Eigencharacter: An Embedding of Chinese Character Orthography

Yu-Hsiang Tseng
Graduate Institute of Linguistics
National Taiwan Unviersity
seantyh@ntu.edu.tw

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan Unviersity
shukaihsieh@ntu.edu.tw

Abstract

Chinese characters are unique in its logographic nature, which inherently encodes world knowledge through thousands of years evolution. This paper proposes an embedding approach, namely eigencharacter (EC) space, which helps NLP application easily access the knowledge encoded in Chinese orthography. These EC representations are automatically extracted, encode both structural and radical information, and easily integrate with other computational models. We built EC representations of 5,000 Chinese characters, investigated orthography knowledge encoded in ECs, and demonstrated how these ECs identified visually similar characters with both structural and radical information.

1 Introduction

Chinese is unique in its logographic writing system. The Chinese scripts consists of a sequence of characters, each carried rich linguistic information on its own. Chinese characters are not only mediums for pronunciations and lexical meanings, they also carry abundant information in the visual patterns.

Chinese orthography has been closely investigated in literature, from structural analysis of *ShuoWenJieZi* in Han dynasty, to contemporary sociolinguistic perspective (Tsou, 1981). Recent behavioral studies even argued that, given the salience of Chinese writing system, the orthographic components are activated first when reading, followed by phonological and semantic activation (Perfetti et al., 2005). However, emphases of previous orthographic approaches were more on radicals, components, or their respective positions in the whole characters and how Chinese readers recognize characters in a processing theory. This paper presents a computational approach, eigencharacter representations, to describe Chinese characters in a vector space. The resulting representations encode lexical knowledge embedded in Chinese characters, therefore provide unique insights on the integration between computational models and linguistics.

2 Previous Works

Chinese characters are visual patterns occupied in a square space. Depending on the strokes of a character, the visual pattern may be simple, such as a single stroke character (一, yī, "one") or complex, such as a 16 stroke character (龜, guī, "turtle"). Psychophysics studies showed that Chinese characters carry more information in high spatial frequency, compared with alphabetic language (Wang and Legge, 2018). Although some Chinese characters are *unique* characters, which there is no further components can be distinguished in the whole character, identifying radicals and components in a character is the most common way to analyze Chinese orthography.

2.1 Components decomposition

In Chinese classic text, *ShuoWenJieZi* identified 540 radicals in Chinese characters, from which 214 of them are derived and used in modern Chinese. The radical often carries a semantic meaning of a character, and rest of the characters form a component which may provide hints of character pronunciation. For example, 燃, rán, "burning" has radical 火, huǒ, "fire", in the left side, which has apparent semantic connection between the whole character. The right side of the character, 然, rán, "then" provides a phonological cue, which is the same as the whole character in this exam-

Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), pages 24–28
Hong Kong, China, November 3, 2019.

ple. The decomposition strategy are especially useful in pedagogical context and behavioral experiment, since it separates meanings from pronunciations, so they can be taught or manipulated separately.

Not all Chinese characters are applicable to this decomposition strategy. Some characters has *unique* structure, which it cannot be easily separated as radical and components. For instance, 東, dōng, "east" has radical of 木, mù, "wood", but the rest of the character, 曰, yuē, "say", is tightly embedded in the character and has no phonological relations with the whole character. There are even some characters cannot be decomposed at all, such as 我, wǒ, "I, me", cannot be decomposed into a component after removing the radical (戈, gē, "weaponery").

Some studies tried to decompose characters into finer components, through which characters can be divided recursively into a component hierarchy (Chuang and Hsieh, 2005). For example, instead of decomposing 燃 into its semantic radical (火) and phonological component (然), the phonological component can be further divided into three components: 夕, xì, "dusk", 犬, quǎn, "dog", 灬, huǒ, "fire". This approach provides a complete description of characters, but is not without caveats. Specifically, it is not easy to find visually similar characters with component hierarchy, (e.g. 已 and 己 are visually similar, while not sharing common components), and the definition of a components is not always clear (e.g. 龍, lóng, dragon could have one, two or three components, depending on different definitions).

2.2 Eigendecomposition of Visual Stimuli

Although decomposing characters into components are advantageous in pedagogical context and in behavioral experiments, the discrete nature of components prevents a simple coding scheme of Chinese orthography. Specifically, there are 214 radicals in modern Chinese, which would require hundreds of dimension in a vector to encode radicals and other components. In addition, were positions of each radicals/components considered, the dimensions needed to encoded a single character would increase exponentially.

An alternative approach to construct a computational representation of Chinese character is leveraging the fact scripts are written in square blocks, each character can be considered as an information-laden visual patterns. The computation task is to extract common components among these patterns (characters), and choose fewest possible number of components to best represent given set of characters. The idea is closely related to eigenface decomposition in face recognition and face processing studies (Sirovich and Kirby, 1987).

Chinese characters and faces are two distant but striking similar concepts, both in computational tasks and in cognitive neuroscience. Face and characters were shown to share similar processing mechanisms and even found to have closely related neural mechanisms (Farah et al., 1995; Zhang et al., 2018). In addition, face and (handwritten) character recognition were both attempted in a low dimensional space (Sirovich and Kirby, 1987; Long et al., 2011). The low dimensional face space (eigenface) was later applied into cognitive science, through which a face space was constructed and was used to explain phenomena concerning face recognition (O'toole et al., 1994).

In this paper, inspired by concepts of eigenface, we tried to construct a eigencharacter space to represent Chinese characters, and investigate the orthographic information implicated in eigencharacters.

Constructing eigencharacters provides unique advantages in computational modeling. These representations are invaluable that they are (1) clearly and automatically defined given a set of characters; (2) helpful when finding similar characters even when not sharing common components; (3) insightful when considering Chinese orthography on their structure and essential components; (4) easily manipulable and conveniently incorporated, since they are inherently a vector, into recent computational models (e.g. neural network models).

3 Eigencharacter

We constructed eigencharacter space with 5,000 most frequently used characters, which was the estimated vocabulary size of average college students in Taiwan (Hue, 2003). Mean strokes of these characters was 12.24, standard

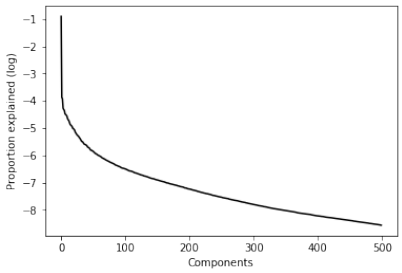

Figure 1: Variance explained under different number of eigencharacter components.

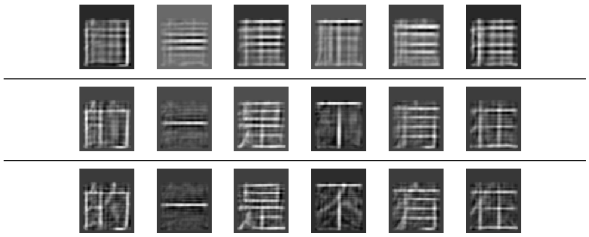

Figure 2: Character reconstructed with different number of eigencharacters. Reconstruction with 10 ECs (upper panel), with 50 ECs (middle panel), and 100 ECs (lower panel).

deviation was 4.48. Character with the fewest stroke (1 stroke) was 一, yī, "one"; the one with most strokes was 籲, yù, "call, implore".

Each character was first drawn with white ink on a binary bitmap of black background. The font used was Microsoft JhenHei, font size was 64. After drawing characters on bitmaps, they were reshaped into column vectors of length 4800 (i.e. 64×75). The resulting character matrix therefore has dimension 4800×5000 matrix.

The character matrix was then decomposed with singular value decomposition:

$$M = U\Sigma V^T$$

where M is the original character matrix, Σ is a diagonal matrix with singular values. To determine the number of singular vectors, or number of eigencharacters(ECs) needed to best represent the character matrix, we first examined the scree plot of singular values normalized by the Frobenius norm of M (Figure 1).

From Figure 1, proportion of variance explained quickly dropped after 50 ECs. To verified the observation, we attempted to reconstruct the character with first 10, 50 and 100 ECs (i.e. the first 10, 50, 100 columns of U). The resulting construction is shown in Figure 2. The reconstruction of first 10 ECs only recovered limited patterns of each character. Interestingly, the patterns recovered were mostly vertical or horizontal stripes. When using 50 ECs, the resulting patterns started to be recognizable, and they were identifiable when using 100 ECs. Basing on the results above, we chose first 50 ECs to construct eigencharacters space.

4 Experiments

Constructed ECs space serves multiple purposes. Among their potential advantages on incorporating orthographic knowledge naturally inherited in Chinese writing system, we demonstrate how ECs reveal structure and component information in Chinese orthography, and how they are particularly effective in finding visually similar characters.

4.1 Rendering Eigencharacters

ECs are abstract mathematical construct extracted from singular value decomposition, which might not be directly interpretable. However, these ECs are essentially the bases best represent 5,000 characters, the actual patterns of these ECs could bear interesting insight on Chinese orthography.

We rendered 50 ECs extracted in previous section, and reconstructed them as if they were normal character. The renderings were shown in Figure 3, ECs are ordered descendingly by their respective singular values.

The rendering showed interesting patterns. By visual inspection, we can observed that (1) first few ECs encode "low spatial frequnecy" information, such as the general character block in EC0, vertical stripes in EC1, EC2, and horizontal stripes in EC4, EC5; (2) they do not correspond directly to radicals, but some important radicals can be identified nevertheless, such as 氵 radical, "water" in EC14, 言 radical, "words" in EC15, and 女 radical, "female" in EC31.

In addition to visual inspection, we could also understand ECs by the characters having the highest or the lowest coefficients in each ECs. By these positively or negatively *loaded*

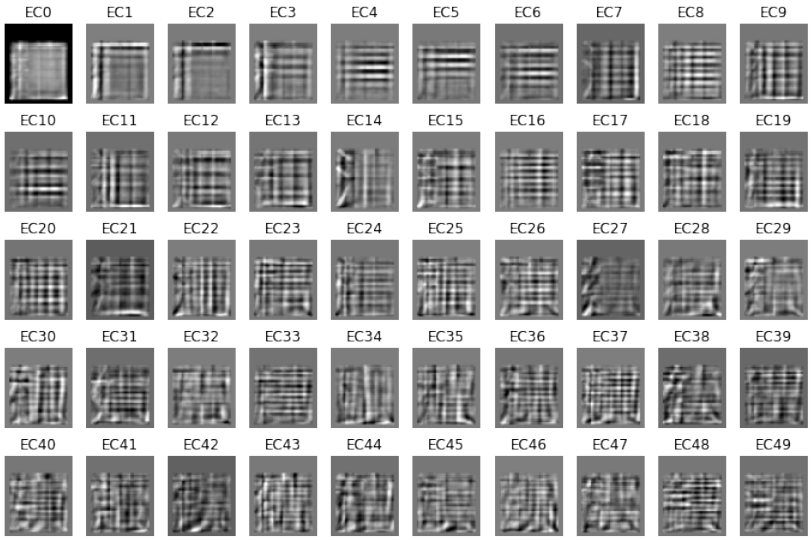

Figure 3: First 50 Eigencharacters.

characters, we could infer the information each EC encodes in character space. For example, the 3 highest loaded characters on EC0 are 轟, 竇 and 鷹, and the 3 lowest loaded characters are 一, 卜, and 二, aided by the EC rendering, we could infer EC0 is a component of "stroke complexity". Likewise, the 3 highest loaded characters of EC1 are 圄, 鬩, and 閘, the 3 lowest loaded are 值, 椿, and 捧. EC1 is then infered to be a component of "enclosing structure". These components of structure echoed the behavioral studies that showed Chinese readers use structural information to judge character similarity (Yeh and Li, 2002). Aside from components of structural representation, there are also components of radicals. For instance, EC31, which shows a 女 radical in rendering, is a component of 女 radical. It has highest loading in characters with 女 radical, such as 媒, 娩, and 妮.

Renderings of ECs and inspection of their loaded characters, suggest ECs are not only abstract mathematical constructs. Instead, they automatically encode and reflect structural and radical aspects in Chinese orthography.

4.2 Finding Similar Characters

Eigencharacters encodes structural and radical information in characters, which would be ideal to find visual similar characters that is otherwise impossible using components decomposition approach.

Table 1 show examples of similar characters identified with eigencharacters. Character similarity is defined as the euclidean distance between two characters in EC space. In first row of table 1, EC space found similar characters with identical radical (犭), components(胡), and remarkably considering the three parts vertical structure simultaneously. In the second row of the table, the similar characters of 語, highlighted another property of EC space: it did not restrict itself on the exact components, but the visually similar components, such as 諮, 晤 and 誤, they either share the same radical/component, or having similar right hand side components. The last row also showed the advantages of EC space in finding visually similar character. For instance, 東 and 泉 are both unique structure, and they share similar patterns (曰 in the middle, and two oblique strokes in lower half) which would be challenged to accommodate were components-based decomposition were used.

These illustrative examples showed EC space, which inherently equipped with knowledge of structural and radical information, provides an ideal representation to explore Chinese orthography.

5 Conclusion

This paper introduces eigencharacters, an embedding representation of Chinese orthography. It provides unique advantages over

Seed	Similar characters
湖	溯潮漏瑚澗
語	諮晤誤診譜
頭	頹頷頸頌頻
龍	鶺廳籠麓蹝
東	泉帛柬車蒐

Table 1: Similar characters found with eigencharacter space.

component-based character decomposition, in that it can be automatically extracted, encodes both structural and radical information, and easily integrates with other computational models. Equipped with EC representations, human knowledge encoded in Chinese orthography becomes easily accessible to downstream NLP applications.

Acknowledgments

This work was supported by Ministry of Science and Technology (MOST), Taiawan. Grant Number MOST 108-2634-F-001-006.

References

D. M. Chuang and C. C. Hsieh. 2005. Database of chinese character orthography: Development and application. In *International conference on Chinese character and globalization*.

Martha J. Farah, Kevin D. Wilson, H. Maxwell Drain, and James R. Tanaka. 1995. The inverted face inversion effect in prosopagnosia: Evidence for mandatory, face-specific perceptual mechanisms. *Vision Research*, 35(14):2089–2093.

C. W. Hue. 2003. Number of characters a college student knows. *Journal of Chinese Linguistics*, 31(2):300–339.

H. Long, X. C. Zhang, and K. E. Ercan. 2011. Handwritten chinese character recognition using eigen space decomposition. *Science China*, 53(1):1–10.

Alice J. O'toole, Kenneth A. Deffenbacher, Dominique Valentin, and Herve Abdi. 1994. Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, 22(2):208–224.

Charles A. Perfetti, Ying Liu, and Li Hai Tan. 2005. The lexical constituency model: Some implications of research on chinese for general theories of reading. *Psychological Review*, 112(1):43–59.

L. Sirovich and M. Kirby. 1987. Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am. A*, 4(3):519–524.

B. K. Y. Tsou. 1981. A sociolinguistic analysis of the logographic writing system of chinese. *Journal of Chinese Linguistics*, 9(1):1–19.

Hui Wang and Gordon E. Legge. 2018. Comparing the minimum spatial-frequency content for recognizing Chinese and alphabet charactersWang amp; Legge. *Journal of Vision*, 18(1):1–1.

Su-Ling Yeh and Jing-Ling Li. 2002. Role of structure and component in judgments of visual similarity of chinese characters. *Journal of Experimental Psychology: Human Perception and Performance*, 28(4):933–947.

Bo Zhang, Sheng He, and Xuchu Weng. 2018. Localization and functional characterization of an occipital visual word form sensitive area. *Scientific Reports*, 8(1).