is not random. Something, either the whole or only a part of the indexing radical block or of the hanzi, shares a common feature of another indexing radical, leading to misclassification.

4.2. Related and Future Works

The radical feature detector implemented by [20,23,24] is comparatively much more complex. For instance, Refs. [23,24] considered the hanzi as linearized hierarchical radical structures to solve the character recognition in handwriting and natural scenes, respectively. An unknown character is hence identifiable by its total decomposition into a string of sub-lexical units followed by a search in a dictionary. Ref. [20] adopted a similar concept to analyze intrinsic emotions in texts according to shared sub-lexical features. Although the hanzi is not explicitly identified in this case, it remains a hanzi recognition problem because

a closed set of sub-lexical elements narrows down the candidate characters. Therefore, despite the same terminology, the radical information they extracted is not one of the 214 indexing radicals from dictionaries, but rather the 485 sub-lexical units the Chinese characters can be decomposed into according to the GB18030 standard.

On the other hand, Refs. [17,18] classified digitized contents according to the indexing radical. For instance, Ref. [18] proposed to tag disease-related texts from non-related ones by searching for the indexing radical that groups the hanzi expressing "disease" or "illness". Albeit exploring the indexing radical information and not the sub-lexical units, both works rely on digitized input data, i.e., the hanzi are expected to be Unicode, thus requiring an extra step of digitizing a given text if it is printed material. Nowadays, OCR technology is a reliable technology to recover the contents of printed material, especially for western languages, but it still presents limitations for Chinese texts: most OCR tools are direction-sensitive, but the biggest challenge is that the Chinese characters are very similar. Moreover, OCR usually supports only the most common press font and begins to fail in detecting the contents even when the font style is just a little flourished, such as brush calligraphy [25,26].

In this sense, although the identifier we proposed is deficient to say which is the input hanzi, it does not require the input data to be digitized Unicode, making it useful when a digital version of the text is unavailable, and OCR fails. This feature is handy considering that modern smartphones all have built-in cameras, and it is often easier to take a picture of a text than to type it in. Since our hanzi extractor is based on the in-line spacing options, however, the routine requires input images with well-aligned vertical and horizontal lines between adjacent hanzi. This routine should be improved if one wants to use any image from natural scenes for varying rotation, scaling, or even randomly distributed characters. This task remains an input extraction problem and does not relate to the indexing radical classification performance.

Afterward, the proposed radical identifier has potential applications in natural language processing and multimedia Chinese learning tools when character recognition is not essential. For instance, radical knowledge contributes to text classification, information extraction, and data storage algorithms. Moreover, since it is fundamental in understanding and meaning inference, the proposed radical identifier can be implemented into a game or an application to aid hanzi learning, e.g., by highlighting the indexing radical of the hanzi from a picture. The image input is relevant in this case because a beginner learner might not have the means to input a hanzi as a raw text element. Finally, the proposed classifier can be extended to other Asian languages with script systems derived from Han characters, such as Japanese and Korean. Furthermore, one can extend the concept to ancient variants of Chinese characters, such as Tangut script or of other ancient Dynasties as well as the individual systems adopted by the Six Kingdoms before the character unification by Qin Shi Huang [33,34].

5. Conclusions

A CNN-based hanzi indexing radical identifier is demonstrated in this work for 15 indexing radicals and 3 font style families, namely regular, brush and handwriting. The identifiers were trained individually for each style and validated with all three calligraphies, achieving an average F-score of ∼95.7% in closed tests when the testing font matches the one in the training session. The performance dropped to different degrees for all identifiers when the training and testing fonts were different due to various factors such as size, thickness, and style variation. For instance, the regular-based identifier classified the brush style and handwriting samples with an average F-score of ∼83.0% and ∼41.8%, respectively. The latter is poorer because the handwriting calligraphy has a sharper size and stroke variation than the brush style, which is still closer to the regular font despite the cursive nature of some brush calligraphies. In this sense, the brush-based identifier showed better performance in different font styles, with an average F-score of ∼70% for both regular and handwriting samples since it holds both the upright boxy and the flourished properties

of these styles. Lastly, the handwriting-based identifier showed the poorest performance for the regular (∼30.9%) and brush samples (∼60.4%), probably due to the overall size variation of the training samples compared to the test set. Aside from the test for different font styles, an open, general test was designed to evaluate the identification robustness to factors such as font, offset, and blurring variables. In this scenario, the identifier showed variable F-scores because its performance depends mainly on how many indexing radicals the model knows. Therefore, it is essential to update the training dataset to encompass more classes, as all 214 indexing radicals have characters of daily use. Moreover, combining all three font style categories in one training dataset should increase the identifier robustness as well.

Author Contributions: Conceptualization, Y.T.W.; methodology, Y.T.W.; software, Y.T.W.; validation, Y.T.W.; formal analysis, Y.T.W.; investigation, Y.T.W.; resources, Y.T.W.; data curation, Y.T.W.; writing—original draft preparation, Y.T.W.; writing—review and editing, E.F. and C.K.S.; visualization, Y.T.W.; supervision, E.F.; project administration, E.F.; funding acquisition, E.F. and C.K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by CNPq, Conselho Nacional de Desenvolvimento Cientifico e Tecnologico—Brazil under the grant number 142580/2018-0.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CJK	Chinese-Japanese-Korean
CNN	Convolutional neural networks
FN	False negative
FP	False positive
OCR	Optical character recognition
TP	True positive

References

1. Lindqvist, C. *China: Empire of Living Symbols*; Addison-Wesley: Reading, MA, USA, 1991.
2. Yamasaki, T.; Hattori, T. Computer calligraphy—Brush written Kanji formation based on the brush-touch movement. In Proceedings of the 1996 IEEE International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems (Cat. No. 96CH35929), Beijing, China, 14–17 October 1996; Volume 3, pp. 1736–1741. [CrossRef]
3. Xu, S. Shuowen Jiezi [Discussing Writing and Explaining Characters]. (Eastern Han). 100–121. Available online: https://ctext.org/shuo-wen-jie-zi/zh (accessed on 11 May 2022).
4. Zhan, H.; Cheng, H.J. The role of technology in teaching and learning Chinese characters. *Int. J. Technol. Teach. Learn.* **2014**, *10*, 147–162.
5. Tao, H.; Tong, S.; Zhao, H.; Xu, T.; Jin, B.; Liu, Q. A radical-aware attention-based model for Chinese text classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5125–5132. [CrossRef]
6. Li, H.; Chen, H.C. Processing of radicals in Chinese character recognition. In *Cognitive Processing of Chinese and Related Asian Languages*; Chen, H.C., Ed.; Chinese University Press: Hong Kong, China, 1997; Chapter 8, pp. 141–160.

7. Ho, C.S.H.; Ng, T.T.; Ng, W.K. A "radical" approach to reading development in Chinese: The role of semantic radicals and phonetic radicals. *J. Lit. Res.* **2003**, *35*, 849–878. [CrossRef]
8. Tong, X.; Tong, X.; McBride, C. Radical sensitivity is the key to understanding Chinese character acquisition in children. *Read. Writ.* **2017**, *30*, 1251–1265. [CrossRef]
9. Lü, C.; Koda, K.; Zhang, D.; Zhang, Y. Effects of semantic radical properties on character meaning extraction and inference among learners of Chinese as a foreign language. *Writ. Syst. Res.* **2015**, *7*, 169–185. [CrossRef]
10. Shen, H.H.; Ke, C. Radical Awareness and Word Acquisition Among Nonnative Learners of Chinese. *Mod. Lang. J.* **2007**, *91*, 97–111. [CrossRef]
11. Wong, Y.K. The role of radical awareness in Chinese-as-a-second-language learners' Chinese character reading development. *Lang. Aware.* **2017**, *26*, 211–225. [CrossRef]
12. Li, J.; Zhou, J. Chinese character structure analysis based on complex networks. *Phys. A Stat. Mech. Appl.* **2007**, *380*, 629–638. [CrossRef]
13. Tong, X.; Yip, J.H.Y. Cracking the Chinese character: Radical sensitivity in learners of Chinese as a foreign language and its relationship to Chinese word reading. *Read. Writ.* **2015**, *28*, 159–181. [CrossRef]
14. Chen, H.C.; Hsu, C.C.; Chang, L.Y.; Lin, Y.C.; Chang, K.E.; Sung, Y.T. Using a radical-derived character e-learning platform to increase learner knowledge of Chinese characters. *Lang. Learn. Technol.* **2013**, *17*, 89–106. [CrossRef]
15. Chen, M.P.; Wang, L.C.; Chen, H.J.; Chen, Y.C. Effects of type of multimedia strategy on learning of Chinese characters for non-native novices. *Comput. Educ.* **2014**, *70*, 41–52. [CrossRef]
16. Liu, H.C. Using eye-tracking technology to explore the impact of instructional multimedia on CFL learners' Chinese character recognition. *Asia-Pac. Educ. Res.* **2021**, *30*, 33–46. [CrossRef]
17. Hu, H.; Du, X. Radical features for Chinese text classification. In Proceedings of the 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, Chongqing, China, 29–31 May 2012; pp. 720–724. [CrossRef]
18. Tan, H.; Yang, Z.; Ning, J.; Ding, Z.; Liu, Q. Chinese Medical Named Entity Recognition Based on Chinese Character Radical Features and Pre-trained Language Models. In Proceedings of the 2021 International Conference on Asian Language Processing (IALP), Singapore, 11–13 December 2021; pp. 121–124. [CrossRef]
19. The Unicode Consortium. *The Unicode Standard*; Technical Report Version 14.0.0; Unicode Consortium: Mountain View, CA, USA, 2021.
20. Ke, Y.; Hagiwara, M. Radical-level Ideograph Encoder for RNN-based Sentiment Analysis of Chinese and Japanese. In Proceedings of the 9th Asian Conference on Machine Learning, Seoul, Republic of Korea, 15–17 November 2017; Volume 77, pp. 561–573.
21. Huang, Y.; He, M.; Jin, L.; Wang, Y. RD-GAN: Few/Zero-shot Chinese character style transfer via radical decomposition and rendering. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 156–172. [CrossRef]
22. Xue, M.; Du, J.; Zhang, J.; Wang, Z.R.; Wang, B.; Ren, B. Radical Composition Network for Chinese Character Generation. In Proceedings of the International Conference on Document Analysis and Recognition, Lausanne, Switzerland, 5–10 September 2021; Springer: Berlin/Heidelberg, Germany, 2021, pp. 252–267. [CrossRef]
23. Wang, T.; Xie, Z.; Li, Z.; Jin, L.; Chen, X. Radical aggregation network for few-shot offline handwritten Chinese character recognition. *Pattern Recognit. Lett.* **2019**, *125*, 821–827. [CrossRef]
24. Zhang, J.; Du, J.; Dai, L. Radical analysis network for learning hierarchies of Chinese characters. *Pattern Recognit.* **2020**, *103*, 107305. [CrossRef]
25. Yin, Y.; Zhang, W.; Hong, S.; Yang, J.; Xiong, J.; Gui, G. Deep learning-aided OCR techniques for Chinese uppercase characters in the application of Internet of Things. *IEEE Access* **2019**, *7*, 47043–47049. [CrossRef]
26. Kim, C.; Kim, J.S.; Kim, U.J. A study on features for improving performance of Chinese OCR by machine learning. In Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference, Guangzhou, China, 22–24 June 2019; pp. 51–55. [CrossRef]
27. Chollet, F.; Scott Zhu, Q.; Rahman, F.; Lee, T.; de Marmiesse, G.; Zabluda, O.; Jin, H.; Watson, M.; Chao, R.; Pumperla, M. Keras. 2015. Available online: https://keras.io (accessed on 10 March 2022).
28. Tian, Y. zi2zi: Master Chinese Calligraphy with Conditional Adversarial Networks. 2017. Available online: https://kaonashi-tyc.github.io/2017/04/06/zi2zi.html (accessed on 10 March 2022).
29. Yan, M.; Richter, E.M.; Shu, H.; Kliegl, R. Readers of Chinese extract semantic information from parafoveal words. *Psychon. Bull. Rev.* **2009**, *16*, 561–566. [CrossRef] [PubMed]
30. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
31. Wu, Y.T.; Fujiwara, E.; Suzuki, C.K. Identification of the radical component from images of Chinese characters. In Proceedings of the 2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Natal, Brazil, 24–27 October 2022; pp. 55–60. [CrossRef]
32. Li, S.P.D.; Law, S.P.; Lau, K.Y.D.; Rapp, B. Functional orthographic units in Chinese character reading: Are there abstract radical identities? *Psychon. Bull. Rev.* **2021**, *28*, 610–623. [CrossRef] [PubMed]

33. Zhang, C.; Liu, X. Feature extraction of ancient Chinese characters based on deep convolution neural network and big data analysis. *Comput. Intell. Neurosci.* **2021**, *2021*, 2491116. [CrossRef] [PubMed]
34. Yalin, M.; Li, L.; Yichun, J.; Guodong, L. Research on denoising method of Chinese ancient character image based on Chinese character writing standard model. *Sci. Rep.* **2022**, *12*, 19795. [CrossRef]