Published in final edited form as:

Br J Psychol. 2010 February ; 101(Pt 1): 33–46. doi:10.1348/000712609X474730.

Do Image Descriptions Underlie Word Recognition in Reading?

Gordon E. Legge¹, Bosco S. Tjan², Susana T.L. Chung³, and Charles A. Bigelow⁴

¹ University of Minnesota, Twin Cities

² University of Southern California, Los Angeles

³ University of California, Berkeley

⁴ Rochester Institute of Technology, Rochester

A major goal of theories of human pattern recognition is to trace the transformations from a retinal image of an object (letter, word, face or other 3D form) to a perceptual decision. The transformations can be described in terms of the resulting intermediate representations. A theory may choose to address one or more of these transformations and the corresponding representations. For the theory to be successful, the properties of the representation(s) must help us understand visual behavior or brain activity. Watt and Dakin have presented an impressive proposal for how stimulus data are encoded by biologically plausible filters (receptive fields), how the filter outputs are segmented into groups by a computationally straightforward algorithm, and how the primitive features in the resulting groups may be categorized for purposes of recognition. The authors apply their general approach to the important and interesting case of word recognition in text. In particular, by selection of an appropriate range of threshold values in the segmentation process, and an appropriate size (scale) for the filters, the authors show how the model accounts for interesting properties of text spacing (between letters, words and lines) and also provides unique representations for words.

We will comment on two aspects of Watt and Dakin's proposal below. First, the general architecture of the model is of substantial interest. The "Utility of image descriptions" in the title refers to the proposal that an early stage of symbolic representation is sufficient for pattern recognition. Second, the specific application to text is also important. We are not aware of any other computational theory that addresses spacing effects in text. While the practical problem of letter and word recognition has been solved computationally in a variety of high-performance optical-character-recognition (OCR) software applications¹, the features underlying human letter and word recognition remain controversial.

Architecture of the Model

In the Watt and Dakin model, visual encoding begins with response functions (sub-images) derived from linear Gabor filters, each narrowly tuned to a particular orientation, spatial frequency (scale), and phase. These filters are commonly assumed to be an idealized linear model for the simple cells found in the primary visual cortex V1. Watt and Dakin observed that the response of any particular Gabor filter to an image, when thresholded, often forms spatially contiguous regions. They modeled these contiguous regions as thin rectangles and describe them in terms of their position, length, orientation, and mass (which along with the

Corresponding Author: Gordon E. Legge, Department of Psychology, University of Minnesota, 75 East River Rd. Minneapolis, MN, 55455, U.S.A. (612) 625-0846, legge@umn.edu.

¹Most OCR programs achieve their high accuracy with the aid of extensive dictionaries of words in the target language. Accuracy drops substantially without this type of lexical support.

length of a region gives its width). Implicitly, the tuning properties of the filter that generated the regions are also part of the description. The key hypothesis of Watt and Dakin is that subsequent visual processing is based on these descriptions of the rectangular regions. To show feasibility, Watt and Dakin applied this model to printed text and demonstrated that (1) such *Image Descriptions* contain enough information to identify the words, and (2) by simply grouping the overlapped regions generated with different Gabor filters, individual words are usually segmented correctly from each other in a page of text, provided that inter-word spacing and inter-line spacing have values typical of typeset text.

The hypothesis that Image Descriptions are constructed from the output of low-level filters to form the basis for subsequent visual processing is motivated by contemporary knowledge of the response properties of neurons in the visual cortex. Most neurons in the primary visual cortex have spatially localized receptive fields and are selective for orientation. Subpopulations of both simple and complex cells also have the end-stopped property (Hubel & Wiesel, 1962, 1965, 1968; Schiller, Finlay & Volman, 1976), which can be thought of as coding feature length and width. It seems that the primary visual cortex indeed possesses all the essential ingredients to construct Image Descriptions. The construction of Image Descriptions requires local pooling between neurons tuned to the same orientation, spatial frequency and phase. There is indeed evidence for pooling between neurons with receptive fields displaced in the direction of their preferred orientation in the primary visual cortex (Fitzpatrick, 1996). It is also evident that neurons in V1 tend to cluster by their spatial frequency preference (Mallik et al., 2008). Unlike pooling across orientation columns, there is currently no anatomical or otherwise conclusive evidence to support pooling across neurons tuned to the same spatial frequency or phase within V1 or between V1 and V2. Although there is electrophysiological evidence for facilitatory interactions between neurons tuned to the same spatial frequency and phase in the primary visual cortex (Maffei & Fiorentini, 1976), these interactions could be due to feedback from a far-away origin instead of lateral connections or feed-forward projections. It is also known that simple cells in V1 are tuned to a variety of spatial-frequency and phase combinations (Mechler, Reich, & Victor, 2002). To sum up, we do not yet have any clear evidence for a neural representation of Image Descriptions in V1 or other low-level visual areas, although the building blocks for such a representation may exist.

Without confirmatory physiological evidence, the value of a representational theory may be judged in terms of its theoretical utility and ability to account for empirical findings of visual behavior. Proposals for specific forms of visual representation are typically controversial. This may be because the visual system consists of a series of processing stages; each stage can be considered as producing a representation of the input at some level of abstraction. Although a common view is that a perceptual decision is made at the end of this chain of processing, it is likely that a perceptual decision can be made anywhere along the processing chain. Such a system effectively has the equivalence of a continuum of representations, each constructed from a weighted use of the discrete set of representations that are actually present in the visual system (Tjan, 2001). The representations in this continuum range from the highly image-specific to the very abstract.

A theoretical reason for preferring a representational continuum over a representational endpoint is that the amount of details and the level of invariance that are required of a representation depend on the task and stimuli (Tjan & Legge, 1998). The flexibility of the visual system in choosing an appropriate representation for a given task makes it difficult to determine empirically if a particular representation is used by the visual system.

A testable theory of representation has to go beyond merely specifying a particular representational scheme. It should also describe how the proposed representational scheme is

derived from other lower-level representations of the input and how higher-level representations can be derived from it.

Watt and Dakin have been explicit about how to construct Image Descriptions in terms of rectangular regions but they have not specified how in general these region descriptors are turned into high-level representations beyond a simple scheme of categorization. Watt and Dakin have shown that a set of categorized region descriptors form unique codes for words, but this demonstration does not mean that the representation contains the appropriate level of invariance for general-purpose word recognition. For example, a change in font, print style, or letter spacing that does not affect the word identity can lead to different Image Description codes, suggesting that Image Descriptions may not be the representation that directly supports human word recognition.

Watt and Dakin did, however, focus on how Image Descriptions support grouping a prerequisite for building high-level representations. Their rule for grouping was pleasingly simple: if regions overlap, they group. They demonstrated the viability of this grouping rule with printed text.

Application to Text

Watt and Dakin contend that spacing in commonly used typefaces has evolved to achieve an optimal tradeoff between legibility (presumably including the demands of interline and interword spacing) and economy of print space. The design and arrangement of newspaper types provides evidence on this hypothesis. In support of the contention, half a century ago, U.S. newspapers had converged on a rather homogeneous format. Half of US newspapers had adopted a single type design, Corona, and many of the others used typefaces similar to it, and the text columns of many newspapers were roughly the same width and contained roughly the same number of characters, and newsprint cost roughly the same for most of the papers. But contrary to the contention, there has recently been a divergence of formats; in news publishing there are now dozens of different newspaper typefaces and layouts.

If Watt and Dakin are correct in their hypothesis about the optimality of spacing in common typefaces, the optimal values may be rather broadly tuned, given the substantial variation in inter-word and inter-line spacing across legible fonts. For example, for 12-pt Times Roman and Courier, with nearly identical x-size, the inter-word spacing for Courier is about twice the inter-word spacing for unjustified Times Roman².

Even within a given font, inter-word spacing can vary substantially to meet the needs of right justification and other typographical design decisions. Technical factors in the history of typography have also had a major influence on spacing, ranging from standard subdivisions of the em-square used in traditional metal typography to algorithmic computation of word spaces for text justified by word-processing or page-layout software. Similarly, interline spacing is influenced by a variety of design factors considered by typographers, including line length, characters per line, typeface x-height, average character width, and the purpose of the text. Just as the common QWERTY keyboard layout was a compromise between function and technical design considerations, and not necessarily optimized for rapid or accurate typing, spacing in typography has been influenced by a variety of factors in addition to visual function.

²Watt & Dakin appear to measure word spacing from the right-hand black edge of the rightmost stem of the last letter in the preceding word, to the left-hand black edge of the leftmost stem of the first letter in the succeeding word. Adopting this definition, and based on calculations by one of us (CB), the word space in unjustified Times Roman text is 1.2 times the n-size. For Courier, the word space ranges from about 1.9 to 3.0 the n-size.

The above cautionary comments aside, can the model generate novel predictions regarding characteristics of printed text? More detailed specification of the properties of the model's four types of filters and the grouping rule might be necessary. Here is one qualitative prediction that seems to follow from the conceptual design. Selection of the threshold level of filter response is made so that letters within a word will group together, i.e., their region descriptions overlap, but image Descriptions for individual words will be segmented from one another because of the inter-word spacing. We would therefore expect that if inter-letter spacing within words increases, the demands of grouping would require a corresponding increase in inter-word spacing. We consider again the case of Times Roman and Courier, with nearly identical x-heights for a 12-pt font. Interletter spacing is greater for Courier (10 letters per inch) than for Times Roman (which averages about 13 characters per inch). As predicted the inter-word spacing for Courier is larger than the inter-word spacing for Times Roman, but by a factor of about two, rather than 1.3 times as might be expected from the proportional difference in inter-letter spacing. Of course, many factors influence this comparison, especially the fixed character width of the mono-spaced Courier font.

An intriguing property of Image Descriptions in printed text is that there is no need for segmentation of letters within words or representation of individual letters in the coding scheme. This raises the long-debated issue of whether letter recognition precedes word recognition, or whether the basic perceptual representation of words relies on some other form of representation such as the proposed Image Descriptions. Rather than attempting to review this debate, we briefly mention three examples of the effects of text manipulation on reading that might be challenging to explain with the image-description representation of words.

First, we comment on the effect of removing inter-word spaces on reading speed. Subjects are capable of impressive reading speeds (often greater than 100 words/minute) when spaces between words are removed, although they slow down by 30 to 50% compared with regularly spaced text (Spragins, Lefton & Fisher, 1976; Epelboim, Booth & Steinman, 1994; Rayner, Fischer & Pollatsek, 1998; Sass, Legge & Lee, 2006). Word codes, based on Image Descriptions, would lose their distinctiveness and would not seem to be viable without word-word segmentation and without letter recognition. Perhaps the reader might reset the filter threshold to permit letter-letter segmentation and then adopt a reading strategy based on Image Descriptions for letters. Indeed, Watt and Dakin acknowledge that reading by letters rather than by Image Descriptions of words would be necessary to deal with unfamiliar words.

As an aside, we note that many early alphabetic manuscripts did not make use of spaces between words. From about 400 B.C. to about 400 A.D., classical Greek scribes and early Christian scribes did not use word spaces (Parkes, 1993; Martin, 1994). And yet, they wrote and read enough to lay the foundations for what we think of as western civilization. Of course, we have no measures of reading speed from these early days, and it is likely that the later adoption of spaces between words in text was motivated by enhanced legibility. An historically important example of a text without word spaces is the oldest surviving Bible with a complete New Testament, The Codex Sinaiticus, written around 350 A.D. and now digitized and available on-line at: <http://www.codexsinaiticus.org/en/>

Another text manipulation that might challenge the utility of Image Descriptions is to increase the spacing between letters. Using the fixed-width Courier font, Chung (2002) measured reading speed for sentences presented as a sequence of words one at a time at the same location on a display screen (rapid serial visual presentation method). Double inter-letter spacing was achieved by inserting an empty space between adjacent letters. In central vision, reading speed decreased by only 11 percent when the extra spaces were inserted. It is not clear that filters of an appropriate scale could be found to provide appropriate word-based Image Descriptions for text with double letter spacing while at the same time retaining a unique code for each word.

Yu et al. (2007) attributed the mild decline in reading speed for text with double letter spacing to a reduction in the size of the visual span, the number of letters recognized with high reliability on each fixation.

Third, the model's reliance on Image Descriptions constructed at a particular scale is consistent with the idea in the literature that recognizing letters or reading text of a given print size is based on visual processing within a relatively narrow band of spatial frequencies (Legge, Pelli, Rubin & Schleske, 1985; Solomon & Pelli, 1994; Chung et al. 2002; Majaj et al, 2002), possibly within a single spatial-frequency channel. But the single-channel theory of reading is challenged by findings from Beckmann, Legge & Luebker (1991), summarized by Legge (2007, p. 64–65). In the Beckmann et al study, text was constructed so that some of the letters were filtered through one narrow band of spatial frequency and the other letters were filtered through a separate, non-overlapping band. Wide difference between the bands, with center-to-center band difference greater than 2 octaves, ensured that the low and high bands would stimulate different spatial-frequency channels. Subjects' reading speeds were compared for three types of text: all low-band characters, all high-band characters, and "ransom notes" in which the characters were randomly assigned to be low band or high band on a letter-by-letter basis. For a character size for which pure low-band and pure high-band reading speeds were roughly matched, there was very little decrement in reading speeds for the corresponding ransom notes. If word recognition in text were based on Image Descriptions at a single scale, we would expect that the mixed-frequency ransom-note text, yielding region descriptions at different scales, would be difficult to read. If the Image Descriptions are at multiple scales to match the disparate frequency bands of the letters, word segmentation would fail, because by Watt and Dakin's analysis, the range of scale that supports proper segmentation is only about one octave wide. The findings imply that word recognition and high-speed reading can be supported by sensory signals from widely separated spatial scales, perhaps difficult to account for with Image Descriptions.

Conclusions

Watt and Dakin have proposed an attractively simple model of visual encoding in which low-level sensory representations, called *Image Descriptions*, form the basis for high-level representations including words. The front-end filters within the model are biologically plausible, although the neural basis of the proposed Image Descriptions is as yet unknown.

The application of Image Descriptions to word recognition and text layout is novel, and provides a framework for theoretical analysis of pattern recognition in reading printed text. The complexity of factors determining text layout may belie the authors' contention that properties of spacing in text provide a stringent test of the utility of Image Descriptions. Empirical findings from the literature on the psychophysics of reading provide challenges for the model and also opportunities for its further elaboration.

Acknowledgments

The preparation of this commentary was supported by NIH research grants R01-EY002934 (GEL), R01-EY017707 (BST) and R01-EY012810 (STLC). We thank Judith Hirsch and Anirvan Nandy for insightful discussions on physiology.

References

- Beckmann PJ, Legge GE, Luebker A. Reading: Letters, words and their spatial frequency content. Society for Information Display 1991 Digest of Technical Papers 1991:106–108.
- Chung STL. The effect of letter spacing on reading speed in central and peripheral vision. Investigative Ophthalmology & Visual Science 2002;43:1270–1276. [PubMed: 11923275]

- Chung STL, Legge GE, Tjan BS. Spatial-frequency characteristics of letter identification in central and peripheral vision. *Vision Research* 2002;42:2137–2152. [PubMed: 12207975]
- Epelboim J, Booth JR, Steinman RM. Reading unspaced text: Implication for theories of reading eye movements. *Vision Research* 1994;34:1735–1766. [PubMed: 7941379]
- Fitzpatrick D. The functional organization of local circuits in visual cortex: insights from the study of tree shrew striate cortex. *Cerebral Cortex* 1996;6:329–341. [PubMed: 8670661]
- Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* 1962;160:106–154. [PubMed: 14449617]
- Hubel DH, Wiesel TN. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology* 1965;28:229–289. [PubMed: 14283058]
- Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology* 1968;195:215–243. [PubMed: 4966457]
- Legge, GE. *Psychophysics of reading in normal and low vision*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 2007.
- Legge GE, Pelli DG, Rubin GS, Schleske MM. Psychophysics of reading. I. Normal vision. *Vision Research* 1985;25:239–252. [PubMed: 4013091]
- Majaj NJ, Pelli DG, Kurshan P, Palomares M. The role of spatial frequency channels in letter identification. *Vision Research* 2002;42:1165–1184. [PubMed: 11997055]
- Mallik AK, Husson TR, Zhang JX, Rosenberg A, Issa NP. The organization of spatial frequency maps measured by cortical flavoprotein autofluorescence. *Vision Research* 2008;48:1545–1553.10.1016/j.visres.2008.04.014 [PubMed: 18511098]
- Martin, H-J. *The History and Power of Writing*. Cochran, LG., translator. University of Chicago Press; 1994.
- Maffei L, Fiorentini A. The unresponsive regions of visual cortical receptive fields. *Vision Research* 1976;16:1131–1139. [PubMed: 969225]
- Mechler F, Reich DS, Victor JD. Detection and discrimination of relative spatial phase by V1 neurons. *Journal of Neuroscience* 2002;22:6129–6157. [PubMed: 12122074]
- Parkes, MB. *Pause and Effect: Punctuation in the West*. University of California Press; Berkeley: 1993.
- Rayner K, Fischer MH, Pollatsek A. Unspaced text interferes with both word identification and eye movement control. *Vision Research* 1998;38:1129–1144. [PubMed: 9666972]
- Sass SM, Legge GE, Lee HW. Low-vision reading speed: Influences of linguistic inference and aging. *Optometry & Vision Science* 2006;83:166–177. [PubMed: 16534459]
- Schiller PH, Finlay BL, Volman SF. Quantitative studies of single-cell properties in monkey striate cortex. I. Spatiotemporal organization of receptive fields. *Journal of Neurophysiology* 1976;39:1288–1319. [PubMed: 825621]
- Solomon JA, Pelli DG. The visual filter mediating letter identification. *Nature* 1994;369:395–397. [PubMed: 8196766]
- Spragins AB, Lefton LA, Fisher DF. Eye movements while reading and searching spatially transformed text: A developmental examination. *Memory & Cognition* 1976;4:36–42.
- Tjan BS. Adaptive object representation with hierarchically-distributed memory sites. *Advances in Neural Information Processing Systems* 2001;13:66–72.
- Tjan BS, Legge GE. The viewpoint complexity of an object-recognition task. *Vision Research* 1998;38:2335–2350. [PubMed: 9798003]
- Watt R, Dakin S. The utility of image descriptions in the initial stages of vision: A case study of printed text. *British Journal of Psychology*. 2009 in press.
- Yu D, Cheung SH, Legge GE, Chung STL. Effect of letter spacing on visual span and reading. *Journal of Vision* 2007;7(2):1–10.