# Using Singular Value Decomposition to Investigate Degraded Chinese Character Recognition: Evidence from Eye Movements During Reading

Hsueh-Cheng Wang[1], Elizabeth R. Schotter[2], Bernhard Angele[2], Jinmian Yang[2], Dan Simovici[1], Marc Pomplun[1], Keith Rayner[2]

1. Department of Computer Science, University of Massachusetts at Boston, USA
2. Department of Psychology, University of California, San Diego, USA

Word Recognition

- The visual system recognizes objects and words through a hierarchical process beginning with *feature* detection (e.g., Biederman, 1987; Gibson et al., 1963; Hubel & Wiesel, 1962, 1963; McClelland and Rumelhart, 1981; Selfridge, 1957).
- These features are **combined** into higher-level, more meaningful *components* (see McClelland & Rumelhart, 1981),
- Word recognition is mediated by an analysis of a word's component *letters* (Balota, Pollatsek & Rayner, 1985; Taft, 1985; McConkie & Zola, 1979; Slattery, Angele & Rayner, 2011).

DARK

D   A   R   K
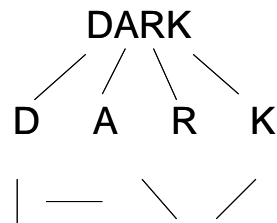
In alphabetic languages

- Not all letters are of equal importance to the word recognition process
  - Changes to *initial* letters are more disruptive than changes to medial or final letters (Rayner & Kaiser, 1975; Rayner, White, Johnson, & Liversedge, 2006).
  - *Exterior* letters are more important than word internal letters (Jordan, Patching, & Thomas, 2003; Rayner et al., 2006).

The orthographic units of a Chinese character

- *Strokes* are simple **features** (e.g., dots, lines, or curves) or combinations of simple features that vary in complexity (Zhang, Wang, Zhang, & Zhang, 2002)
- Strokes order generally follows the order of left to right, top to bottom, and exterior to interior.

场 一 丨 乚 弓 丿 丿

Stroke Removal Studies

- Not all strokes within a character have equal status during character identification (see Tseng, Chang, and Wang, 1965; Yan, Bai, Zang, Bian, Cui, Qi, Rayner, & Liversedge, 2012)
    - Most disruptive: **beginning** strokes removed
    - Moderately disruptive: **ending** strokes removed
    - Least disruptive: **configuration retaining**
- "Configuration" is important

Is there something privileged about the first-written strokes?

- The stroke written order is correlated to stroke position.
- Does deleting first-written strokes alter a greater proportion of the configuration of a character, leading to more impaired reading?
- Are first-written strokes consistent with "the least redundant component" determined by SVD (see below)?

Singular Value Decomposition
(*SVD*, Strang, 1993)

- SVD, similar to Principal Component Analysis (PCA), is a dimension reduction method in linear algebra to retain **the least redundant components** contained in a matrix (see Elden, 2007)
- These dimension reduction methods have been extensively used in pattern recognition (Face, Scene, …)
- SVD has *no* information about writing order

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

$$= \begin{bmatrix} | & & | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r & \cdots & \mathbf{u}_m \\ | & & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} - & \mathbf{v}_1 & - \\ & \vdots & \\ - & \mathbf{v}_r & - \end{bmatrix},$$

$$\underbrace{}_{m \times m} \quad \underbrace{}_{m \times n} \quad \underbrace{}_{n \times n}$$

SVD reduction for Chinese characters

Decomposing
Strokes into Segments

- Four oriented filters decompose each character into simple **segments** that roughly map on to **features**
- Reducing bias toward longer or more complex strokes seeming more important

(a) 他每天早晨都到操场上锻炼身体

Experiment

- *Subjects.* Twelve students at the University of Massachusetts at Boston

- *Materials.* Sentences were taken from the 48 sentences in Yan et al. (2012) in four experimental conditions
    1. All segments retained,
    2. The least important 30% of segments removed,
    3. The most important 30% of segments removed,
    4. 30% of segments randomly selected to be removed.

Stimuli

- We reconstructed the characters using only a subset of the segments.

1 他每天早晨都到操场上锻炼身体
2 他每天早晨都到操场上锻炼身体
3 他每天早晨都到操场上锻炼身体
4 他每天早晨都到操场上锻炼身体

Total Sentence Reading Time
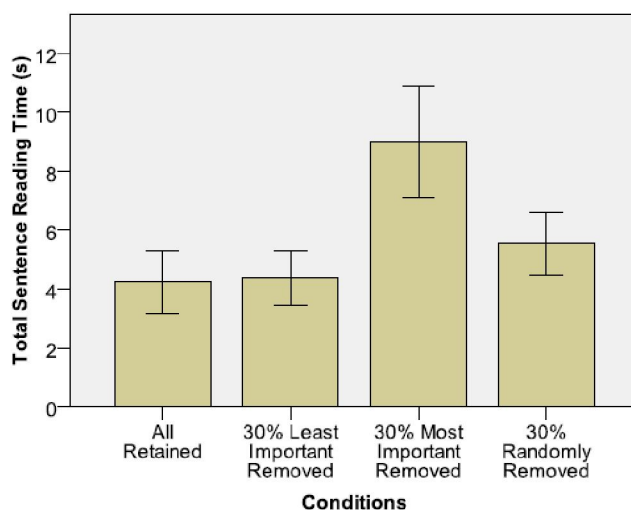
# Numbers of Forward and Regressive Saccades per Sentence
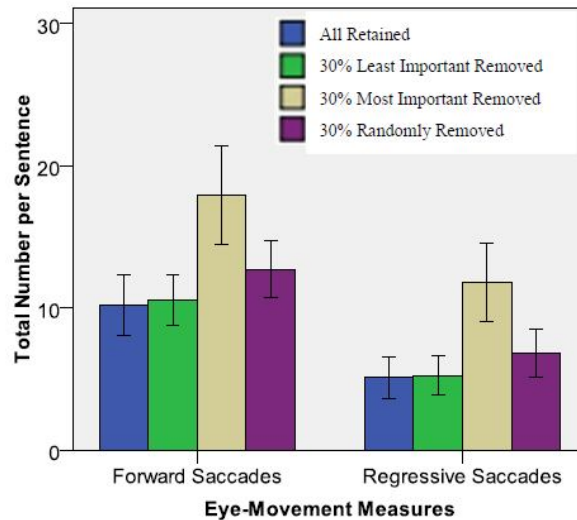
Discussion

- The overall results indicate that the mathematical method SVD captured the most informative segments of Chinese characters.
- Is it possible that SVD is identifying the same strokes that were deleted in the Yan et al. study? What is it about those segments that are most informative?
  - *Distribution of Degradation Position*
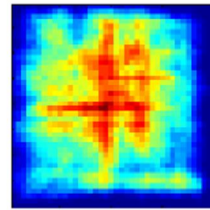 - *Measuring Character Configuration using Contour*

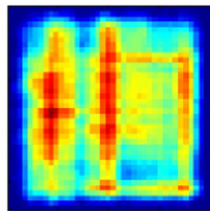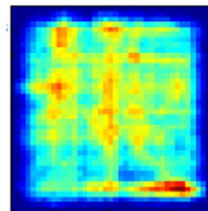# Distribution of Removed Strokes/ Segments
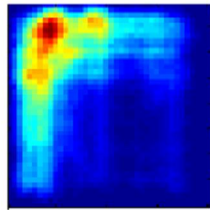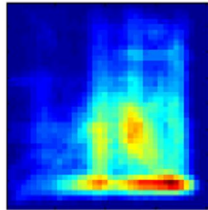
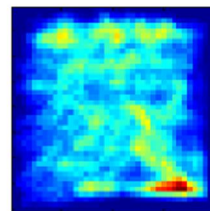(a) Most disruptive (b) Moderately disruptive (c) Least disruptive
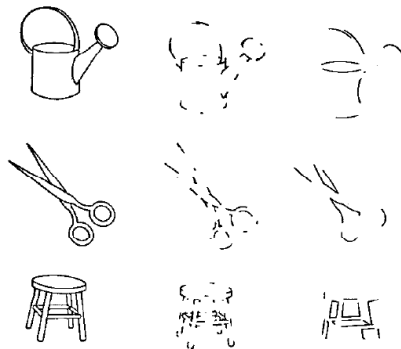
(d) Most disruptive (e) Moderately disruptive (f) Least disruptive

In Object Recognition

- **Contour** is important for successful recognition of degraded objects; observers are more accurate at identifying degraded objects when **vertices** are retained than when the midsections of lines are retained (see Biederman, 1987, for a review)

Contour of Characters using Convex Hull

Without Removal	Most Disruptive	Moderately Disruptive	Least Disruptive

(a) (b) (c) (d)

(e) (f) (g) (h)
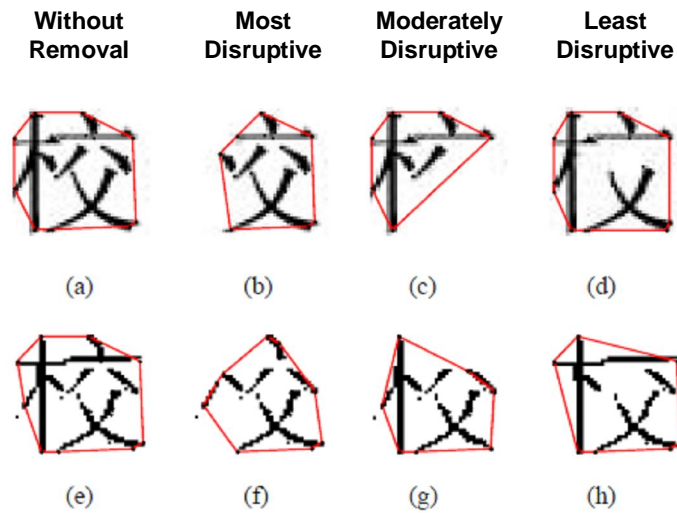
Similarity measures between undegraded and degraded characters

- The least disruptive conditions yielded the highest similarity with the original characters, while the most disruptive conditions yielded the lowest similarity

	Yan et al. (2012)		Current study	
	Proportion of Overlapping Vertices	Proportion of Overlapping Perimeters	Proportion of Overlapping Vertices	Proportion of Overlapping Perimeters
Most Disruptive	0.58 (0.17)	0.46 (0.18)	0.71 (0.18)	0.47 (0.26)
Moderately Disruptive	0.73 (0.16)	0.52 (0.23)	0.78 (0.18)	0.61 (0.27)
Least Disruptive	0.86 (0.15)	0.75 (0.23)	0.80 (0.15)	0.63 (0.24)

Conclusion

- These data are similar to the data reported by Yan et al. (2012)
  - The most important strokes tended to be located on the **left** side of the character and the least important tended to be located in the **bottom right** portion.
  - Semantic radicals usually located on the left or top side
  - Deleting these important strokes would delete a greater proportion of the character, leading to more impaired reading
- There is a correlation between first-written strokes and important segments determined by SVD, which has no information about writing order.