Visual Attention is Attracted by Text Features:
Experimental Data and Computational Model

Hsueh-Cheng Wang[1], Shijian Lu[2], Joo-Hwee Lim[2], and Marc Pomplun[1]

[1] Department of Computer Science, University of Massachusetts at Boston, USA

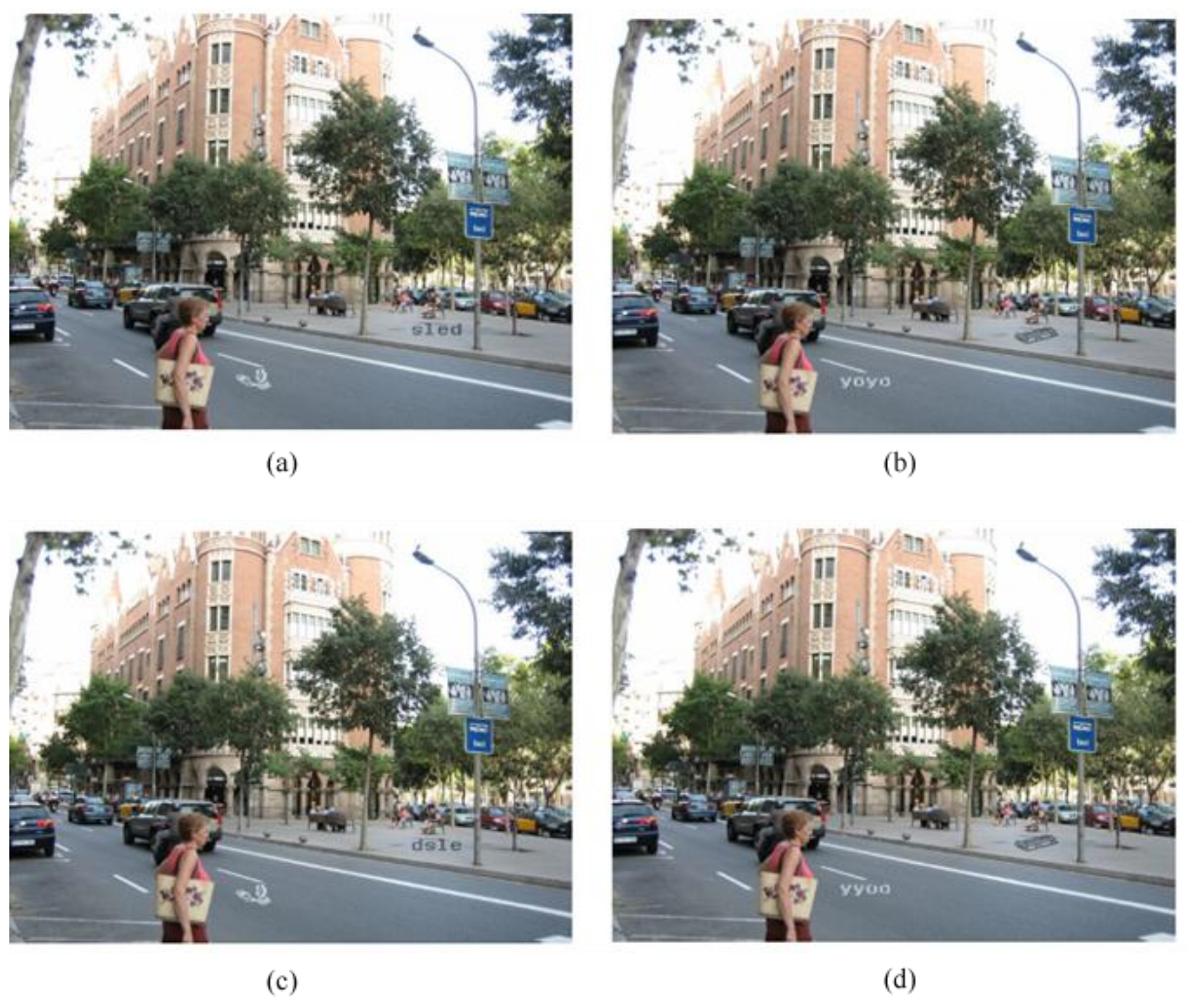
[2] Institute for Infocomm Research, A*STAR, Singapore

- Viewers' attention has been biased toward:
  - Low-level saliency (Itti & Koch, 2001)
  - Center of the screen (Tatler, 2007)
- Adding object locations enhances the ability of the saliency model to predict eye fixations in natural images
  - Add manually-defined regions of faces, texts, and cellphones (Cerf, Frady, & Koch, 2009)
  - Add automatic object detectors:
    - Face, person, & car (Judd, Ehinger, Durand, & Torralba, 2009)
    - Face (Zhao & Koch, 2011)
- Attention is disproportionately attracted by texts (Wang & Pomplun, 2012)
  - Expected locations, text features
- Automatic text detector (Lu, Wang, Lim, & Pomplun, submitted):
  - Specialized text features, e.g., histograms of edge width and edge density, trained with Support Vector Machine (SVM) classifiers.
- Can adding text detector to saliency model improve the prediction of viewers' fixations?
- Do viewers develop a "biological text detector" in visual system?

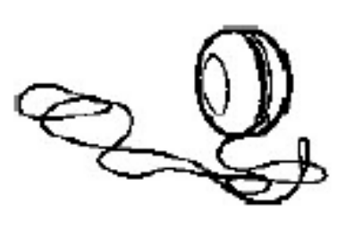
Experimental Data

Computational Model

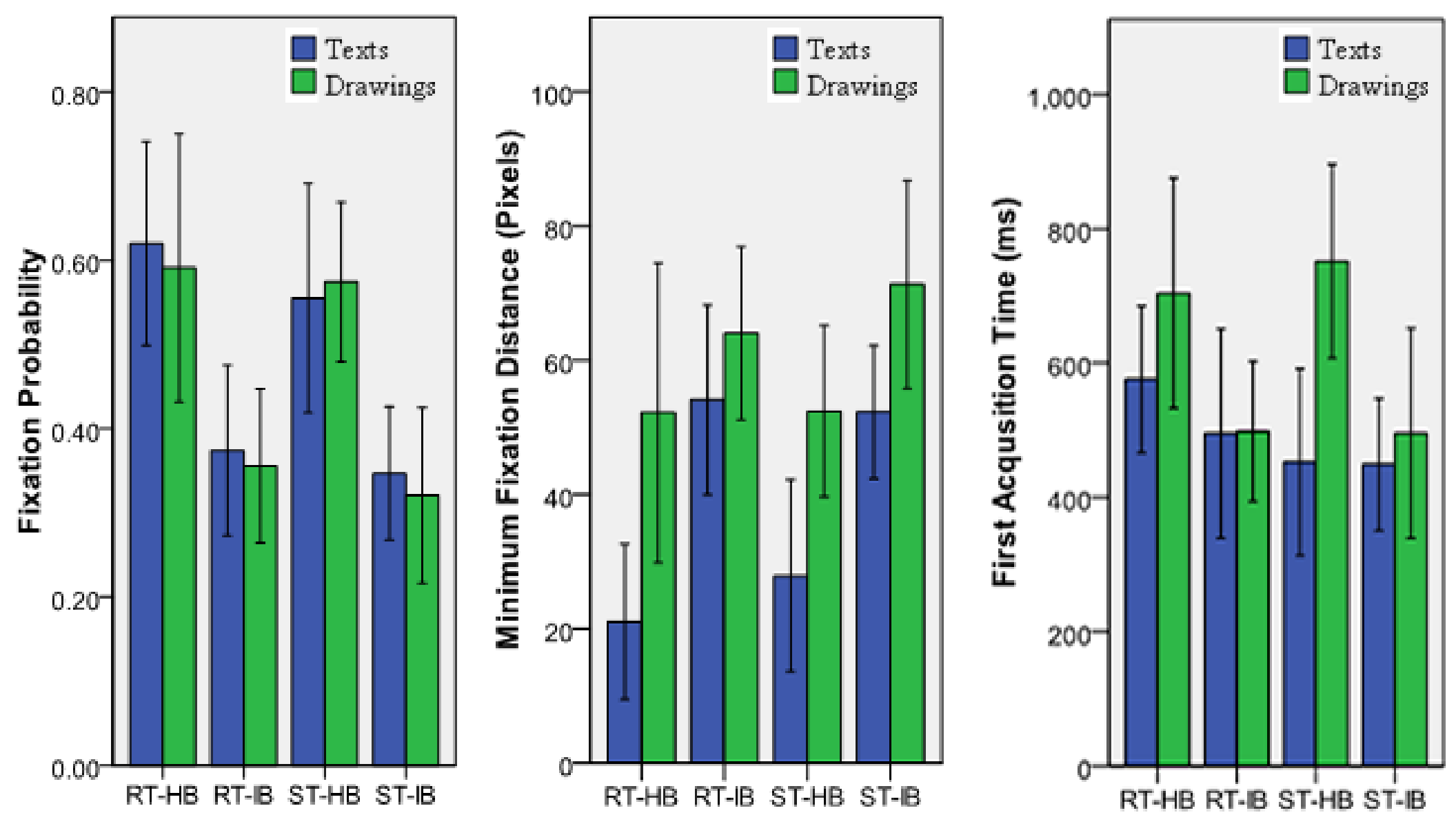
Data Set 1: Texts, Scrambled Texts and Drawings



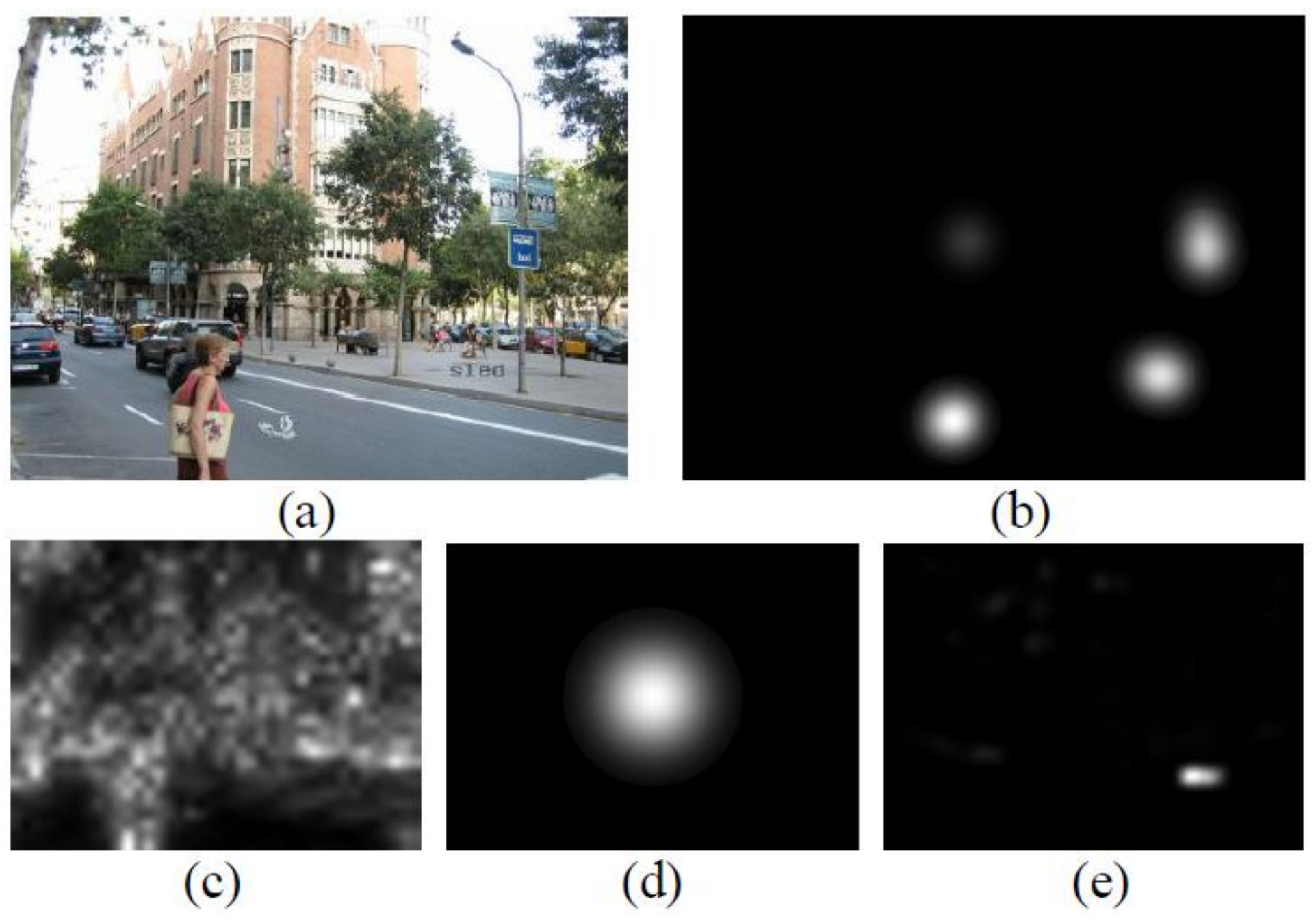
(a) (b) (c) (d)

	Item A	Item B
Word (scrambled word)	sled (dsle)	yoyo (yyoo)
Object drawing		

- Item:
  - Text vs. Object Drawing
- Text-type:
  - Regular (RT) vs. Scrambled (ST)
- Background:
  - Homogeneous (HB) vs. Inhomogeneous (IB)



- Texts received more attention than drawings, suggesting that the specific visual features of texts cause their attractiveness advantage.
- No statistical differences between words and scrambled words.
- Features of texts are operating at low level.

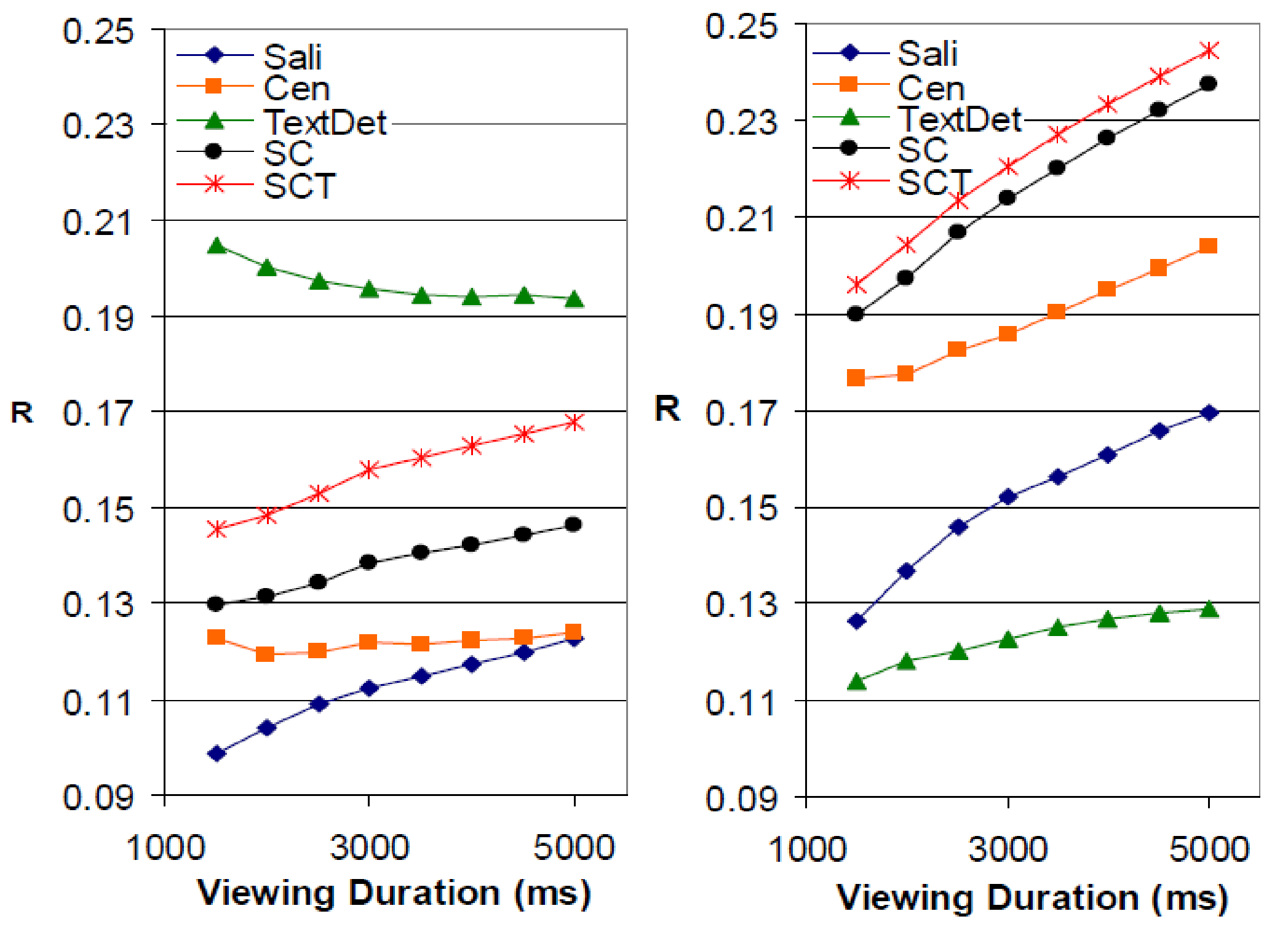


(a) (b) (c) (d) (e)

	Sali	Cen	TextDet	SC	SCT
R -All	0.14	0.16	0.15	0.18	0.20
Text-Present	0.11	0.12	0.20	0.14	0.16
HB	0.09	0.10	0.24	0.10	0.12
IB	0.14	0.15	0.15	0.17	0.19
Text-Absent	0.15	0.19	0.12	0.21	0.22
ROC - All	0.65	0.63	0.63	0.69	0.72
Text-Present	0.61	0.61	0.66	0.64	0.70
HB	0.55	0.60	0.67	0.58	0.67
IB	0.67	0.62	0.64	0.70	0.72
Text-Absent	0.67	0.64	0.62	0.72	0.73

(a) stimulus image
(b) attention (3-second viewing)
(c) Saliency (Sali)
(d) center-bias (Cen)
(e) text-detector maps (TextDet)

- SC (Sali + Cen)
- SCT (Sali + Cen + TextDet)



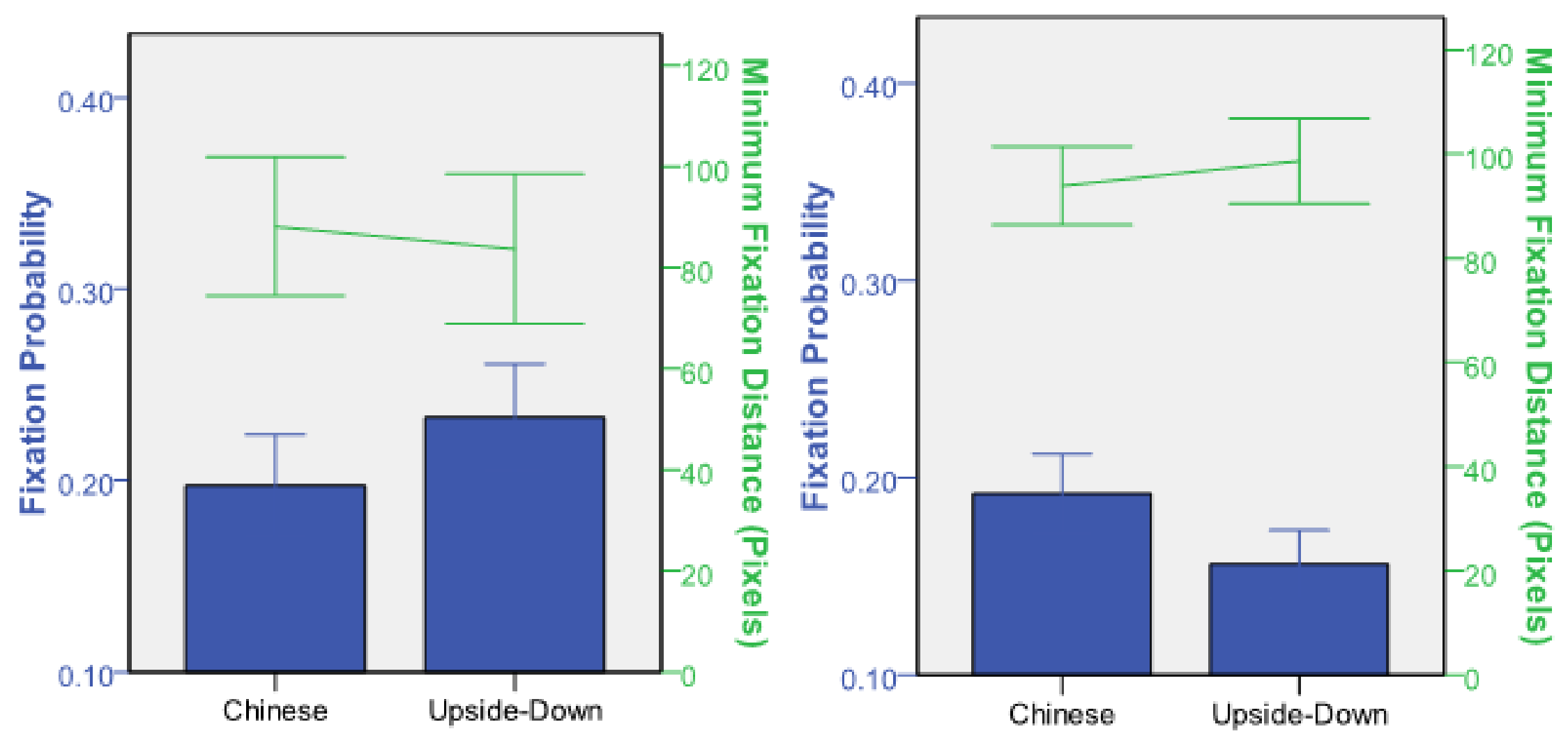
(a) (b)

- Text detector improved the prediction of viewers' visual attention.
- SCT obtained higher R and ROC than SC even in text-absent scenes.
  - Text-like features (e.g., edge density) catch attention.
- HB images obtained higher values than IB images
- Texts are detected early, and later viewers tended to be guided more strongly by saliency

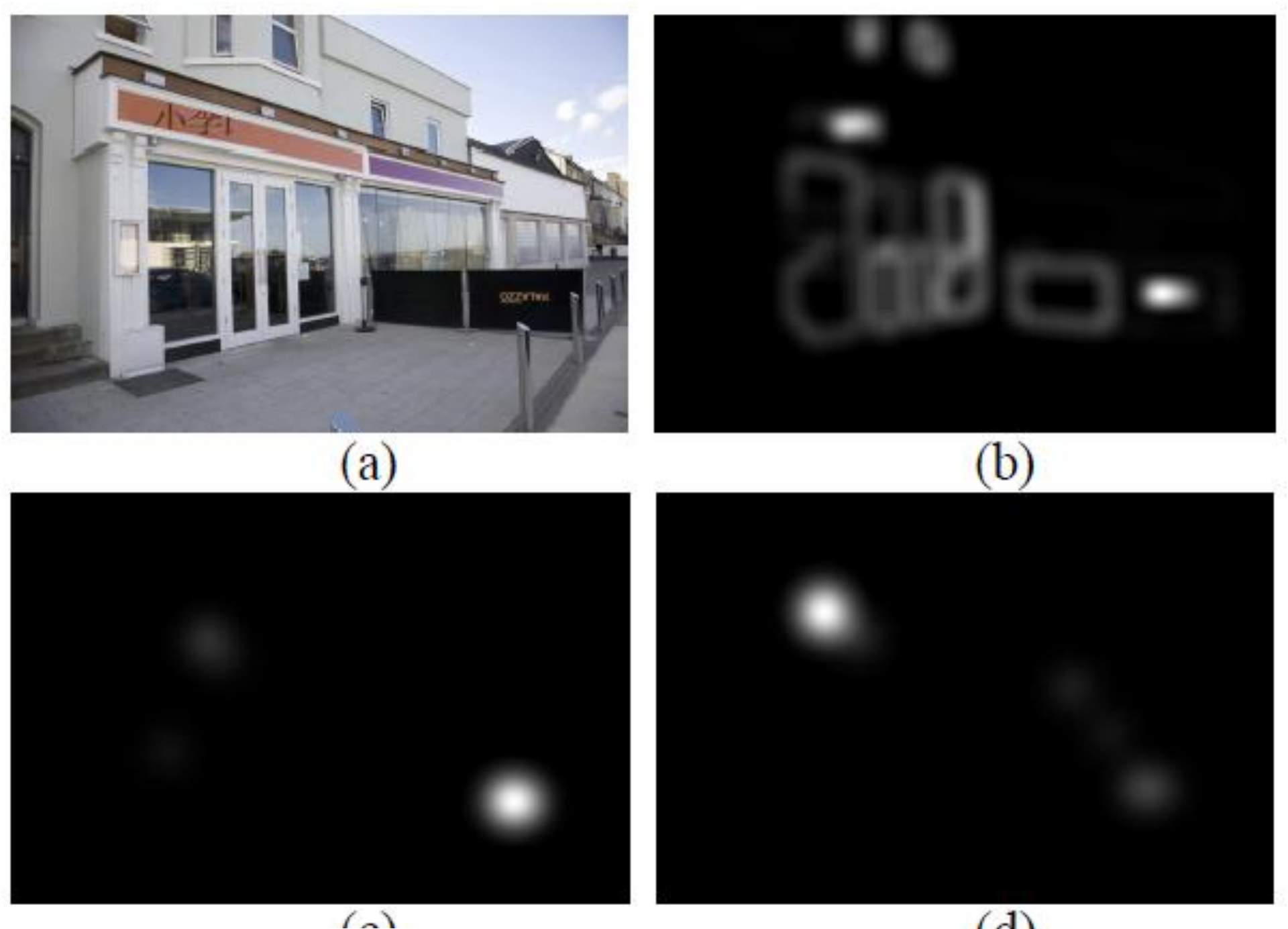
Data Set 2: English vs. Chinese Texts and Native Speakers

(a) (b)

- Texts were either rotated to upside-down or replaced by Chinese texts.
- The stimuli were presented to non-Chinese English speakers and Chinese speakers.

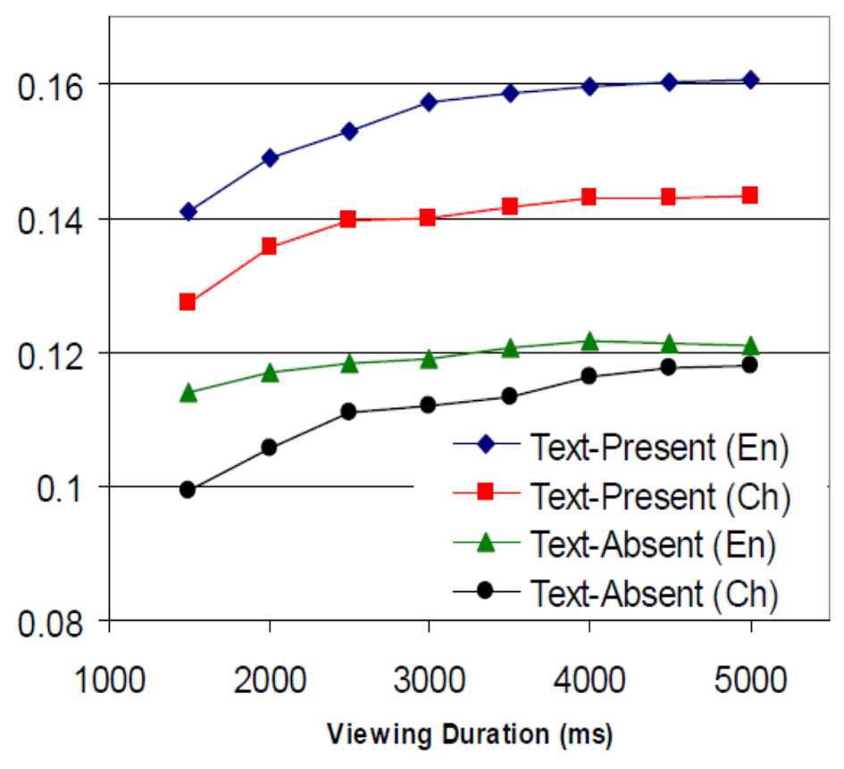


- Text attraction depends on the observer's familiarity with the writing system.
- The results may support the hypothesis that viewers have developed a "text detector" because they are exposed to texts every day and become sensitive to text patterns.



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

(a) stimulus image
(b) text detector map
(c) Attentional map of an English-speaking viewer
(d) Attentional map of a Chinese-speaking viewer



	Sali	Cen	TextDet	SC	SCT
R (En)	0.17	0.17	0.14	0.20	0.21
Text-Present	0.15	0.16	0.16	0.19	0.21
Text-Absent	0.18	0.17	0.12	0.21	0.22
R (Ch)	0.17	0.16	0.12	0.19	0.20
Text-Present	0.15	0.15	0.14	0.18	0.19
Text-Absent	0.18	0.17	0.11	0.20	0.21
ROC (En)	0.69	0.61	0.60	0.72	0.73
Text-Present	0.68	0.62	0.63	0.71	0.73
Text-Absent	0.69	0.61	0.59	0.72	0.73
ROC (Ch)	0.68	0.60	0.60	0.70	0.71
Text-Present	0.67	0.61	0.62	0.69	0.71
Text-Absent	0.68	0.60	0.58	0.70	0.70

- English (En) vs. Chinese (Ch)
  - Common: edge width
  - Ch: more vertical, horizontal, and diagonal strokes
  - En: more curves ("O" or "G").
- Again, text detector improved the prediction of viewers' visual attention, even in text-absent images
- Text detector map predicted English-speaking viewers' attention better than Chinese-speaking viewers' attention

Discussion and Conclusions

- Adding a text detector to an attention model improved its prediction of viewers' visual attention, even in text-absent images.
- Text detector designed for English texts predicted English-speaking viewers' attention better than Chinese-speaking viewers', supporting the hypothesis that viewers have developed a "biological text detector" that is sensitive to text patterns they are exposed to every day and familiar with.
- Further research needs to identify the visual features that underlie this effect. This could be achieved by using text detection algorithms for different writing systems and test their individual components.
- Human viewers can easily locate texts in natural scenes, performing clearly better than current text-detection techniques even when the texts are degraded by noise, rotated, or distorted. Consequently, the results of this line of research are potentially important for developing more efficient and general text detection algorithms.