Threshold Determination for Chinese Character Image Processing in Multimodal Information Fusion

Li Weigang
Computer Science Department
University of Brasilia
Brasília, Brazil
weigang@unb.br

Rafael Marconi Ramos
Computer Science Department
University of Brasilia
Brasília, Brazil
rafaelmr@gmail.com

Pedro Carvalho Brom
Computer Science Department
University of Brasilia
Brasília, Brazil
pcbrom@gmail.com

***Abstract*—Multimodal information fusion is gaining traction in Chinese Natural Language Processing (CNLP), particularly for phono-semantic compound comprehension and character identification. Existing research often overlooks the impact of varying pixel sizes, scales, and stroke counts on character image processing, leading to potential noise. This paper addresses this gap by analyzing our prepared dataset of Chinese characters with varying stroke counts (1-64) at different pixel resolutions (12, 16, 24, 35, 60, 96) and including up to 100 characters per stroke count. We identify a processing threshold for character images based on stroke count and resolution, a first in the field. Using Euclidean near-graphic similarity and ResNet50 image embedding similarity analyses, we establish thresholds such as 12 strokes for 16-pixel images and 26 strokes for 24-pixel images. These findings offer valuable insights for enhancing the robustness of multimodal information fusion for Chinese character recognition in NLP.**

***Index Terms*—Chinese character image, CNLP, multimodal information fusion, similarity, threshold.**

I. INTRODUCTION

The emergence of Large Language Models (LLMs) has ushered in a paradigm shift in Natural Language Processing (NLP). This shift can be viewed as two distinct eras: the pre-LLM era and the LLM era. The pre-LLM era was characterized by a diversity of approaches. Different language models relied on various artificial intelligence techniques and datasets (corpora). However, advancements in machine learning, particularly the rise of Transformer models and the pre-training of massive parameters [1], exemplified by models like ChatGPT and Le Chat, led to the dominance of large prediction models within the field [2]. Currently, in the LLM era, NLP development heavily leverages these advanced models. Researchers and developers derive specific applications for various NLP tasks by fine-tuning these powerful LLMs.

In Chinese NLP research, significant progress includes the development of Chinese LLMs by major companies and open-source initiatives like LLaMA X [3], [4]. Notable studies in this area leverage multimodal technology to enhance Chinese language processing by representing the "shape, sound, and meaning" characteristics of Chinese characters. However, issues related to the processing thresholds of images with varying pixels, scales, and strokes, which can introduce noise in multimodal information fusion, remain unresolved.

Chinese multimodal processing involves the comprehensive handling of pinyin, phonology, images, and the ideographic information of Chinese characters, extending further to words and sentences. When evaluating the similarity between modal information, computational techniques [5], the deviation in Chinese character representation range [6], and the Chinese character representation threshold should be considered [7]. Additionally, noise problems may arise during modal information processing [8]. Although these issues have been discussed to varying degrees in many previous research works, they have not been satisfactorily resolved, nor is there a unified solution.

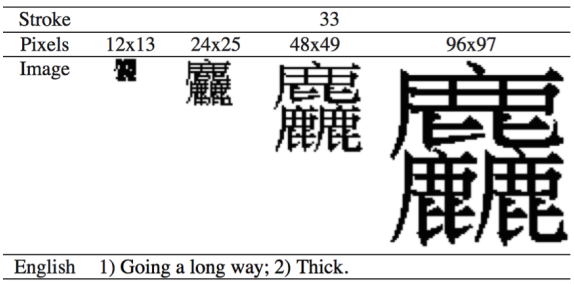

Fig. 1. A traditional Chinese character "Cu," which is composed of three character "deer." This composite character has 33 strokes and is displayed at resolutions of 12, 24, 48, and 96 pixels.

Chinese character image processing (CCIP) is a crucial aspect of Chinese multimodal processing, as thoroughly considered in previous research [9]. Human recognition of Chinese characters primarily relies on visual perception. From a language processing perspective, CCIP involves enabling

This work has been partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq) under the grant number 309545/2021-8.

machines to achieve human-like literacy. The recognition of Chinese character images depends on the scale, resolution, stroke count, and quality of the images. Figure 1 illustrates the parameters of the traditional Chinese character "Cu," which is composed of three character "deer." This composite character has 33 strokes and is displayed at resolutions of 12, 24, 48, and 96 pixels. As the resolution decreases (fewer pixels), recognizing the character becomes increasingly difficult. Conversely, higher resolutions incur greater computational costs. Therefore, it is essential to analyze the relationship between character stroke count and image resolution to determine an appropriate threshold and minimize noise in Chinese NLP.

This paper addresses this gap by exploring the analysis method for image processing thresholds based on the similarity between Chinese character images, using deep learning and other techniques. The study focuses on Song-style Chinese characters from the "Full Character Library (FCL)," [10] with each stroke count ranging from 1 to 64 strokes, selecting 100 characters per stroke count (or the available number of characters if fewer than 100). Experimental research was conducted on these images at sizes of 12, 16, 24, 35, 60, and 96 pixels (px). We analyze the relationships between strokes, pixels, and word frequency using Euclidean near-graphic similarity and ResNet50 image embedding similarity analyses. Our results establish a processing threshold of 12 strokes for 16-pixel characters, 26 strokes for 24-pixel characters, and others.

This study's primary contribution is a novel procedure for identifying the processing threshold of Chinese characters based on pixel size, stroke and frequency. We establish threshold values for various image resolutions, providing valuable insights for optimizing multimodal information fusion in Chinese character recognition for CNLP.

II. RELATED WORK

This section reviews recent research on Chinese multimodal NLP, addressing key issues related to Chinese character image processing, including similarity calculation and processing thresholds.

He and Schomaker (2018) presented a method for open set Chinese character recognition using multi-typed attributes, leveraging various attributes of Chinese characters to create a comprehensive representation [11]. In 2020, Cao et al. proposed a zero-shot learning framework for handwritten Chinese character recognition, utilizing hierarchical decomposition embedding to capture the structure of characters [9]. This method significantly improves recognition accuracy in zero-shot scenarios by decomposing unseen characters into known components.

Sun et al. (2021) introduced ChineseBERT, a pretraining model integrating both glyph and Pinyin information to enhance Chinese language understanding [12]. By incorporating these multimodal features, ChineseBERT achieves superior performance in various NLP tasks compared to models relying solely on textual data.

Cui et al. (2023) proposed an efficient text encoding method for Chinese LLaMA and Alpaca models, focusing on optimizing the encoding process to handle Chinese text more effectively [4]. This enhancement in text encoding mechanisms leads to better performance in understanding and generating Chinese text.

HierCode, a lightweight hierarchical codebook was designed for zero-shot Chinese text recognition [13]. This method leverages a hierarchical structure to encode Chinese characters efficiently, enabling the recognition of unseen characters.

Weigang et al. (2024) proposed Six-Writings, a multimodal processing framework incorporating pictophonetic coding to enhance Chinese language models [6]. The pictophonetic coding integrates visual and phonetic features, improving the model's ability to understand and generate Chinese text accurately.

Li et al. (2024) introduced DRMSpell, a dynamically reweighting multimodal framework for Chinese spelling correction [14]. This method dynamically adjusts the weights of different modalities to correct spelling errors in Chinese text, demonstrating high accuracy in identifying and correcting spelling mistakes.

These models extract the morphological feature by traditional image processing methods instead of linguistic knowledge, which introduces the connection of characters, and the noise from the unclarity of image processing algorithms [8].

III. BASIC CORPORA

In this research, we will use two types of corpora: text corpora and Chinese character images. Two types of text corpora have been prepared:

- The ZH-SC8105 corpus is derived from the "Table of General Standard Chinese Characters"[1] issued in 2013 by the State Language Commission. This corpus includes 8,105 characters with associated stroke and frequency information.
- The ZH-TC96858 corpus is built upon the CNS11643 Chinese standard interchange code in "Full Character Library (FCL)"[2] used in Taiwan province. It includes 96,858 traditional Chinese characters with associated stroke and frequency information.

Chinese character image corpus is derived from traditional Chinese characters in "Full Character Library (FCL)". A web crawler was used to collect character images from website of

[1]https://www.hanyuguoxue.com/zidian/guifanhanzi
[2]https://www.cns11643.gov.tw/

this library [10], forming the ZH-TC-IM 96858 corpus. This corpus includes 96,858 Chinese characters, with stroke counts k ranging from 1 to 64. The collection process for Chinese character images followed these methods:

- For each stroke count, up to 100 Chinese character images were captured. If there are fewer than 100 characters for a particular stroke count, the actual number of characters is used (e.g., for 64-stroke characters, only two are available).
- Each captured Chinese character image is processed into the following pixel sizes: 12x13, 16x17, 24x25, 35x36, 36x37, 48x47, 60x67, and 96x97 pixels. For simplicity, the image pixel sizes will be abbreviated, for example, with 12x13 pixels referred to as 12px, and others. For each size, the black-and-white pixel ratio is calculated. The average black-and-white pixel ratio for Chinese characters with k strokes is then determined based on the collected characters. In most cases, for each k, the mean calculation is based on up to 100 characters.

Figure 2 illustrates the black-and-white pixel ratio of Chinese character images at different pixel sizes and the distribution of character strokes. For example, the red curve labeled "perc16" represents the black-and-white pixel ratio of 16-pixel Chinese character images, showing the stroke distribution. The average black-and-white pixel ratio for single-stroke characters is 0.0971, while for characters with 43 strokes, it is 0.9128. Similarly, the green curve labeled "perc24" represents 24-pixel Chinese character images, with an average black-and-white pixel ratio of 0.1 for single-stroke characters and 0.8645 for characters with 64 strokes.

Overall, for images with smaller pixel sizes, the average black-and-white pixel ratios tend to be higher. The curves generally decrease in order from 12, 16, 24, 35, 36, 60, to 96 pixels. However, the curves for 32 and 48 pixels do not follow this pattern, and their average black-and-white pixel ratios are relatively lower.

IV. METHODS

This section outlines the methods used in this research, specifically focusing on Euclidean similarity and ResNet50 embedding similarity methods. We use a 96 px image as a benchmark and calculate the similarity between this benchmark and images of various resolutions, such as 12 px (12x13 pixels), 16 px (16x16 pixels), 24 px (24x25 pixels), and others.

A. The Euclidean Distance Similarity between Two Images

The Euclidean distance is a measure of the similarity between two images by calculating the straight-line distance between their feature vectors in a multi-dimensional space [15]. Given two images I_1 and I_2, let their feature vectors be represented as $\mathbf{f}_1$ and $\mathbf{f}_2$, respectively.

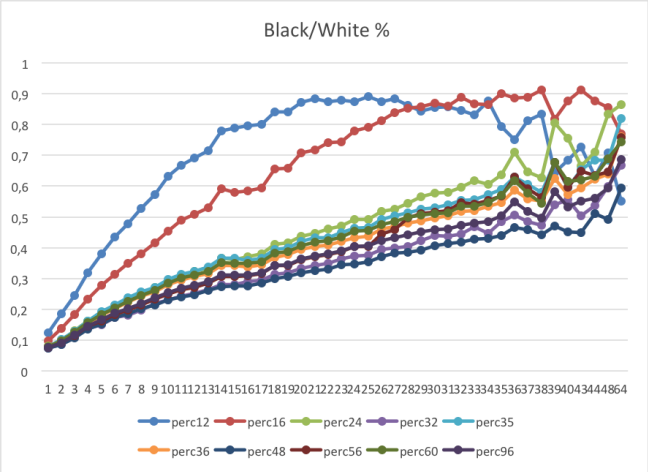

Fig. 2. The black-and-white pixel ratio of Chinese character images at different pixel sizes along the strokes from ZH-TC-IM 96858 corpus.

If the feature vectors are of dimension n, then:

$$\mathbf{f}_1 = [f_{1,1}, f_{1,2}, \ldots, f_{1,n}]; \mathbf{f}_2 = [f_{2,1}, f_{2,2}, \ldots, f_{2,n}]$$

The Euclidean distance d between these two feature vectors is calculated using the Equation 1:

$$d(\mathbf{f}_1, \mathbf{f}_2) = \sqrt{\sum_{i=1}^{n} (f_{1,i} - f_{2,i})^2} \tag{1}$$

The Euclidean distance d quantifies the similarity or dissimilarity between two images based on their feature vectors. A smaller distance indicates greater similarity, while a larger distance indicates greater dissimilarity. Figure 3 shows the results of the similarity between the benchmark (96px, as image I_1) and images of various resolutions, such as 12px (as image I_2, see SE12x96perc curve in the figure), 16px (SE16x96perc), 24px (SE24x96perc), and others. The highest similarity (1.00) is observed between the benchmark and itself. The lowest similarity is observed between the benchmark and the 12x13 pixel image. The trend also clearly shows that as the number of strokes k increases, the similarity decreases.

B. ResNet50 Embedding Similarity Between Two Images

ResNet50 is a deep convolutional neural network that is widely used for image recognition tasks. It consists of 50 layers and employs residual connections to improve training and performance. The similarity between two images can be evaluated by comparing their feature embeddings extracted from a pre-trained ResNet50 model [16].

Given two images I_1 and I_2, we pass them through the ResNet50 network to obtain their respective feature embeddings. Let the embeddings be represented as $\mathbf{e}_1$ and $\mathbf{e}_2$.

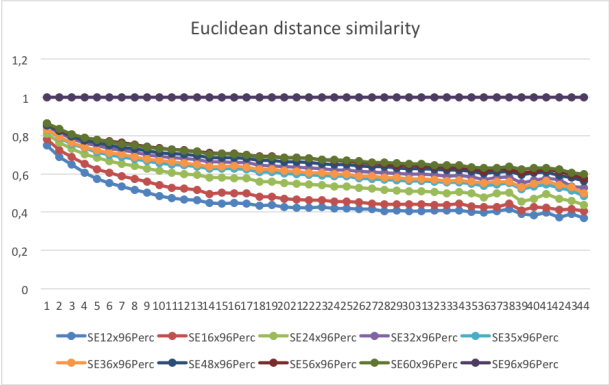

Fig. 3. The Euclidean distance similarity between two images.

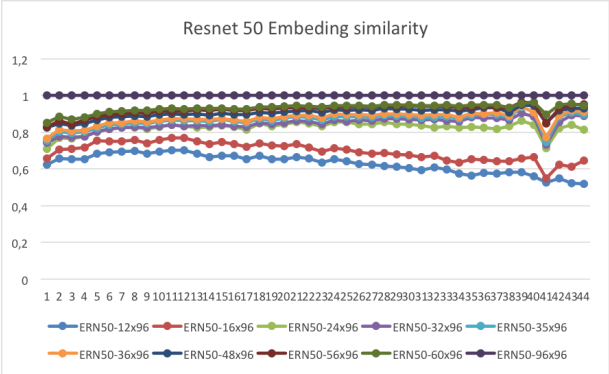

Fig. 4. The ResNet50 embedding similarity between two images.

The feature embeddings are typically high-dimensional vectors extracted from the last fully connected layer before the softmax layer of the network. Suppose the dimension of these embeddings is d. Then:

$$\mathbf{e}_1 = [e_{1,1}, e_{1,2}, \ldots, e_{1,d}]; \mathbf{e}_2 = [e_{2,1}, e_{2,2}, \ldots, e_{2,d}]$$

To measure the similarity between these two embeddings, we use the Euclidean distance, which is given by Equation 2:

$$d(\mathbf{e}_1, \mathbf{e}_2) = \sqrt{\sum_{i=1}^{d} (e_{1,i} - e_{2,i})^2} \tag{2}$$

Alternatively, we can also use the cosine similarity, which measures the cosine of the angle between two vectors. In practice, these similarity measures can be used to compare the embeddings of two images and determine how similar they are in the feature space learned by ResNet50. A higher cosine similarity or a lower Euclidean distance indicates greater similarity between the images.

Figure 4 shows the results of the similarity between the benchmark (96px) and images of various resolutions, such as 12px (SE12x96perc curve in the figure), 16px (SE16x96perc), 24px (SE24x96perc), and others. Similar to the Euclidean distance similarity, the highest similarity (1.00) is observed between the benchmark and itself, while the lowest similarity is observed between the benchmark and the 12x13 pixel image. The difference between Figures 3 and 4 is that the trend of decreasing similarity with increasing number of strokes k is observed only in the cases of 12px and 16px.

V. MODELING AND EXPERIMENTS

This section first introduces the character length (strokes) and character frequency (CLCF) model, followed by the character splitting (GCLCF) model. Subsequently, we present the Chinese character image binary ratio and the Chinese character frequency (CIBRCF) model.

A. CLCF and character splitting models

The relationship index between Chinese character length (strokes) and character frequency is defined as CLCF [7]. It describes the number of strokes k in a Chinese character and the evolutionary pattern of character frequency based on the stroke count, as illustrated in Equation 3:

$$CLCF = \int_1^K log_2(k) dP(k) \tag{3}$$

Here's a breakdown of the notation used: k: Number of strokes in a character; K: Maximum number of strokes found in characters within the corpus; $n(k)$: Number of characters in the corpus with k strokes; N: Total number of characters in the corpus; $p(k)$: Character frequency for k strokes $p(k) = n(k)/N$; $P(k)$: accumulated Character frequency for $k = 1, 2, .., K, P(k) = \sum p(k)$. For example, consider the ZH-SC8105 corpus, where the maximum number of strokes K is 36 and the total number of characters N is 8105.

Based on the CLCF model, the granularity index $GCLCF$ of a Chinese character splitting method is defined as [7]:

$$GCLCF = \int_1^K \frac{log_2(k)}{log_\theta(k)} dP(k) \tag{4}$$

θ is the number of strokes that Chinese character coding can represent. θ_T is the threshold and its value is based on the coding method of Chinese character representation, such as Four-corner number, Wubi, Cangjie, and other codes. GCLCF is the first model to indicate the Chinese coding threshold in CNLP [7].

B. CIBRCF model

Chinese character images are generally composed of black and white pixels. Therefore, we define two parameters: 1) ρ is the black-and-white pixel ratio, if the black pixel area is B and the white pixel area is W, $\rho = B/W$; 2) λ is the ratio of

black pixels to total pixels, $\lambda = B/(B+W)$. For certain stroke number k, there is $\lambda(k) = B(k)/(B(k)+W(k))$. The Chinese character image binary ratio and the character frequency model, CIBRCF, can be defined as:

$$CIBRCF = \int_1^K log_2(\lambda(k) \times 100) dP(k) \tag{5}$$

Most of the parameters in the Equation 5 are the same as those in Equations 3 and 4. $log_2(\lambda(k) \times 100)$ is to avoid only $log_2(\lambda(k))$ taking a negative value. $dP(k)$ is same as shown in Equation 3.

C. Chinese character image processing threshold model

For a dataset comprising N Chinese characters, with a maximum stroke count of K, the threshold index $TCIBRCF$ of a Chinese character image processing is defined as:

$$TCIBRCF = \int_1^K \frac{log_2(\lambda(k) \times 100)}{log_\theta(\lambda(k) \times 100)} dP(k) \tag{6}$$

θ represents the number of strokes in a Chinese character image of a certain scale. Its value is determined based on the black-to-white (B/W) ratio of the Chinese character. Depending on the scale and the number of strokes, there is a threshold θ_T that needs to be determined. This threshold helps in establishing the processing limits for Chinese character images at different resolutions and stroke counts.

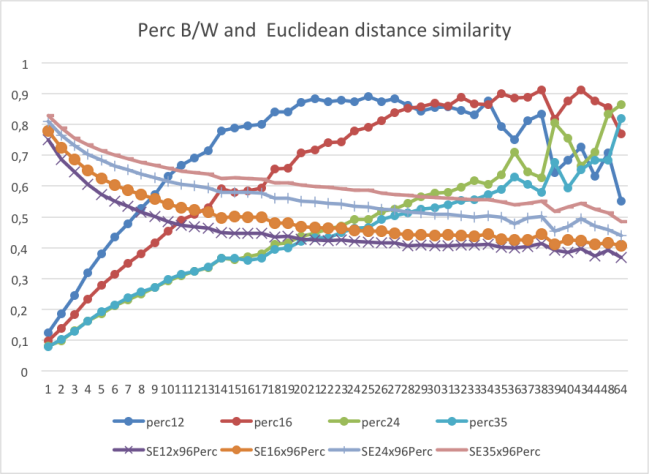

Fig. 5. To determine the threshold by B&W ratio and similarity curves.

Based on the above analyses, we design a procedure to determine the threshold k_T and then θ_T.

- *The curve of the average black-and-white pixel ratio ($\rho = B/W$) of Chinese characters with different strokes:* List the average black-and-white pixel ratio ($\rho = B/W$) curves for 100 Chinese characters with different strokes ($k = 1, ..., 64$). For each stroke count, if there are fewer than 100 characters, use the actual number available. Refer to Figure 2. In Figure 5, the x-axis represents the number of strokes, and the y-axis represents the ratio from 0 to 1. The curves for perc12 (12px image), perc16 (16px image), perc24 (24px image), and perc35 (35px image) are shown.
- *Euclidean similarity curve between different pixel sizes of different strokes ($k = 1, ..., 64$) and 96 pixels for each stroke:* List the Euclidean similarity curves between the images of different pixel sizes and the 96-pixel images for each stroke count k, using 100 samples per stroke if available, refer to Figure 3. In Figure 5, SE12x96perc (Euclidean similarity between the 12px image and the 96px image for characters with k strokes), SE16x96perc, SE24x96perc, and SE35x96perc are shown.
- *The intersection of the above two kinds of curves for the same pixel size determines the threshold of Chinese character processing at that pixel size:* For example, in Figure 4, the intersection of perc35 (35px image) and SE35x96perc (35px image) occurs at stroke k value of 33, indicating that the processing threshold k_T for 35-pixel Chinese character images is 33 strokes. Similarly, the intersection of perc24 (24px image) and SE24x96perc (24px image) occurs at stroke k value of 26, so the processing threshold k_T for 24-pixel Chinese character images is 26 strokes. Refer to Figure 5 for other results.

After determining the threshold for CCIP, the results from Equations 5 and 6 can be visualized in Figure 6. In this figure, the x-axis represents the cumulative frequency of Chinese characters, and the y-axis represents $log_2(\lambda(k) \times 100)$ as described in Equation 5.

The blue curve in the figure corresponds to the 12-pixel image data, with the dashed line indicating its threshold. This threshold signifies that representing Chinese characters with a 12-pixel image becomes problematic after 8 strokes. Similarly, the green curve represents the 16-pixel image data, with its dashed line showing a threshold of 12 strokes. For the ZH-SC8105 corpus, 30.76% of Chinese characters have more than 12 strokes, while for the ZH-TC96858 corpus, 59.27% of Chinese characters exceed 12 strokes. Thus, recognizing Chinese characters with 16-pixel images is generally infeasible.

Figure 7 offers a detailed view of the horizontal axis for cases with k more than 20 strokes (where the cumulative character frequency exceeds 90.62%). The orange curve in this figure represents the 24-pixel image data, with a threshold indicated by the dashed line. After 26 strokes, representing Chinese characters with a 24-pixel image becomes problematic. For the ZH-TC96858 corpus, 1.1% of Chinese characters exceed 26 strokes. The red curve represents the 35-pixel image data. It indicates that representing Chinese characters with 35-

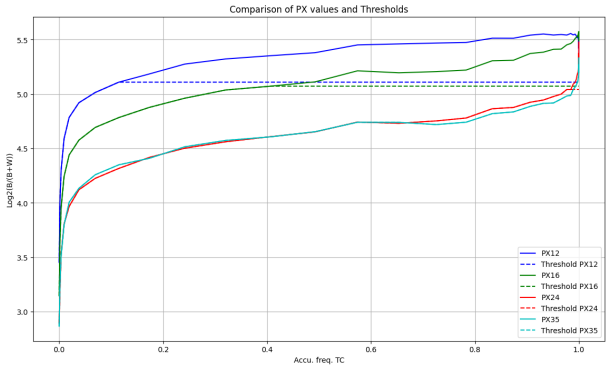

Fig. 6. CCIP threshold: 8 strokes for 12px and 12 strokes for 16px.

pixel images becomes problematic after 33 strokes. However, since characters with more than 33 strokes are rare (less than 0.1% cumulative frequency), the threshold is not marked with a dashed line in Figure 7.

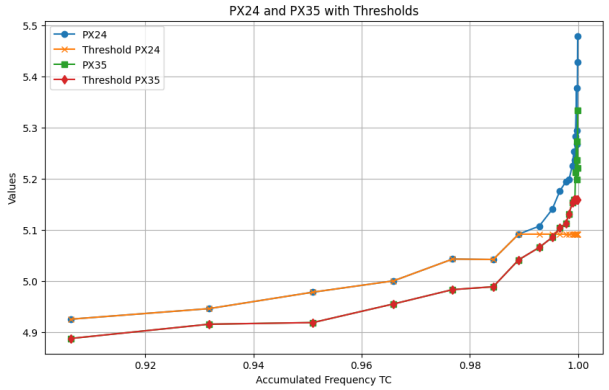

Fig. 7. CCIP threshold: 26 strokes for 24px and 33 strokes for 35px.

VI. CONCLUSIONS

This study investigated the processing thresholds for Chinese character images, considering both stroke count and pixel size. We employed Euclidean near-graphic similarity and ResNet50 image embedding analyses to explore the relationships between these factors and character frequency. Our findings reveal CCIP thresholds of 8 strokes for 12-pixel, 12 strokes for 16-pixel, 26 strokes for 24-pixel, and 33 strokes for 35-pixel images of the related characters, respectively. Notably, for ZH-SC8105, 30.6% of characters have more than 12 strokes, and for ZH-TC96858, this percentage is 59.27%, indicating that 16-pixel image processing needs to pay more attention to the processing threshold.

Weigang et al. (2024) proposed the representation threshold of Chinese character coding [7]. Building on this, we now introduce the processing threshold of Chinese character images in multimodal processing. These results provide valuable insights for optimizing Chinese character image processing and inform the development of robust multimodal information fusion approaches in CNLP.

Future research could extend the analysis to include more than 100 characters for each stroke count. For example, in the ZH-TC96858 corpus, there are 8,406 characters with 12 strokes, which could improve the generalizability of the findings.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] J. Zhou, P. Ke, X. Qiu, M. Huang, and J. Zhang, "Chatgpt: potential, prospects, and limitations," *Frontiers of Information Technology & Electronic Engineering*, pp. 1–6, 2023.

[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[4] Y. Cui, Z. Yang, and X. Yao, "Efficient and effective text encoding for chinese llama and alpaca," *arXiv preprint arXiv:2304.08177*, 2023.

[5] H. Steck, C. Ekanadham, and N. Kallus, "Is cosine-similarity of embeddings really about similarity?" in *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 887–890.

[6] L. Weigang, M. C. Marinho, D. L. Li, and V. V. De Oliveira, "Six-writings multimodal processing with pictophonetic coding to enhance chinese language models," *Frontiers of Information Technology & Electronic Engineering*, vol. 25, no. 1, pp. 84–105, 2024.

[7] L. Weigang, P. C. Brom, D. L. Li, and V. Di Oliveira, "Llm-swpc: A new paradigm for large language models using six-writings pictophonetic representation," *submitted to Frontiers of Information Technology & Electronic Engineering*, 2024.

[8] H. Jin, Z. Zhang, and P. Yuan, "Improving chinese word representation using four corners features," *IEEE Transactions on Big Data*, vol. 8, no. 4, pp. 982–993, 2021.

[9] Z. Cao, J. Lu, S. Cui, and C. Zhang, "Zero-shot handwritten chinese character recognition with hierarchical decomposition embedding," *Pattern Recognition*, vol. 107, p. 107488, 2020.

[10] S.-J. Wu, C.-Y. Yang, and J. Y.-j. Hsu, "Calligan: Style and structure-aware chinese calligraphy character generator," *arXiv preprint arXiv:2005.12500*, 2020.

[11] S. He and L. Schomaker, "Open set chinese character recognition using multi-typed attributes," *arXiv preprint arXiv:1808.08993*, 2018.

[12] Z. Sun, X. Li, X. Sun, Y. Meng, X. Ao, Q. He, F. Wu, and J. Li, "Chinesebert: Chinese pretraining enhanced by glyph and pinyin information," *arXiv preprint arXiv:2106.16038*, 2021.

[13] Y. Zhang, Y. Zhu, D. Peng, P. Zhang, Z. Yang, Z. Yang, C. Yao, and L. Jin, "Hiercode: A lightweight hierarchical codebook for zero-shot chinese text recognition," *arXiv preprint arXiv:2403.13761*, 2024.

[14] L. Yinghao, H. Heyan, W. Baojun, and G. Yang, "Drmspell: Dynamically reweighting multimodality for chinese spelling correction," *Frontiers of Information Technology & Electronic Engineering*, 2024.

[15] L. Wang, Y. Zhang, and J. Feng, "On the euclidean distance of images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1334–1339, 2005.

[16] P. S. Satya Sreedhar and N. Nandhagopal, "Classification similarity network model for image fusion using resnet50 and googlenet." *Intelligent Automation & Soft Computing*, vol. 31, no. 3, 2022.