Ancient Chinese Glyph Identification Powered by Radical Semantics

Yang Chi[1], **Fausto Giunchiglia**[1,2,3], **Chuntao Li**[4,5,*], **Hao Xu**[1,2,5,*]
[1]School of Artificial Intelligence, Jilin University, Changchun, China
[2]College of Computer Science and Technology, Jilin University, Changchun, China
[3]DISI, University of Trento, Trento, Italy
[4]School of Archaeology, Jilin University, Changchun, China [5]Key Laboratory of Ancient Chinese Script, Culture relics and Artificial Intelligence, Jilin University, Changchun, China
yangchi19@mails.jlu.edu.cn, {xuhao,lct33}@jlu.edu.cn
fausto.giunchiglia@unitn.it

Abstract

The ancestor of Chinese character – the ancient characters from about 1300 BC to 200 BC are not fixed in their writing glyphs. At the same or different points in time, one character can possess multiple glyphs that are different in shapes or radicals. Nearly half of ancient glyphs have not been deciphered yet. This paper proposes an innovative task of ancient Chinese glyph identification, which aims at inferring the Chinese character label for the unknown ancient Chinese glyphs which are not in the training set based on the image and radical information. Specifically, we construct a Chinese glyph knowledge graph (CGKG) associating glyphs in different historical periods according to the radical semantics, and propose a multimodal Chinese glyph identification framework (MCGI) fusing the visual, textual, and the graph data. The experiment is designed on a real Chinese glyph dataset spanning over 1000 years, it demonstrates the effectiveness of our method, and reports the potentials of each modality on this task. It provides a preliminary reference for the automatic ancient Chinese character deciphering at the glyph level.

1 Introduction

The ancient Chinese characters before 200 BC are largely different from the modern ones, which mainly include the Oracle bone script (Oracle) in about 1300 BC (Boltz, 1986), the Chinese bronze script (Bronze, about 1000 BC) (Shaughnessy, 1991) and the script belonging to the Warring States period (States, about 400 BC) (Qiu, 2014).

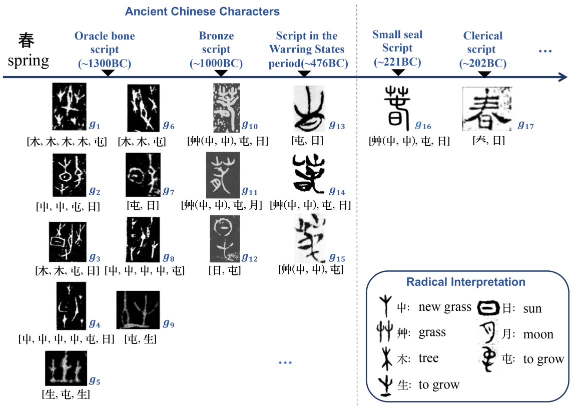
Ancient Chinese Characters
春
spring
Oracle bone script (~1300BC)
Bronze script (~1000BC)
Script in the Warring States period(~476BC)
Small seal Script (~221BC)
Clerical script (~202BC)
...
g_1 [木, 木, 木, 木, 屯]
g_6 [木, 木, 屯]
g_{10} [艸(屮, 屮), 屯, 日]
g_{13} [屯, 日]
g_{16} [艸(屮, 屮), 屯, 日]
g_{17} [𡗗, 日]
g_2 [屮, 屮, 屯, 日]
g_7 [屯, 日]
g_{11} [艸(屮, 屮), 屯, 月]
g_{14} [艸(屮, 屮), 屯, 日]
g_3 [木, 木, 屯, 日]
g_8 [屮, 屮, 屮, 屮, 屯]
g_{12} [日, 屯]
g_{15} [艸(屮, 屮), 屯]
g_4 [屮, 屮, 屮, 屮, 屯, 日]
g_9 [屯, 生]
g_5 [生, 屯, 生]
...
Radical Interpretation
屮: new grass
艸: grass
木: tree
生: to grow
日: sun
月: moon
屯: to grow

Figure 1: The 17 representative glyphs (g_1 - g_{17}) of the Chinese character "春" (spring) which are distributed on 5 time stages of character evolution (Oracle, Bronze, States, Small Seal and Clerical). Their radicals are annotated below the image.

The glyphs of ancient character are not fixed. Figure 1 shows 17 glyphs of the character "春" (spring) distributed in 5 time stages of evolution (Oracle, Bronze, States, Small Seal and Clerical). Although they look different, there are potential semantic relations between their radicals: their radical systems both express the "spring" meaning by portraying a picture of "plants are growing under the sun". Thus, many of the glyphs contain semantic radicals of "屮" (grass), "木" (tree) and "日" (sun), and nearly all of glyphs contain the phonetic radical "屯" ("屯" and "春" are homophones).

At present, about half of the ancient glyphs have not been deciphered, which means that we do not know the modern Chinese characters they corresponded, and can not completely interpret the historical documents containing them. Experts have to compare all related glyphs in history for reasoning. However, there are tens of thousands of glyphs distributed in various time stages with complex relationships between their radicals and shapes, which makes deciphering a difficult work that heavily relies on human memory and experience.

Please imagine a real scenario: given a unknown ancient Chinese glyph, we can obtain its images and radical information, and learn the radical semantics and potential usage patterns from the known glyphs in history. Then, can we automatically speculate the Chinese character to which the unknown glyph belongs? And to what extent the

[*] Corresponding authors

Findings of the Association for Computational Linguistics ACL 2024, pages 12065–12074
August 11-16, 2024 ©2024 Association for Computational Linguistics

visual and radical information can support this task respectively? Answering these questions has significance for ancient Chinese character deciphering.

In this background, our contributions include: (1) we propose an innovative task of ancient Chinese glyph identification, to infer the Chinese character label for unknown glyphs in training set based on the image and radical information (Figure 2); (2) we devise a radical semantics powered multimodal method, including constructing a Chinese glyph knowledge graph (CGKG) and proposing a glyph identification framework (MCGI) fusing the visual, textual and graph features; (3) we evaluate the method on a glyph dataset spanning over 1000 years in both synchronic and diachronic views. It proves the validity of our method compared with the baselines, and gives the conclusion of the contributions for each modality to this task.

This paper provides a preliminary reference for the ancient character deciphering. It has potential applications in ancient Chinese, ancient character, history, and other related fields. For instance, to help experts discover the top-k possible character labels for the unknown glyph, as well as their relevant glyphs in history; and to improve the effects of optical character recognization through radical semantics.

The organization of this paper is as follows: Section 2 presents the related works; Section 3 describes the key knowledge and the task definition; Section 4 introduces CGKG; Section 5 introduces our glyph identification framework (MCGI); the evaluation is proposed in Section 6; Section 7 describes the limitations and future works; and Section 8 concludes this paper.

2 Related Works

Related works on Chinese characters: Most of the related works concentrated on optical character recognition (OCR) (Cao et al., 2020; Diao et al., 2023; Huang et al., 2022; Zhang et al., 2018). However, different from them, our focus is not on image recognition, but to identify the unknown ancient glyphs based on image and radical semantics. Chi et al. (2022) introduced an ancient Chinese glyph similarity measurement method that extracted radical semantic features from the graph, but the method is not designed for glyph identification. Zhang et al. (2021) presented an unknown image as a query, to receive a set of similar images from an adjacent writing system with associated scholarly information, and so help guide the deciphering of the query. Chang et al. (2022) presented an image-to-image translation networks to generate corresponding modern Chinese forms by simulating and restoring the real historical evolutionary process of Oracle characters, which enables archaeologists to use the generated modern characters to infer possible lexical natures of Oracle characters. Although these works also can serve for deciphering, the specific tasks are not identical with ours. And compared with these works, because of introducing radical semantics, our method is more suitable for identifying the unknown glyphs who have different radical composition to their corresponding modern glyph or other variant glyphs. It is consistent with the theory of human deciphering: in the deciphering works, experts not only need to compare with visual similar glyphs but also discover and reason those glyphs that are related in semantics and pronunciations based on the radicals.

Multimodal Methods: the related works are distributed in domains of vision-and-language models (Kim et al., 2021; Radford et al., 2021; Xu et al., 2023), NLP and multimodal knowledge graph representation (Li et al., 2023; Wang et al., 2022; Zhang et al., 2023). Most of the multimodal models follow an universal workflow: generating the embedding for each modality through the pre-trained uni-modal encoders, such as BERT (Devlin et al., 2019), the Deep Residual Network (ResNet) (He et al., 2016), and the Graph Attention Network (GAT)(Velickovic et al., 2018), respectively for textual, visual and graph modality. These embeddings will be fed into a cross-modal encoder for data fusion, which can be a simply dot product operation, multimodal attention mechanisms or more complicated transformers. ViLT (Kim et al., 2021) took shallow embedding layers and used transformer on textual and visual embedding interactions. In the knowledge graph domain, HRGAT (Wang et al., 2022) proposed a multimodal fusion method based on text and visual co-attention. IMF (Li et al., 2023) proposed a two-stage interactive multimodal fusion framework for link prediction.

3 Key Information and Task Definition

3.1 Chinese Radicals

Chinese characters are composed of radicals, which are also characters or variants of character. The number of radicals is less than that of character significantly because they are commonly shared by different glyphs. Radical can express the semantic or phonetic information, therefore, it is the smallest units to describe Chinese glyphs.

3.2 Variant Glyphs

The variant glyphs (异体字) refer to a group of glyphs with the same meaning, pronunciation, usage but different in their graphics. The representative example is the traditional Chinese character (e.g., "鳥", bird) and the corresponding simplified Chinese character (e.g.,"鸟"). They belong to the same Chinese character label in this paper.

We divide the variant glyphs based on the following three indicators: (1) **different in radical types**; (2) **different in radical quantity, location or glyph** (e.g., "虎"(tiger) and "厇"(tiger)); (3) **different in living time stages**. We take the Oracle, Bronze and States as three ancient time stages, and others after them as the Modern stage.

In addition, we call a group of variant glyphs as the **synchronic variant glyphs** if they lived in the same time stage, otherwise, we call them **diachronic variant glyphs**. Correspondingly, the testing glyphs in this work is divided into two groups: the **synchronic testing glyph**, if it has at least one synchronic variant glyphs in the training set, otherwise, it is **diachronic testing glyph**. We will separately show the results of the two groups in evaluation, because the latter is more difficult to be identified and is common in practical applications.

3.3 ZiNet Knowledge Base

The available data for this work is limited, because most of the Chinese character image resources do not integrate at the glyph level. The dataset of this work (Section 6.1) is from ZiNet (Chi et al., 2022). ZiNet is a diachronic knowledge base describing relationships and evolution of Chinese characters. Up to now, it records up to 16000 glyphs in various historical periods of Oracle, Bronze and States, and associates them with the character, radicals and images.

3.4 Task Definition

Our task is shown in Figure 2. In this task, there is a character set $C = \{c_i|i = 1, 2, ..., |C|\}$ containing Chinese character labels; a training glyph set $G_{train} = \{g_i|i = 1, 2, ..., |G_{train}|\}$ containing Chinese glyphs distributed in all of time stages; and a testing glyph set $G_{test} = \{g_i|i = 1, 2, ..., |G_{test}|\}$ containing Chinese glyph samples in ancient time stages, which is unseen for the model, $G_{train} \cap G_{test} = \emptyset$. Each Chinese glyph belongs to one Chinese character label. Thus, the training dataset is $\{(g, c)|g \in G_{train}, c \in C\}$ and the testing dataset is $\{(g, c)|g \in G_{test}, c \in C\}$.

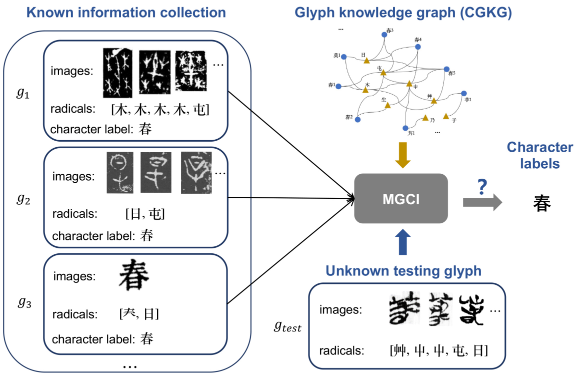

Figure 2: Definition of the ancient Chinese glyph identification task in this paper.

The Chinese glyph in this task is described by two kinds of data: the images and a short radical description text (Figure 2). Each glyph g_i has one or more than one images, it's image set is defined as P_{g_i}. And it also has a radical sequence R_{g_i}, which contains the radicals of ranking in the writing order.

There is a graph KG, $KG = (V, Rel)$, $V = V_g \cup V_r$, where V is the entity node set, V_g is the glyph entity node set and V_r is the radical entity set. Rel is the edge set between entities. KG contains all glyphs in G_{train} and their radicals, as well as the edges between them.

The goal of this task is to predict the corresponding character label c, $c \in C$ for the glyph g, $g \in G_{test}$: $f(P_g, R_g, KG|\theta) = c$, where θ is the parameters of the model gotten from trainings. All of the supervised learnings are based on the training dataset and KG.

4 Chinese Glyph Knowledge Graph

As shown in Figure 3, the CGKG has two types of entities of glyph and radical, and two relations: the inclusion relation between glyph and radical $Rel_1(g, r)$, and the semantic relation between two radicals $Rel_2(r, r)$. Two radicals would be linked with Rel_2 if they met the following relationships:

- They belong to the same character (e.g., "虎"(tiger) and "厇"(tiger)).
- They are derived from the same mother character (e.g., "東" (bag) and "束"(tie)).
- They have the indication relationship (指事) (e.g., "生" (to grow) and "屮"(grass)).

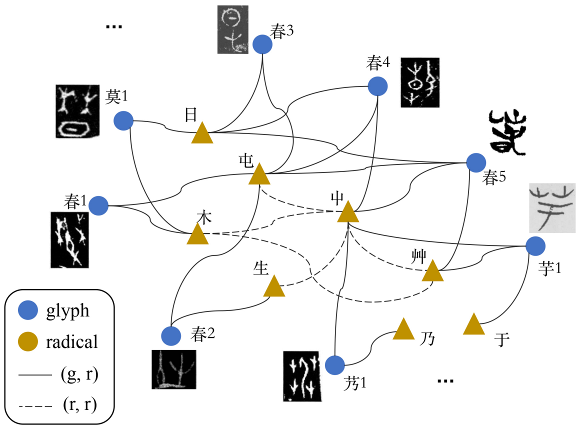

Figure 3: The Chinese glyph knowledge graph (CGKG), which contains two types of entities: glyph (g) and radical (r); and two relations: $Rel(g, r)$ and $Rel(r, r)$.

- They express the same or similar meaning (e.g., “屮”(grass) and “木”(tree)).
- They have the same or similar pronunciation and usually can be mutual borrowing used (通假) in the ancient Chinese (e.g., “匕” (female ancestor) and “比”(close to each other)).
- They are the interchangeable radicals (“屮” (grass) and “木” (tree) is a pair of interchangeable radical observed in the variant glyph pair “屮, 屮, 屯, 日” and “木, 木, 屯, 日” of the character “春”).

The first five relations were annotated by our experts, and the last one was automatically extracted from dataset. We automatically extracted interchangeable radical pairs from all of variant glyph pairs in dataset: if only the target radical pair are replaced with each other and other parts are the same, they will be set as the candidate, the radical pairs appeared over 2 times will be added into KG.

These relations are not independent but highly relevant. We do not specify the relationship categories, if two radicals satisfy any of above professional indicators, we will add an edge between them in KG, and finally construct an undirected graph. Ultimately, there are 1907 edges between 657 radical nodes in KG.

5 Chinese glyph identification Method

5.1 Method Overview

The architecture of MCGI is shown in Figure 4. Given a glyph g_k, the inputs include an image set P_{g_k}, a radical token sequence T_{g_k}, which is the textual representation of R_{g_k} and the graph KG.

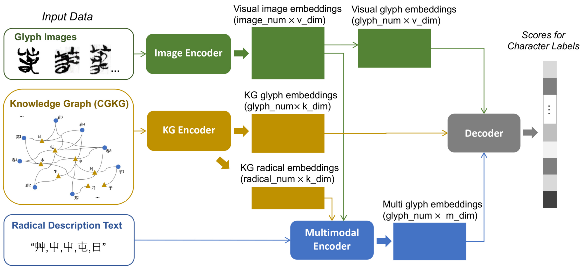
Input Data
Glyph Images
Image Encoder
Visual image embeddings (image_num × v_dim)
Visual glyph embeddings (glyph_num × v_dim)
Scores for Character Labels
Knowledge Graph (CGKG)
KG Encoder
KG glyph embeddings (glyph_num× k_dim)
KG radical embeddings (radical_num × k_dim)
Decoder
Radical Description Text
“艸,屮,屮,屯,日”
Multimodal Encoder
Multi glyph embeddings (glyph_num × m_dim)

Figure 4: The MCGI framework for glyph identification. It has three encoders to generate visual, textual and multimodal glyph embeddings and one decoder to get the scores for character labels from these embeddings.

MCGI consists of three encoder modules: the image encoder (IE) encodes images in P_{g_k} into high-dimension embeddings and generates the visual glyph embedding $\boldsymbol{e^{v}_{g_k}}$; the graph encoder (KGE) represents glyph and radical entities in KG as the KG glyph embedding $\boldsymbol{e^{kg}_{g_k}}$ and the radical embeddings $\boldsymbol{e_r}$ for all radical entities; the multimodal encoder (ME) generates the multimodal glyph embedding $\boldsymbol{e^{m}_{g_k}}$, which fuses the initialized token embeddings of T_{g_k} provided by BERT with the image and radical embeddings from IE and KGE.

Finally, the $Decoder$ module outputs the result based on the glyph representations from three encoders: $Decoder(\boldsymbol{e^{v}_{g_k}}, \boldsymbol{e^{kg}_{g_k}}, \boldsymbol{e^{m}_{g_k}}) = \{s(g_k, c_i)|i = 1, 2, ..., |C|\}$, $s(g, c)$ is the score between the glyph and the character label.

In this section, we will introduce the IE in Section 5.2, the KGE in Section 5.3, the ME in Section 5.4, the $Decoder$ in Section 5.5, and the working steps of MCGI in Section 5.6.

5.2 Image Encoder

For a target glyph g_k, IE generates the embedding set $E(P_{g_k})$ of its images: $E(P_{g_k}) = \{\boldsymbol{e^{v_i}_{g_k}}|\boldsymbol{e^{v_i}_{g_k}} \in \mathbb{R}^{d_v}, i = 1, 2, ..., |P_{g_k}|\}$, d_v is the dimension of the visual feature space. And it's visual glyph representation $\boldsymbol{e^{v}_{g_k}}$ is set as the average of $E(P_{g_k})$:

$$\boldsymbol{e^{v}_{g_k}} = \frac{1}{|P_{g_k}|} \sum_{p_i \in P_{g_k}} IE(p_i) \quad (1)$$

IE is a computer vision model, such as ResNet. It is pre-trained on the image classification task, which classifies image samples into the character label: $f_{p->c}$ using the images of the glyphs in training dataset.

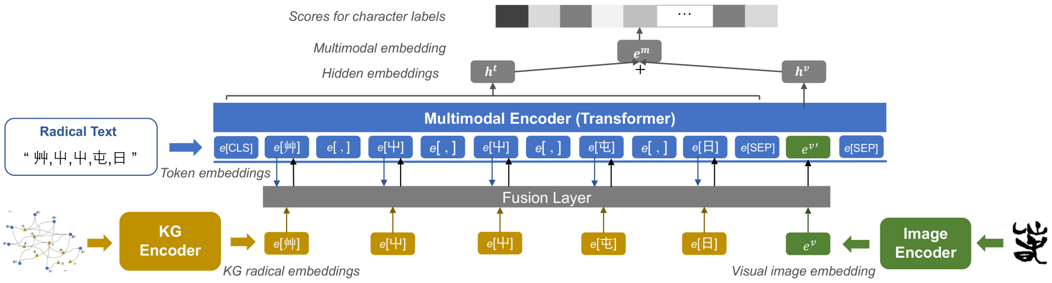

Figure 5: Architecture of the Multimodal Encoder. It takes three kinds of input: the token embeddings initialized from BERT, the image embedding acquired from IE, and the radical embeddings acquired from KGE. These embeddings are fused in the fusion layer and fed into BERT to output the multimodal representation.

5.3 Knowledge Graph Encoder

KGE embeds the glyph and radical entities in KG into $E(KG_g) = \{\boldsymbol{e^{kg}_{g_i}} | \boldsymbol{e^{kg}_{g_i}} \in \mathbb{R}^{d_{kg}}, i = 1, 2, ..., |V_g|\}$ for glyphs, and $E(KG_r) = \{\boldsymbol{e_{r_i}} | \boldsymbol{e_{r_i}} \in \mathbb{R}^{d_{kg}}, i = 1, 2, ..., |V_r|\}$ for radicals, where d_{kg} is the dimension of the graph feature space, V_g and V_r are the glyph and radical node set, respectively, which are the subsets of V.

The KGE is a graph representation model. Here we use node2vec (Grover and Leskovec, 2016) to initialize the entity embeddings and further train a GAT network on the glyph entity classification task $f_{g->c}$ based on G_{train}.

5.4 Multimodal Encoder

The architecture of ME is shown in Figure 5, which is based on BERT. For each glyph g_k, ME inputs each of its image embedding $\boldsymbol{e^{v_i}_{g_k}}$, $\boldsymbol{e^{v_i}_{g_k}} \in E(P_{g_k})$; the radical token embeddings $E(T_{g_k})$ initialized from BERT; and the entity embeddings for radicals of g_k: $E(KG_{r_{(g_k)}})$. It outputs the fused multimodal representation $\boldsymbol{e^{m_i}_{g_k}}$:

$$\boldsymbol{e^{m_i}_{g_k}} = ME(\boldsymbol{e^{v_i}_{g_k}}, E(T_{g_k}), E(KG_{r_{(g_k)}})) \tag{2}$$

We get the multimodal glyph representation $\boldsymbol{e^m_{g_k}}$ by averaging all embeddings in $\{\boldsymbol{e^{m_i}_{g_k}} | \boldsymbol{e^{m_i}_{g_k}} \in \mathbb{R}^{d_m}, i = 1, 2, ..., |P_{g_k}|\}$ gotten from ME, d_m is the dimension of the multimodal feature space.

Specifically, for T, we arrange the radicals in order from top to bottom, from left to right and from outside to inside to the maximum extent, and “,” is the spacing symbol. The radical of the single glyph is set as itself. The T should contain all levels of radicals, for instance, T of the glyph g_{14} (Figure 1) is {艸, 屮, 屮, 屯, 日}, in which “屮, 屮” are the second level radicals making up “艸”.

The specific symbols for BERT are added to T: $\{[CLS], t_2, ..., t_{n-3}, [SEP], [IMG], [SEP]\}$, where n is $|T|$, $[CLS]$, and $[SEP]$ are the start and end symbols, $[IMG]$ is the position where will be filled with the image embedding $\boldsymbol{e^v}$, and t_2 to t_{n-3} are the tokens. The initialized embedding sequence of T is: $E(T) = \{\boldsymbol{e^{t_i}} | \boldsymbol{e^{t_i}} \in \mathbb{R}^{d_t}, i = 1, 2, ..., n\}$, d_t is the dimension of the textual feature space.

The fusion layer fuses $E(T)$ with $\boldsymbol{e^v}$ and radical embeddings from $E(KG_r)$, and generates an unified input embedding sequence $E(O)$, $E(O) = \{\boldsymbol{e^{o_i}} | \boldsymbol{e^{o_i}} \in \mathbb{R}^{d_t}, i = 1, 2, ..., n\}$:

$$\begin{cases} \boldsymbol{e^{o_i}} = Linear(\boldsymbol{e^v}), & i = n-1 \\ \boldsymbol{e^{o_i}} = Linear(\boldsymbol{e^{t_i}} || \boldsymbol{e_{r_{(t_i)}}}), & r_{(t_i)} \in V'_r \\ \boldsymbol{e^{o_i}} = \boldsymbol{e^{t_i}}, & Otherwise \end{cases}$$

where $Linear$ is the linear layer and $||$ is concatenation operation, $r_{(t_i)}$ is the radical entity corresponding to the token t_i and $\boldsymbol{e_{r_{(t_i)}}}$ is the radical embedding of it. V'_r is the set of radical entities that have Rel_2 relation in KG, $V'_r \subset V_r$.

$E(O)$ is fed to ME and outputs the hidden visual representation $\boldsymbol{h^v}$ of o_{n-1}, and the hidden textual representation $\boldsymbol{h^t}$, which is the average of the outputs of o_1 to o_{n-2}, at the last layer. The multimodal embedding $\boldsymbol{e^m}$ is:

$$\boldsymbol{e^m} = \boldsymbol{h^v} + \boldsymbol{h^t} \tag{3}$$

We pre-train ME on the classification task based on the training data. The $\boldsymbol{e^m}$ will be mapped to the scores for character labels through a linear layer and softmax function.

5.5 Decoder

We take a unsupervised $Decoder$. Every character label c_i corresponds to a set of glyphs $G_{(c_i)} = \{g^{c_i}_j | g^{c_i}_j \in G_{train}, j = 1, 2, ..., m\}$, m is the number of glyph of c_i in G_{train}. We calculate the Cosine Similarity between each glyph embedding pair in G_{test} and G_{train} and set the score of (g_k, c_i) as the max score between g_k and the glyphs in $G_{(c_i)}$.

There are three similarity scores between g_k and c_i: $s^m(g_k, c_i)$, $s^v(g_k, c_i)$ and $s^{kg}(g_k, c_i)$, corresponding to the three input glyph embeddings of $\boldsymbol{e^m_{g_k}}$, $\boldsymbol{e^v_{g_k}}$, and $\boldsymbol{e^{kg}_{g_k}}$. The final score $s(g_k, c_i)$ is:

$$s = \alpha s^m + \beta s^v + \gamma s^{kg} \tag{4}$$

where α, β and γ are hyperparameters.

For comparison, we also realized a supervised $Decoder$ which is trained on glyph classification: $f_{g->c}$, in which the hidden glyph representation is $\boldsymbol{e_{g_k}} = \alpha Linear(\boldsymbol{e^m_{g_k}}) + \beta Linear(\boldsymbol{e^v_{g_k}}) + \gamma Linear(\boldsymbol{e^{kg}_{g_k}})$, where α, β and γ are parameters gotten from training.

5.6 Working Steps and Loss Function

MCGI follows a staged training and predicting framework: in the pre-training stage, it separately trains encoders of IE, KGE and ME, and acquires the visual, graph and multimodal glyph embeddings for glyphs in G_{train}, as well as the radical entity embeddings.

In the predicting stage, given a glyph g_{test}: (1) it acquires image embeddings and visual glyph embedding through IE; (2) adds g_{test} into the KG, including creating a new glyph node and linking it to the corresponding radical nodes, and trains KGE to get the KG glyph embeddings for g_{test} and glyphs in G_{train}; (3) acquires the multimodal embedding through ME; (4) outputs the character label through $Decoder$.

We use the cross-entropy loss on the training of three encoders. Given a (x, c_i) pair, x is the target glyph or image, $s(x, c_i)$ is the score between x and the character c_i, the loss function is:

$$loss = -s(x, c_i) + log \sum_{l=1}^{|C|} exp(s(x, c_l)) \tag{5}$$

6 Evaluation

6.1 Dataset Information

The statistics is shown in Table 1. It has 6941 character labels and 1974 radicals. There are 14931 glyphs and 53452 images for training, and 1279 ancient glyphs and 5319 images for testing. The glyph samples are distributed in different time stages (Oracle, Bronze, States and Modern).

There is an average of only 2.15 (14931/6941) glyph samples per character label for training. It is because of the incompleteness of the current dataset in this field, and the real-world discovered variant glyphs themselves are limited in quantity. Thus, in the current research, the testing glyphs only overs 805 commonly used character labels, and each testing glyph has at least one variant glyph in the training set.

Dataset	Data	Statistics
-	Character (C)	6941
	Radical (R)	1974

Train	Glyph (G_{train})	14931
		(Oracle: 2478;
		Bronze: 2839;
		States: 6042;
		Modern: 3572)

	Image (P_{train})	53452

Test	Glyph (G_{test})	1279
		(Oracle: 414;
		Bronze: 468;
		States: 397)

	Image (P_{test})	5319

Table 1: The statistic information of dataset.

	Consistent	Inconsistent	Sum
S	145	453	598
D	110	571	681
Sum	255	1024	1279

Table 2: Statistics of the categories of testing glyphs (“S” represents the synchronic testing glyphs; “D” is the diachronic testing glyphs).

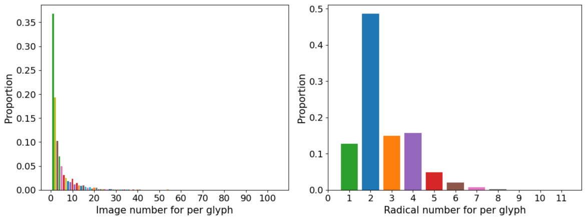

Figure 6: The proportion distributions of training glyph samples for per number of images and radicals.

And the proportion distributions of image and radical quantities for training glyph samples are shown in Figure 6. The average number of image for per glyph is 3.73 and radical is 2.62. The image dataset has the long tail effect due to the differences in usage frequency. It conforms to the real world distribution for ancient Chinese character.

The categories of testing glyphs is shown in Table 2, among the 1279 testing glyphs, there are

Method	Synchronic testing group					Diachronic testing group				
	R@1	**R@3**	**R@10**	**MR**	**MRR**	**R@1**	**R@3**	**R@10**	**MR**	**MRR**
(v) DenseNet121	29.9%	39.6%	50.5%	410.9	0.369	23.8%	35.1%	47.9%	591.0	0.316
(v) ResNet50(ours)	35.6%	48.7%	61.4%	291.6	0.441	28.8%	40.2%	53.2%	368.9	0.369
(k) GAT_CGKG	58.7%	77.8%	85.5%	33.9	0.691	36.9%	51.5%	68.0%	153.0	0.469
(t) LCS	50.5%	66.1%	76.9%	58.4	0.599	28.6%	43.3%	52.9%	479.4	0.378
(t) SikuBERT	67.4%	79.4%	85.5%	85.7	0.742	35.7%	47.6%	57.7%	482.0	0.435
(t) BERT(ours)	70.6%	82.8%	88.0%	82.0	0.777	40.4%	53.5%	62.4%	453.5	0.485
(t+v) ViLT (Kim et al., 2021)	66.7%	81.1%	87.5%	69.5	0.743	51.7%	63.3%	70.8%	348.3	0.587
(t+v) HRGAT (Wang et al., 2022)	65.2%	74.4%	81.6%	185.3	0.710	36.1%	46.0%	53.3%	721.7	0.424
(t+v+k) IMF (Li et al., 2023)	63.0%	76.1%	85.6%	68.4	0.710	38.6%	54.8%	68.4%	156.7	0.448
(t+v+k) GlyphSim (Chi et al., 2022)	69.6%	85.8%	91.6%	**9.3**	0.784	44.1%	62.0%	79.1%	**62.2**	0.557
MCGI (ME(t+v))	68.6%	80.3%	85.5%	84.1	0.755	55.9%	64.3%	72.1%	260.1	0.618
MCGI (ME(t+k))	71.9%	84.1%	89.0%	69.8	0.782	41.1%	53.7%	61.7%	362.7	0.487
MCGI (ME(t+v+k))	69.2%	79.3%	85.8%	93.6	0.754	**56.5%**	65.9%	73.3%	270.7	0.625
MCGI (ME(t+v+k)+IE)	69.4%	83.6%	89.0%	32.1	0.772	50.4%	65.8%	75.3%	130.7	0.594
MCGI (ME(t+v+k)+KGE)	70.7%	85.3%	90.5%	19.4	0.785	52.7%	69.3%	79.1%	96.8	0.625
MCGI (ME(t)+KGE+IE)	72.1%	87.1%	92.8%	13.0	0.801	51.7%	68.1%	80.9%	72.9	0.619
MCGI (ME(t+v)+KGE+IE)	72.6%	88.3%	93.0%	11.9	0.808	54.8%	71.1%	**82.5%**	68.6	0.648
MCGI ($\boldsymbol{ME}$(t+v+k)+$\boldsymbol{KGE}$+$\boldsymbol{IE}$)	73.5%	**88.5%**	**93.0%**	13.1	**0.812**	**56.5%**	**71.8%**	82.1%	70.0	**0.659**
MCGI (ME(t+v+k)+KGE+IE-(s))	**73.7%**	86.5%	92.0%	29.9	0.810	55.7%	70.6%	80.8%	90.9	0.646

Table 3: The results of the quantitative comparisons with uni-modal and multimodal baseline methods on the synchronic and diachronic testing groups, respectively. **MCGI ($\boldsymbol{ME}$(t+v+k)+$\boldsymbol{KGE}$+$\boldsymbol{IE}$)** is our full method. The best results are highlighted in **red** color, and the best results in four sub groups are highlighted in blue color.

598 synchronic testing glyphs (S, Section 3.2) and 681 diachronic ones (D). In addition, we count the consistency of radical types (Section 3.2) between testing glyphs and their variant glyphs that are in training set. The testing glyph in "Consistent" group has at least one training variant glyph with the same radical types as it (they are only different in radical location, quantity, glyph or time stage), while the testing glyph in "Inconsistent" group changed the radical types. There are 1024 inconsistent testing glyphs and only 255 consistent ones, which makes a certain difficulty for this task.

6.2 Indicators and Baseline Models

Our evaluation indicators include: **R@1**, **R@3** and **R@10**, which are the average proportion of top-n correct label rankings; the **Mean Rank (MR)** and the **Mean Reciprocal Ranking (MRR)**. The potential application scenario of this work is to recommend experts with a set of most possible (top-n) character labels to inspire their thoughts and reduce the candidate character scope. So the ranking score is most suitable to evaluate this task and the MR score can show the human workload in the real applications.

We classified the models according to the modality of data they used. The uni-modal models includes ((v) is visual modality; (t) is textual modality; (k) is graph modality):

- (v): (1) DesNet121; (2) ResNet50 (ours): we pre-trained a ResNet model on the image dataset of ancient Chinese characters based on unsupervised contrastive learning.
- (k): GAT_CGKG, the GAT network on CGKG.
- (t): (1) LCS (longest common subsequence) of the radical text; (2) SikuBERT[1], which was trained on ancient Chinese corpus; (3) BERT (ours): we trained it from SikuBERT on the unearthed ancient Chinese contexts.

All the uni-modal models were trained on the glyph or image classification tasks. For the (v) models, the score between (g, c) is set to the sum of the (p, c) scores of all images of the glyph g.

There are four multimodal baseline methods (Section 2): (t+v) ViLT (Kim et al., 2021); (t+v) HRGAT (Wang et al., 2022); (t+v+k) IMF (Li et al., 2023); and (t+v+k) GlyphSim (Chi et al., 2022). Apart from GlyphSim, their tasks are very different from ours. We only realized their multimodal

[1] https://huggingface.co/SIKU-BERT/sikubert

fusion parts and trained their models on glyph classification task, to show the effects of the potential available methods on our dataset and task.

For our MCGI, the IE, KGE and ME encoders used the uni-modal models of ResNet50 (ours), GAT_CGKG and BERT(ours) respectively. To do the ablation study, we firstly individually evaluated the three kinds of ME encoder that use different combinations of T, KG_r and P information: ME(t+v+k); ME(t+v); and ME(t+k), in which ME(t+v+k) is our full method. And then, we evaluated our full MCGI method: ME(t+v+k)+KGE+IE in $Decoder$ and several combinations of three encoders: ME(t+v+k)+IE; ME(t+v+k)+KGE; ME(t)+KGE+IE; ME(t+v)+KGE+IE. We also evaluated the supervised $Decoder$ for comparison: ME(t+v+k)+KGE+IE-(s).

6.3 Configuration

In pre-training of IE, the batch size was 64, the learning rate was 0.001, epoch was 150 and d_v was 768; For ME pre-training, we used the basic BERT with 768 dimensions, batch size was 64, the learning rate was 0.00002, d_m was 768, and we trained it for 100 epochs. For KGE, the embeddings were initialized by using the opened OpenNE[2] tool, we used their default parameters and the dimension of the output vector was 1000. And we trained GAT network through 5 epochs, the learning rate was 0.0005 and batch size was 256, the d_k was 2000. The hyperparameters α, β and γ in $Decoder$ were 0.4,0,5, and 0,7 respectively, which were set by testing 50 extra validate glyph samples on the combinations between 0.1 and 1.0. For all supervised models, we ran for 3 times and chose the best score as the result. The training time of our ME and IE is about 8 hours on the NVIDIA RTX 3090 GPU.

6.4 Results and Discussions

Validity of our method: The results are shown in Table 3, our model achieves the best results in most indicators except MR compared with the baselines.

The results of ME encoder (ME(t+v+k)), on the D group, are improved from the uni-modal and also better than the multimodal data fusion baselines (Kim et al., 2021; Li et al., 2023; Wang et al., 2022). However, we observed slight declines for S group after fusing the visual features (comparing BERT(ours) and ME(t+v)).

The $Decoder$ module further enhances the overall performance. The ablation study proves the validity of all encoders: (1) it shows the contribution of our KGE encoder and the graph CGKG (comparing ME(t+v+k)+IE and ME(t+v+k)+KGE+IE; ME(t+v+k) and ME(t+v+k)+KGE). It improves the top-n rankings especially for S group, and reduces the MR indicator observably; (2) by comparing directly using BERT (ME(t)+KGE+IE) and using ME encoder (ME(t+v+k)+KGE+IE)), it proves the effectiveness of our ME encoder, especially for improving the results for D group; (3) we did not observe the improvement of supervised $Decoder$ (ME(t+v+k)+KGE+IE-(s)) in most indicators. It may because that one character label only has limited glyph samples for training in our specific task; (4) the unsupervised method GlyphSim (Chi et al., 2022) has better MR scores compared with ours, however, we have much better top-n rankings through pre-training of encoders.

Effectiveness for each modality: There are overlarge pictorial differences between ancient variant glyphs, which limited the performance of (v) methods. The results of (t) and (k) models are better than the visual models, the (t) models have better scores in the top-n ranks and (k) model suits to reduce the mean rankings. However, we also observed limitations of them, most glyphs contain less than 3 radicals, which is insufficient to distinguish features of similar characters. We randomly selected 150 error samples for ME(t+k) model, and found that 108 of them were classified into character labels with very similar radical combinations. The multimodal fusion can leverage the advantages of each modality, in which the radical semantics has a better effect on the S testing group (ME(t+k)), while the D group more requires visual features (ME(t+v)). The reason we analyzed is that there is a greater radical semantic correlation between S variant glyphs, making them benefit more from our glyph knowledge graph.

7 Limitations

The limitations of this work includes:

(1) We only tested 1279 deciphered glyphs in 805 commonly used character labels. The uncommon or undeciphered glyphs should be more difficult to be reasoned, thus the effectiveness would be reduced. And the datasets in this field are imperfect currently for those uncommon characters.

[2] https://github.com/hengdos/OpenNE

(2) Our method only suits for the glyph who has variant glyphs in training set. And the effects of the method were observed slightly worse for those glyphs with single radical, as well as the glyphs whose radicals are both newfound and not shared with training glyphs, but this situation is rare in this task.

(3) Except the radical and visual features, the contexts of historical documents, dictionary records, pronunciations and so on are very important, which should be considered in the future.

Currently, we are associating Chinese glyphs with more textual resources such as the historical documents, dictionary records and so on. The ultimate goal is to provide a complete methodology and refined model for ancient Chinese character deciphering. And in application level, we will develop the glyph recognization and recommendation serves based on this work.

8 Conclusion

This paper introduces an innovative research of glyph identification of ancient Chinese characters and devises a knowledge powered multimodal method based on visual and radical semantic information for this task. The results proved the effectiveness of our glyph knowledge graph and the multimodal method, which achieves the best results in most indicators, and shows the potentials of visual and radical information in this task for the first time. This work can be applied in the related fields of ancient Chinese linguistics and characters.

Acknowledgements

This article is a phased achievement of the National Social Science Foundation project “Research on Chinese Pre Qin Language and Culture Based on Jin-wen Data”(number:23VRC033), and has been supported by the Youth Project of the Philosophy and Social Science Research Innovation Team of Jilin University “Inheritance of Chinese Character Culture and New Forms of Human Civilization”(number:2023QNTD02).

References

William G. Boltz. 1986. Early chinese writing. *World Archaeology*, 17(3):420–436.

Zhong Cao, Jiang Lu, Sen Cui, and Changshui Zhang. 2020. Zero-shot handwritten chinese character recognition with hierarchical decomposition embedding. *Pattern Recognition*, 107:107488.

Xiang Chang, Fei Chao, Changjing Shang, and Qiang Shen. 2022. Sundial-gan: A cascade generative adversarial networks framework for deciphering oracle bone inscriptions. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 1195–1203.

Yang Chi, Fausto Giunchiglia, Xiaolei Diao Daqian Shi, Chuntao Li, and Hao Xu. 2022. Zinet: Linking chinese characters spanning three thousand years. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022*, page 3061–3070.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 4171–4186.

Xiaolei Diao, Daqian Shi, Hao Tang, Qiang Shen, Yanzeng Li, Lei Wu, and Hao Xu. 2023. Rzcr: Zero-shot character recognition via radical-based reasoning. In *In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 654–662.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1063–6919.

Guanjie Huang, Xiangyu Luo, Shaowei Wang, Tianlong Gu, and Kaile Su. 2022. Hippocampus-heuristic character recognition network for zero-shot learning in chinese character recognition. *Pattern Recognition*, 130:108818.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, PMLR*, page 5583–5594.

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. Imf: Interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*, page 2572–2580.

Jane Qiu. 2014. Ancient times table hidden in chinese bamboo strips. *Nature*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, PMLR*, pages 8748–8763.

Edward L. Shaughnessy. 1991. *Sources of Western Zhou History: Inscribed Bronze Vessels*. University of California Press, Berkeley, Los Angeles, Oxford.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR*.

Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain*, page 938–948.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI-23)*, pages 10637–10647.

Gechuan Zhang, Dairui Liu, Barry Smyth, and Ruihai Dong. 2021. Deciphering ancient chinese oracle bone inscriptions using case-based reasoning. In *Proceedings of the 29th International Conference on Case-Based Reasoning, ICCBR 2021*, page 309–324.

Jianshu Zhang, Yixing Zhu, Jun Du, and Lirong Dai. 2018. Radical analysis network for zero-shot learning in printed chinese character recognition. In *In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME)*.

Ying Zhang, Wenbo Fan, Kehui Song, Yu Zhao, Xuhui Sui, and Xiaojie Yuan. 2023. Incorporating object-level visual context for multimodal fine-grained entity typing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 15380–15390.