Journal of Research in Reading
UKLA
The United Kingdom Literacy Association

Journal of Research in Reading, ISSN 0141-0423 *DOI:10.1111/j.1467-9817.2013.01558.x*
Volume 36, Issue S1, 2013, pp S35–S50

Using singular value decomposition to investigate degraded Chinese character recognition: evidence from eye movements during reading

Hsueh-Cheng Wang

Department of Computer Science, University of Massachusetts at Boston, USA

Elizabeth R. Schotter, Bernhard Angele and Jinmian Yang

Department of Psychology, University of California, USA

Dan Simovici and Marc Pomplun

Department of Computer Science, University of Massachusetts at Boston, USA

Keith Rayner

Department of Psychology, University of California, USA

Previous research indicates that removing initial strokes from Chinese characters makes them harder to read than removing final or internal ones. In the present study, we examined the contribution of important components to character configuration via singular value decomposition. The results indicated that when the least important segments, which did not seriously alter the configuration (contour) of the character, were deleted, subjects read as fast as when no segments were deleted. When the most important segments, which are located in the left side of a character and written first, were deleted, reading speed was greatly slowed. These results suggest that singular value decomposition, which has no information about stroke writing order, can identify the most important strokes for Chinese character identification. Furthermore, they also suggest that contour may be correlated with stroke writing order.

The visual system recognises visual stimuli such as objects and words through a hierarchical process that includes a part-based stage beginning with independent feature detection (e.g., Biederman, 1987; Hubel & Wiesel, 1962, 1963; McClelland & Rumelhart, 1981; Selfridge, 1959; Taft, Zhu & Peng, 1999). These features are combined into higher-level, more meaningful components that contribute to activation of the overall concept (i.e., the object or word). According to the word recognition model proposed by McClelland and Rumelhart (1981), English words are represented by analysis of visual input at the feature, letter and word levels. Successful word recognition is achieved by activating nameable or functional components of the word, such as letters (see Martelli, Majaj & Pelli, 2005, for a review). However, some early studies suggested that, rather than individual letter analysis, the visual configuration of a word is important for word recognition (Garner, 1981; Healy & Drewnowski, 1983; Healy, 1994). Similarly, Chinese words are processed at multiple levels: feature, radical, character and multi-character word (Taft et al., 1999); but the differences between the orthographic representations of English and Chinese raise the question, what are the functional components of Chinese characters?

In languages with alphabetic orthographies, word recognition is mediated, to some extent, by an analysis of a word's component letters (Balota, Pollatsek & Rayner, 1985; Evett & Humphreys, 1981; Gough, 1972; McClelland & Rumelhart, 1981; McConkie & Zola, 1979; Slattery, Angele & Rayner, 2011; Taft, 1985). Moreover, not all letters are of equal importance to the word recognition process. When a word's letters are replaced by other letters, changes to initial letters are more disruptive than changes to medial or final letters (Rayner & Kaiser, 1975; Rayner, White, Johnson & Liversedge, 2006). Furthermore, exterior letters are more important than word internal letters (Jordan, Thomas, Patching & Scott-Brown, 2003; Rayner et al., 2006). These data suggest that, in alphabetic languages, the position of the letters within a word is very important and some letter positions are more important than others. What about component features of a very different orthography, like Chinese, which does not use letters?

In contrast to alphabetic languages, the orthographic units of a Chinese character are written within a single box, not in a linear order. These units, *strokes*, are simple features (e.g., dots, lines or curves) or combinations of simple features that vary in complexity (Zhang, Wang, Zhang & Zhang, 2002). For example, the strokes of the character 场 in its writing order are ㇐, ㇑, ㇀, ㇋ ㇒and ㇒, where the fourth stroke, ㇋, is composed of five segments. Strokes are written in a defined order that generally follows the order of left to right, top to bottom and exterior to interior. Thus, each stroke's contribution to the character is less clear in Chinese. As in alphabetic orthographies, one or more lines represent a letter, which in turn represents a sound (although alphabetic languages differ in the degree to which letters correspond directly to sounds), in Chinese, each stroke (or combination of a few strokes) does not necessarily correspond to a sound or other unit of representation, but rather the configuration of all the strokes together contributes to the character's overall meaning.

Not surprisingly, without all the strokes of a character intact, subjects have difficulty identifying or remembering the character, depending on the extent and nature of stroke removal. Tseng, Chang and Wang (1965) removed different portions of the strokes of characters ranging from 10% to 60%, using different methods: from the beginning or from the ending of the canonically written stroke order, or strokes that did not greatly change the character's visual configuration. Subjects were required to fill in the strokes that had been removed and performed better when the ending strokes were removed than when beginning strokes were removed and best when the basic configuration of the character was retained.

Yan et al. (2012) found similar results to Tseng et al. (1965) for subjects reading sentences in which the characters had 15%, 30% or 50% of their strokes removed either at the beginning or the end of the writing order, or strokes that did not contribute to the character's configuration. They found that when 30% or 50% of strokes were removed, reading was disrupted. Furthermore, in those conditions, characters in which the strokes that do not contribute to the configuration were removed were easiest to read, characters with ending strokes removed were more difficult to read, and characters with beginning strokes removed were the most difficult to read. Taken together, these data suggest that the strokes of a character that are written first are more important than the strokes that are written last or the ones that do not contribute to character configuration. The results are generally consistent with the findings of alphabetical languages that beginning or exterior letters seem to be more important than ending or interior ones.

However, these studies with Chinese raise the question of whether there is something privileged about the first-written strokes or whether another aspect of the strokes at the beginning of the writing order is what causes them to be more important for character identification. To test this, we turned to *singular value decomposition* (SVD; Strang, 1993) to investigate whether the contribution of these strokes to the configuration of the character drives their importance for identification. Importantly, SVD defines a character in terms of its visual form and is completely agnostic to the prescribed writing order of Chinese. Thus, with this method of stroke deletion, we were able to directly test the contribution of visual form (choosing the segments identified by SVD as the least or most informative, i.e., those that contribute the most or least redundant information to the character's overall visual form, respectively). SVD, similar to principal component analysis, is a dimension-reduction method in linear algebra to retain the least redundant components contained in a matrix (Elden, 2007). The mathematical detail of SVD is briefly described in Appendix A. These dimension-reduction methods have been extensively used to describe stimuli in many pattern recognition and vision problems, including face recognition (Craw & Cameron, 1991; Turk & Pentland, 1991), scene recognition (see Torralba & Oliva, 2003, for a review), and most importantly for our present purposes, handwritten character recognition (Hastie & Simard, 1998). SVD is particularly well suited to decompose and summarise Chinese characters because the orientations of strokes are typically vertical, horizontal, diagonal or combinations thereof. Thus, linear algebra and SVD are easily adapted to Chinese characters, which can easily be defined by oriented lines and could prove to be informative in that it can identify those that are most important to the visual configuration of a character while remaining agnostic about canonical writing order.

In this study, we used four oriented filters to decompose each character into simple segments that roughly map on to features such as oriented lines of vertical, horizontal and diagonal orientations, as presented in Figure 1. Representing characters by their oriented line vectors may be well adapted to a biologically plausible model (such as the feature level by Taft et al., 1999) of character identification for early visual processing. To combine these features into higher-level, more meaningful components that contribute to activation of the overall concept, we used SVD to determine which segments contribute the most information to a Chinese character's visual form. The actual input into SVD is a matrix containing 40 × 40 greyscale pixels of the original character, which provide only low-level visual information (such as features, orientations and spatial structure) but not higher-level linguistic information (such as writing order or the composition of radicals).

The least redundant components determined by SVD were considered the most important ones for recognition in this study. As mentioned previously, strokes may be simple

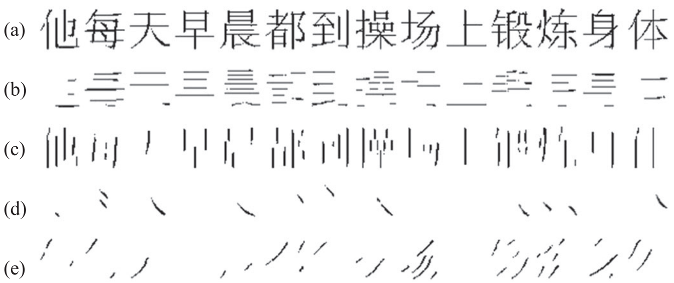


Figure 1. Decomposition of characters. (a) all segments, (b) horizontal segments, (c) vertical segments, (d) diagonal (top-left to bottom-right) segments and (e) diagonal (top-right to bottom-left) segments.

features or combinations thereof that vary in length and complexity. In contrast to previous studies, the SVD method of stroke deletion avoids the problem of stroke complexity by decomposing strokes into segments, thus reducing bias toward longer or more complex strokes seeming more important or disruptive if deleted.

In the present study, we investigated whether the most important strokes can be identified by SVD. Importantly, if SVD identifies that the most disruptive strokes to delete are the same as those identified by Tseng et al. (1965) and Yan et al. (2012), it would suggest that there is something about the characters in those positions that are more informative (less redundant) than other strokes: the visual configuration of the character because visual configuration is the only input to the SVD algorithm. We had three hypotheses: (1) when the most important segments are removed, reading would be most disrupted; (2) when the least important segments are removed, reading would be least disrupted; and (3) the most important strokes would be those that retained the character configuration and the least important to be those that were less related to character configuration.

Similar to Yan et al. (2012), we had subjects read sentences with characters that had elements deleted. In contrast to Yan et al., we deleted segments, as opposed to strokes and only used one degree of removal: 30%. In the present study, we removed (1) the least important segments according to SVD, (2) the most important segments according to SVD or (3) randomly selected segments. We also included a control condition in which no segments were deleted.

To delete segments, we first determined the important components identified by SVD as most (highest ranked) or least (lowest ranked) important (Figure 2). Using the same principles, we then reconstructed the characters using only a subset of the components. As demonstrated in Figure 3, reconstruction of characters using some components from SVD results in some pixels from a given segment being retained, but not all. To create the characters that were ultimately used in the experiment, we completely retained a segment when the SVD method reconstructed at least 35% of the original, or completely removed it if the SVD method reconstructed less than 35% of the original.

Method

Subjects

Sixteen students at the University of Massachusetts at Boston were recruited for the experiment. All subjects were native speakers of Chinese between the ages of 19 and 30 years

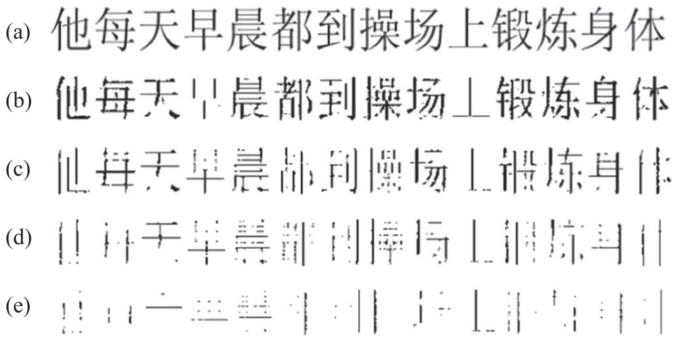


Figure 2. Singular value decomposition (SVD) reduction for Chinese characters. Sentence (a) is the original sentence without any reduction. Sentences (b) to (e) are the sentences after removing the least important 20%, 40%, 60% and 80% information as determined by SVD.

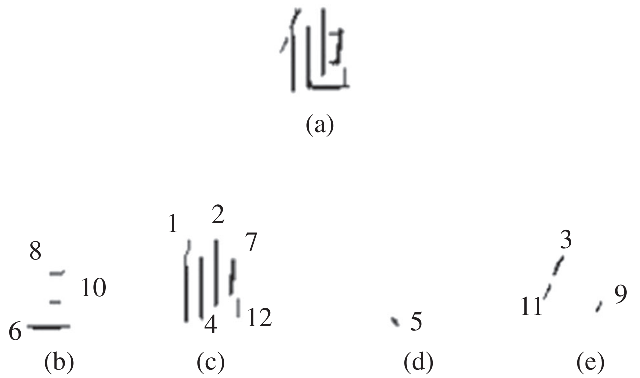


Figure 3. Determining the importance of segments by singular value decomposition (SVD). (a) Original character, (b) horizontal segments, (c) vertical segment, (d) to (e) diagonal segments. Numbers 1 to 12 represent from the most (highest ranked) to least (lowest ranked) important segments identified by SVD.

with normal or corrected-to-normal vision. Each participant received 10 dollars for participation in a half-hour session.

Materials

Sentences were taken from the 50 sentences of Yan et al. (2012); two sentences were used as practice stimuli, and 48 were used in the experiment. Each sentence contained 14 characters for which character frequency ranged from 21 to 38,075 per million (mean: 3,343); frequencies were taken from the China Newspaper Database (Yan et al., 2012). The 48 sentences for the experiment were presented in four experimental conditions: (1) all segments retained; (2) the least important 30% of segments removed; (3) the most important 30% of segments removed; and (4) 30% of segments randomly selected to be removed. The degree of deletion was set to be as close to 30% of the total number of pixels as possible. This method, deleting characters based on the number of pixels, does not consider spatial information such as orientation or length of segments. For example, two segments with the same length but different thicknesses would not be considered equivalent sizes. Stimuli were counterbalanced so that each subject read 12 sentences in each condition, and across all subjects, each sentence was seen equally often in each condition. Each sentence was followed by a true/false comprehension question. Examples of these stimuli are shown in Figure 4.

Apparatus

Eye movements were recorded using an SR Research EyeLink 1000 system with a sampling frequency of 1000 Hz run in desktop mode. After calibration, the average calibration error was 0.5°. Stimuli were presented on a 19-inch Dell P992 CRT monitor (Round Rock, TX, USA) with a refresh rate of 85 Hz and a screen resolution of 1,024 × 768 pixels. Each character subtended about 1.5° visual angle.

Procedure

Subjects were instructed to read the degraded sentences and try to understand them as accurately as possible. At the beginning of the experiment, a standard three-point calibration and

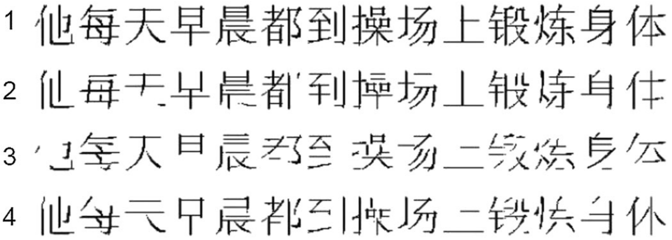


Figure 4. Four conditions of reading stimuli. Sentence 1 represents the all retained condition, sentence 2 represents the least important 30% segments removed condition, sentence 3 represents the most important 30% segments removed condition and sentence 4 represents the 30% randomly selected segments condition.

validation of the gaze recording were completed. Following two practice sentences, subjects viewed 48 sentences in random order. Each sentence was followed by a true or false question. At the start of each sentence, a calibration box appeared at the position of the first character, and once a fixation was detected inside the box, the sentence appeared. If the fixation did not trigger the sentence to appear, the experimenter recalibrated the tracker and continued the experiment. Subjects read the sentences at their own pace and pressed a button to indicate they had finished. Subsequently, the stimulus disappeared and the question appeared, and subjects responded yes or no by pressing the corresponding button.

Results

We examined a number of *global reading measures*, which index reading efficiency (Rayner, 1998, 2009). These measures include *comprehension accuracy*, *total sentence reading time*, *mean fixation duration*, *number of progressive saccades*, *number of regressive saccades* and *mean forward saccade length*. In general, reading efficiency is positively related to comprehension accuracy and mean forward saccade length and negatively related to total sentence reading time, mean fixation duration, number of progressive saccades and number of regressive saccades. In addition to global reading measures, we also report local reading measures on the words and individual characters in the sentence: *first fixation duration*, *gaze duration* and *total time* (Appendix B).

Trials with a total sentence reading time of less than 1,000 ms were considered as accidentally terminated (see means and standard deviations in Table 1), and trials with total reading time above four standard deviations from the mean were considered outliers. These trials were excluded from analysis, which removed 5.8% of the data. Means and standard deviations of global reading measures are shown in Table 1. Repeated measures one-way ANOVAs by-subjects (F_1) and by-items (F_2) for four experimental conditions were performed, and then post hoc tests using SPSS corrected Bonferroni adjusted p-values were carried out for paired comparisons between conditions.

Comprehension accuracy

Comprehension accuracy was high in all conditions: 90% in the all retained condition, 89% in the least important removed condition, 80% in the most important removed condition and 85% in the randomly removed condition. One-way ANOVAs revealed an overall difference in accuracy across the four conditions, $F_1(3, 45) = 3.37$, $p < .05$, $F_2(3, 141) = 4.61$, $p < .01$. Post hoc tests indicated that texts in the least important removed condition were comprehended better than in the most important removed condition by subjects and by items

Table 1. Means and standard deviations of global reading measures across experimental conditions. Standard deviations are shown in parentheses.

	Removal condition			
	All retained	Most important	Least important	Randomly selected
Comprehension accuracy	0.90 (0.08)	0.80 (0.13)	0.89 (0.08)	0.85 (0.08)
Total sentence reading time (s)	3.98 (1.30)	9.90 (4.12)	4.29 (1.18)	5.73 (1.59)
Mean fixation duration (ms)	209 (40)	254 (36)	219 (34)	226 (36)
Number of progressive saccades	10.00 (3.29)	19.76 (7.81)	10.52 (2.59)	13.09 (3.43)
Number of regressive saccades	4.96 (2.09)	12.98 (5.82)	5.27 (1.98)	7.23 (2.44)
Mean saccade length (degrees)	2.20 (0.59)	1.76 (0.43)	1.93 (0.48)	1.87 (0.49)

(both ps $< .05$). The comparison between the all retained condition and the most important removed condition was significant by items ($p < .05$) but not by subjects ($p = .21$). None of the other comparisons were significant (all ps $> .21$). These results indicate that subjects read and comprehended the sentences well, but that removing the segments identified as the most important via SVD caused some comprehension difficulty.

Total sentence reading time

One-way ANOVAs revealed significant differences across the conditions (Figure 5), $F_1(3, 45) = 31.02$, $p < .001$, $F_2(3, 141) = 54.21$, $p < .001$. Post hoc tests showed that subjects read slowest in the most important removed condition compared with each of the other conditions (all ps $< .01$). Subjects read slower in the randomly removed condition than in the all retained condition (both ps $< .001$) and the least important removed conditions by subjects ($p < .001$) and marginally by items ($p = .06$). There was no significant difference between the least important removed condition and the all retained conditions (both ps $> .12$). These results indicate that reading fluency was reduced not only in the most important removed condition but also the randomly removed condition. In contrast, readers

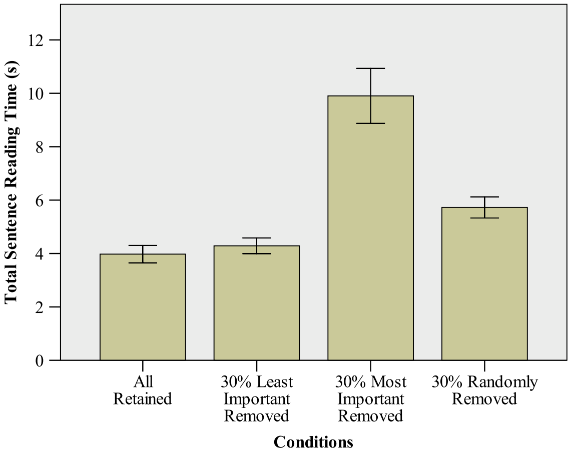


Figure 5. Total sentence reading time as a function of condition. Error bars are based on standard errors.

read the sentences as efficiently in the least important removed condition as in the all retained condition.

Mean fixation duration

One-way ANOVAs revealed a significant overall effect of condition on mean fixation duration, $F_1(3, 45) = 47.20$, $p < .001$, $F_2(3, 141) = 25.98$, $p < .001$. Post hoc tests revealed that the mean fixation durations in the most important removed condition were significantly longer than each of the other conditions (all *p*s < .001). Mean fixation duration in the randomly removed condition was longer than that in the all retained condition ($p < .01$). Mean fixation duration in the least important removed condition was longer than in the all retained condition in the by-items analysis ($p < .05$) but not in the by-subjects analysis ($p = .09$). None of the other comparisons were significant (all *p*s > .09).

Number of saccades

Similar to the total sentence reading time analyses, one-way ANOVAs revealed a significant overall effect of condition on the number of progressive saccades, $F_1(3, 45) = 28.20$, $p < .001$, $F_2(3, 141) = 48.71$, $p < .001$ (Figure 6). Post hoc tests revealed that there were more progressive saccades in the most important removed condition compared with each of the other conditions (all *p*s < .01). More progressive saccades were produced in the randomly removed condition than in the all retained and the least important removed conditions (both *p*s < .05). There was no significant difference between the least important removed and the all retained conditions ($p > .13$).

One-way ANOVAs revealed an overall effect of condition on the number of regressive saccades to be significant, $F_1(3, 45) = 29.66$, $p < .001$, $F_2(3, 141) = 51.41$, $p < .001$. Post hoc tests showed that there were more regressive saccades in the most important removed condition compared with each of the other conditions (all *p*s < .01). More regressive saccades were found in the randomly removed condition than in the

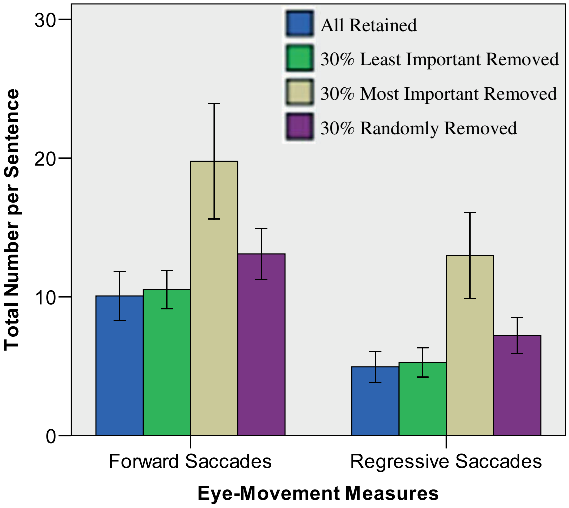


Figure 6. Number of forward and regressive saccades per sentence. Error bars are based on standard errors.

all retained and the least important removed conditions (both $p\text{s} < .05$). There was no significant difference between the least important removed and the all retained conditions ($p > .52$).

Mean forward saccade length

One-way ANOVAs revealed a significant overall effect of condition, $F_1(3, 45) = 29.62$, $p < .001$, $F_2(3, 141) = 7.12$, $p < .001$. Post hoc tests of both by-subjects and by-items analyses indicated that forward saccade length in the all retained condition was longer than in the least important removed, the most important removed and randomly removed conditions in the by-subjects analysis (all $p\text{s} < .001$) and in the by-item analysis ($p = .06$, $p < .01$ and $p < .05$, respectively). Forward saccade length in the least important removed condition was significantly longer than in the most important removed condition in the by-subjects analysis ($p < .05$) but not in by-items analysis ($p > .24$). Forward saccade length in the randomly removed condition was marginally longer than in the most important removed condition in the by-subjects analysis ($p = .07$) but not in the by-items analysis ($p > .50$).

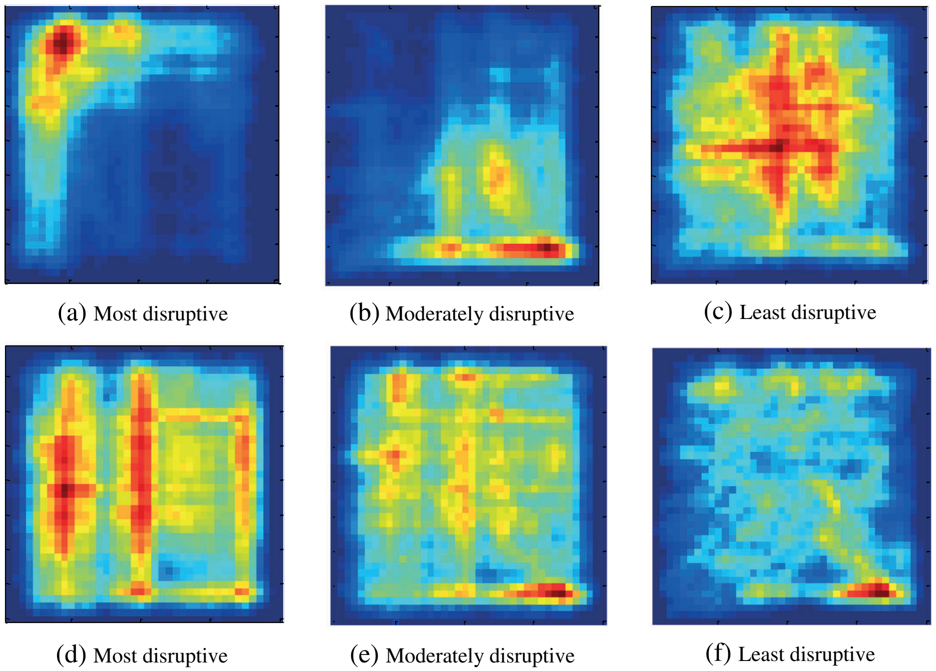
Taken together, these results suggest that reading was most disrupted when the most important segments were removed, moderately disrupted when randomly selected segments were removed and least disrupted (equivalent to reading with all strokes retained) when the least important strokes were removed. Our data are similar to the data reported by Yan et al. (2012), in that some conditions (beginning strokes removed in their study and most important segments removed in our study) were more disruptive than others (ending strokes removed in their study and randomly selected strokes in our study). Furthermore, both studies found that removing certain character elements (those that did not contribute to the configuration in their study and the least important segments in our study) did not alter reading compared with reading intact characters.

Additional analyses

The overall results indicate that the mathematical method SVD captured the most informative segments of Chinese characters. However, it is possible that SVD identifies the same strokes that were deleted in the study of Yan et al. Alternatively, SVD may identify a different set of segments in the characters, but if so, what is it about those segments that is most informative and impair reading most when deleted?

Distribution of degradation position

Because stroke order is correlated to stroke position (e.g., beginning strokes tend to be located at the top-left position of characters), we compared the distribution of deleted elements from the different conditions by Yan et al. (2012) and the conditions used in the current study. As shown in Figure 7, the distribution of the positions where strokes were removed in the study of Yan et al. (2012) is different from the distribution in the current study. As expected for the materials of Yan et al., the top-left positions tended to be removed in the beginning stroke removal condition, the bottom-right positions tended to be removed in the ending strokes removal condition, and character internal positions tended to be removed in the configuration-retaining condition. In contrast, although SVD identified the most important segments



(a) Most disruptive (b) Moderately disruptive (c) Least disruptive

(d) Most disruptive (e) Moderately disruptive (f) Least disruptive

Figure 7. (a) to (c): Distributions of removed strokes (30%) by Yan et al. (2012), which were the most, moderately and the least disruptive. (a) Beginning strokes, (b) ending strokes and (c) configuration-retaining strokes. (d) to (f): Distributions of removed segments (30%) in this study which were the most, moderately and the least disruptive. (d) Most important segments removed, (e) random segments removed and (f) least important segments removed.

as those located on the left side of the character, they were more widely distributed than those in the study of Yan et al. Similarly, SVD identified the least informative segments as those in the bottom-right location of the character, again, with a wider distribution than in the study of Yan et al. Lastly, the configuration-retaining condition (from the study of Yan et al.) and the randomly selected condition (from the current study) are similarly centred around the middle of the character but, again, with a wider distribution with the SVD method. Thus, it seems that the beginning strokes/segments (i.e., those located in the upper and left-side positions of the character) tend to be more important for character identification than those that are located in the bottom and right-side positions.

Because these locations were also identified by SVD, which has no information about the order of writing strokes, the correlation between writing order and character configuration may be the cause of these strokes' importance. However, the configuration of a Chinese character is not well defined. In the following analyses, we propose a computational method for representing character configuration and measuring the degradation percentage of characters of Yan et al. (2012) and this study.

Measuring character configuration using contour

In object recognition, contour is important for successful recognition of degraded objects; observers are more accurate at identifying degraded characters when vertices are retained than when the midsections of lines are retained (see Biederman, 1987, for a review). We extracted vertices from each segment and simplified the contour of the character using

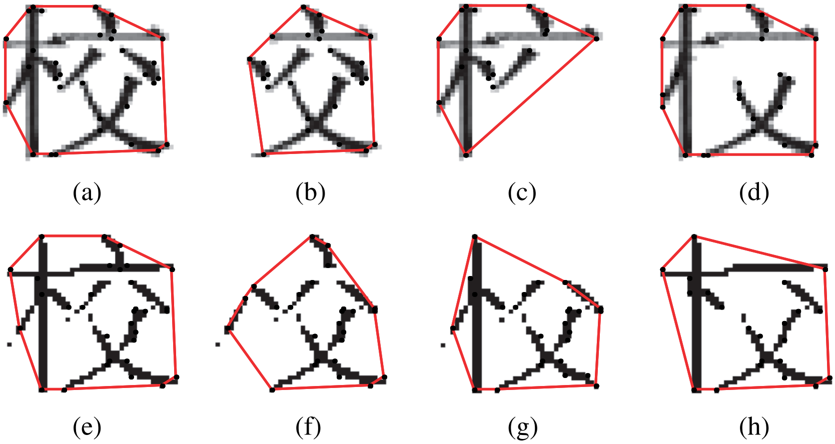


Figure 8. Defining the contours of characters using convex hull. (a) to (d) are sample characters by Yan et al. (2012). (a) Character without removal, (b) the most disruptive removal (beginning strokes), (c) moderately disruptive removal (ending strokes) and (d) the least disruptive removal (configuration-retaining). (e) to (h) are sample characters in this study. (e) All retained, (f) the most disruptive removal (most important segments), (g) moderately disruptive removal (randomly selected segments) and (h) the least disruptive removal (least important segments).

convex hull – the shape formed by a 'tight rubber band' that surrounds all the vertices, shown in Figure 8.

Two similarity measures were computed between the original character and each of the degraded characters for the stimuli of Yan et al. and our own stimuli: the *proportion of overlapping vertices* and the *proportion of overlapping perimeters*. Overlapping vertices are the number of matching vertices, and overlapping perimeters are the sum of length of matching edges. As shown in Table 2, the results indicate that the least disruptive conditions yielded the highest similarity with the original characters, whereas the most disruptive conditions yielded the lowest similarity. A one-way ANOVA for the degree of disruption (most, moderately and least disruptive conditions; 48 sentences each condition) was performed, and then post hoc tests using SPSS corrected Bonferroni adjusted *p*-values were conducted for paired comparisons between conditions.

In the analysis for the stimuli of Yan et al. (2012), the results showed significant main effects of the proportion of overlapping vertices and perimeters, $F(2, 94) = 589.89$, $p < .001$

Table 2. The means and standard deviations of similarity measures of contour using convex hull between undegraded and degraded characters of Yan et al. (2012) and the current study.

	Yan et al. (2012)		Current study	
	Proportion of overlapping vertices	Proportion of overlapping perimeters	Proportion of overlapping vertices	Proportion of overlapping perimeters
Most disruptive	0.59 (0.04)	0.46 (0.05)	0.71 (0.05)	0.47 (0.07)
Moderately disruptive	0.73 (0.04)	0.52 (0.05)	0.78 (0.05)	0.61 (0.07)
Least disruptive	0.86 (0.04)	0.75 (0.07)	0.80 (0.04)	0.63 (0.07)

and $F(2, 94) = 390.51$, $p < .001$, respectively. Post hoc tests revealed that the proportion of overlapping vertices and perimeters of the most disruptive conditions was significantly lower than that of the moderately and least disruptive conditions (all ps $< .001$), and that of the moderately disruptive conditions was significantly lower than that of the least disruptive conditions (all ps $< .001$). In the current study, there was also an overall effect across conditions for the proportion of overlapping vertices, $F(2, 94) = 52.72$, $p < .001$, and perimeters, $F(2, 94) = 68.60$, $p < .001$. Post hoc tests revealed that the moderately and least disruptive conditions contained characters with higher proportions of overlapping vertices and perimeters than the most disruptive conditions (all ps $< .001$). However, although the least disruptive conditions had slightly more overlapping vertices and perimeters than the moderately disruptive conditions, this difference was not statistically significant (both ps $> .34$). These results indicate that the least disruptive conditions are those that retained most of the contours of the original characters. Thus, degrading the contour of the original character is likely what made the beginning stroke condition (in the Yan et al. study), and the most important removed condition (in the present study) lead to the poorest reading efficiency. The nonsignificant results between least and moderately disruptive conditions in the current study may imply that contour is not the only factor that influences the degree of disruption produced by deleting strokes; degradation position may also be important.

General discussion

In this study, we investigated how readers recognised Chinese characters that were degraded using SVD to identify the most important and least important segments. We found that reading was most impaired when subjects read sentences with the most important segments removed. Reading was not impaired when the least important segments were removed and reading was moderately impaired when randomly selected segments were removed. Our data suggest that SVD is a powerful tool in determining what the most informative segments of Chinese characters are.

Comparisons between spatial distributions of deleted strokes between the most disruptive and least disruptive conditions for the present study and the study of Yan et al. (2012) revealed that the most important strokes in both studies tended to be located on the left side of the character and the least important tended to be located in the bottom-right portion of the character. In both studies, the most important strokes/segments for Chinese character identification are the ones that retained the character configuration. When these elements are removed, the contour of the character changes most dramatically, and consequently, reading is most impaired. Conversely, the least important strokes/segments are those that, when deleted, do not greatly change the contour of the character and therefore cause no reading impairment. We suggest that the convex hull might be a useful tool for measuring the configuration of a character.

Written order and low-level visual information

Here, we have suggested that low-level character configuration influences character recognition. However, the study of Yan et al. (2012) suggests that the important strokes are determined by writing order. In fact, there may be a fundamental relationship between the visual configuration of the character and the stroke order children are taught to write. Our study cannot distinguish whether the correlation between low-level visual configuration information and stroke written order is coincidental or cognitively determined. Further studies should directly assess this issue.

There are important educational implications of this research. For hundreds of years, learning written stroke order has been considered essential to learning Chinese characters. However, it is unclear whether learning stroke order facilitates character recognition and production (Law, Ki, Chung, Ko & Lam, 1998) or whether such practices are just a vestige of traditional teaching practices. With the advent of computers, writing Chinese has transitioned from handwriting to using word processing software. Presently, many Chinese typing methods do not require the knowledge of written stroke order but rather knowledge of the phonological form (Pinyin) of the word. Thus, with the growing use of computers and mobile devices to produce electronic writing, it is important to know whether learning traditional written stroke order is beneficial to Chinese children and second language learners. The results of the present study suggest that the first-written strokes of Chinese characters constitute important visual (configural) information, which may support successful Chinese character learning (Law et al., 1998).

Why left hand and exterior strokes are important: position, radicals, foveal and parafoveal processing

In addition to the factors mentioned previously, the left-hand and exterior strokes may be important for several possible reasons. First, semantic radicals tend to be located on the left or top side of Chinese characters, and there is much evidence from a range of paradigms to suggest that reading a complex character involves the processing of its component radicals (Taft et al., 1999; Zhou, Ye, Cheung & Chen, 2009). Therefore, when the strokes/segments that contribute to these radicals are missing, reading is more impaired than when other strokes/segments are missing. Alternatively, radicals that are on the top and left-hand side of the character tend to have fewer strokes than radicals on the bottom or right-hand side. Therefore, deleting these strokes would delete a greater proportion of the character, leading to more impaired reading.

Alternatively, external strokes might be important because they may be more susceptible to lateral masking (Bouma, 1973) from adjacent characters (see White, Johnson, Liversedge & Rayner, 2008, for an investigation in English) because there are no spaces between words or characters in Chinese. This lateral masking may lead to poorer parafoveal identification of the upcoming character and thus a greater importance of these strokes when the character is actually fixated.

In short, the present study demonstrates that SVD can identify the most and least informative strokes of Chinese characters. When these strokes are deleted, reading is impaired more or less, respectively. These data are similar to the data reported by Yan et al. (2012), using a different method. Both methods identify left strokes/segments as being the most informative for Chinese character identification.

Acknowledgements

Parts of the data were presented at the 16th European Conference on Eye Movements (ECEM 2011) and the Fifth China International Conference on Eye Movements (CICEM 2012). Preparation of the article was supported by NIH grants EY021802 to Marc Pomplun, HD026765 to Keith Rayner and a fellowship to Elizabeth R. Schotter from DC000041.

References

Balota, D.A., Pollatsek, A. & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology*, 17, 364–390.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2), 115–147.

Bouma, H. (1973). Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Research*, 13, 767–782.

Craw, I. & Cameron, P. (1991). Parameterising images for recognition and reconstruction. In P. Mowforth (Ed.), *Proceedings of the British Machine Vision Conference, 1991*. Berlin: Springer Verlag.

Elden, L. (2007). Matrix methods in data mining and pattern recognition. Society of Industrial and Applied Mathematics, Cambridge.

Evett, L.J. & Humphreys, G.W. (1981). The use of abstract graphemic information in lexical access. *Quarterly Journal of Experimental Psychology*, 33A, 325–350.

Garner, W.R. (1981). The role of configuration in the identification of visually degraded words. *Memory & Cognition*, 9(5), 445–452.

Gough, P.B. (1972). One second of reading. In J.F. Kavanagh & I.G. Mattingly (Eds.), *Language by eye and by ear*. Cambridge, MA: MIT Press.

Hastie, T. & Simard P.Y. (1998). Metrics and models for handwritten character recognition. *Statistical Science*, 13, 54–65.

Healy, A.F. (1994). Letter detection: A window to unitization and other cognitive processes. *Psychonomic Bulletin & Review*, 1, 333–344.

Healy, A.F. & Drewnowski, A. (1983). Investigating the boundaries of reading units: Letter detection in misspelled words. *Journal of Experimental Psychology: Human Perception and Performance*, 9(3), 413–426.

Hubel, D.H. & Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160, 106–154.

Hubel, D.H. & Wiesel, T.N. (1963). Receptive fields of cells in striate cortex of very young, visually inexperienced kittens. *Journal of Neurophysiology*, 26, 994–1002.

Jordan, T.R., Thomas, S.M., Patching, G.R. & Scott-Brown, K.C. (2003). Assessing the importance of letter pairs in initial, exterior, and interior positions in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(5), 883–393.

Law, N., Ki, W.W., Chung, A.L.S., Ko, P.Y. & Lam, H.C. (1998). Children's stroke sequence errors in writing Chinese characters. *Reading and Writing: An Interdisciplinary Journal*, 10, 267–292.

Martelli, M., Majaj, N.J. & Pelli, D.G. (2005). Are faces processed like words? A diagnostic test for recognition by parts. *Journal of Vision*, 5, 58–70.

McClelland, J.L. & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88, 375–407.

McConkie, G.W. & Zola, D. (1979). Is visual information integrated across successive fixations in reading? *Perception & Psychophysics*, 25(3), 221–224.

Rayner, K. (1998). Eye movement in reading and information processing: 20 years of research. *Psychological Bulletin*, 24, 372–422.

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62, 1457–1506.

Rayner, K. & Kaiser, J.S. (1975). Reading mutilated text. *Journal of Educational Psychology*, 67, 301–306.

Rayner, K., White, S., Johnson, R. & Liversedge, S. (2006). Raeding wrods with jubmled lettres: There is a cost. *Psychological Science*, 17, 192–193.

Selfridge, O.G. (1959). Pandemonium: A paradigm for learning. In D.V. Blake & A.M. Uttley (Eds.), *The mechanisation of thought processes*. (pp. 511–529). London: HM Stationery Office.

Slattery, T.J., Angele, B. & Rayner, K. (2011). Eye movements and display change detection during reading. *Journal of Experimental Psychology: Human Perception and Performance*, 37, 1924–1938.

Strang G. (1993). *Introduction to Linear Algebra*. (2nd edn). Wellesley, MA: Cambridge University Press.

Taft, M. (1985). The decoding of words in lexical access: A review of the morphographic approach. In D. Besner, T.G. Waller & G.E. MacKinnon (Eds.), *Reading research: Advances in theory and practice*. (pp. 83–126). New York: Academic Press.

Taft, M., Zhu, X. & Peng, D. (1999). Positional specificity of radicals in Chinese character recognition. *Journal of Memory and Language*, 40, 498–519.

Torralba, A. & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, 14, 391–412.

Tseng, S.C., Chang, L.H. & Wang C.C. (1965). An informational analysis of the Chinese language: I. The reconstruction of the removed strokes of the ideograms in printed sentence-texts [in Chinese]. *Acta Psychologica Sinica*, 10, 299–306.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3, 71–86.

White, S.J., Johnson, R.L., Liversedge, S.P. & Rayner, K. (2008). Eye movements when reading transposed text: The importance of word beginning letters. *Journal of Experimental Psychology: Human Perception and Performance*, 34, 1261–1276.

Yan, G., Bai, X., Zang, C., Bian, Q., Cui, L., Qi, L. et al. (2012). Using stroke removal to investigate Chinese character identification during reading: Evidence from eye movements. *Reading and Writing*, 25, 951–979.

Zhang, J.J., Wang, H.P., Zhang, M. & Zhang, H.C. (2002). The effect of the complexity and repetition of the strokes on the cognition of the strokes and the Chinese characters. *Acta Psychologica Sinica*, 34(5), 449–453 (in Chinese).

Zhou, X., Ye, Z., Cheung, H. & Chen, H.-C. (2009). Processing the Chinese language: An introduction. *Language & Cognitive Processes*, 24, 929–946.

Appendix A

The SVD (Strang, 1993) is a powerful linear algebra factorisation technique of a rectangular matrix. Any $m \times n$ matrix A can be decomposed into a product of three matrices:

$$\begin{aligned} A &= U\sum V^T \\ &= \underbrace{\begin{bmatrix} | & | & | \\ u_1 \ldots & u_r \ldots & u_m \\ | & | & | \end{bmatrix}}_{m\times m} \underbrace{\begin{bmatrix} \sigma_1 & & \\ & \vdots & \\ & & \sigma_r \end{bmatrix}}_{m\times n} \underbrace{\begin{bmatrix} - & v_1 & - \\ & \vdots & \\ - & v_r & - \end{bmatrix}}_{n\times n}, \end{aligned}$$

where U is an $m \times m$ matrix of orthonormal columns, V^{T} is an $n \times n$ matrix of orthonormal rows and ⊠ is a non-negative $m \times n$ matrix with singular values $_q$, …, σ_r. Most software packages for numerical calculations such as MATLAB contain SVD.

Appendix B

Appendix Table B1. Means and standard deviations of local reading measures on all words and individual characters in the sentence across experimental conditions. Standard deviations are shown in parentheses.

	Removal condition			
	All retained	Most important	Least important	Randomly selected
Character-based				
First fixation duration (ms)	203 (24)	235 (35)	215 (35)	218 (34)
Gaze duration (ms)	218 (42)	280 (57)	236 (45)	245 (43)
Total time (ms)	339 (81)	688 (236)	368 (80)	435 (89)
Word-based				
First fixation duration (ms)	202 (33)	234 (35)	213 (32)	216 (34)
Gaze duration (ms)	259 (68)	402 (105)	300 (71)	321 (81)
Total time (ms)	447 (125)	1001 (388)	493 (119)	593 (130)

Hsueh-Cheng Wang received his PhD in Computer Science from the University of Massachusetts at Boston in 2013. He is now a postdoctoral associate at the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT). His research deals with reading and scene viewing using eye-movement analysis and computational models, as well as assistive technology, machine perception and robotics.

Elizabeth R. Schotter is completing her dissertation research in Experimental Psychology at the University of California, San Diego. Her primary research interests focus on language processing, reading, speech production and visual decision-making.

Bernhard Angele is completing his dissertation research in Experimental Psychology at the University of California, San Diego. His primary research interests focus on the use of parafoveal information during reading.

Jinmian Yang obtained her PhD in Cognitive Psychology from the University of Massachusetts, Amherst in 2010. She is currently a post-doctoral researcher at the University of California, San Diego. Her research has primarily focused on the nature of parafoveal preview processing during the reading of Chinese and English, and the time course of semantic and syntactic processing in reading Chinese.

Dan Simovici is Professor and the Graduate Program Director of Computer Science at the University of Massachusetts at Boston. His research interests are information-theoretical and linear methods in data mining, semantic models in databases and algebraic aspects of multiple-valued logic.

Marc Pomplun is Professor of Computer Science at the University of Massachusetts at Boston and the Director of the Visual Attention Laboratory. His work focuses on analysing, modelling and simulating aspects of human vision.

Keith Rayner is the Atkinson Professor of Psychology at the University of California, San Diego. His research focuses on reading, but he also has interests in scene perception and visual search.

Received 3 September 2012; revised version received 29 January 2013.

Address for correspondence: Hsueh-Cheng Wang, Department of Computer Science, University of Massachusetts at Boston, 100 Morrissey Blvd,Boston, MA 02125-3393, USA. E-mail: *hchengwang@gmail.com*