https://doi.org/10.2352/ISSN.2470-1173.2017.6.MOBMU-297

A Billion Words to Remember

George Nagy. Professor Emeritus,
Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA

Abstract

Current wearable camera and computer technology opens the way for preservation of every printed, computer mediated and spoken word that an individual has ever seen or heard. Text images acquired autonomously at one frame per second by a 20 megapixel miniature camera and recorded speech, both with GPS tags, can be uploaded and stored permanently on available mobile or desktop devices. After culling redundant images and mosaicking fragments, the text can be transcribed, tagged, indexed and summarized. A combination of already developed methods of information retrieval, web science and cognitive computing will enable selective retrieval of the accumulated information. New issues are engendered by the potential advent of microcosms of personal information at a scale of about 1:1,000,000 of the World Wide Web.

Introduction

Surrounded as we always are by natural and computer-mediated visual and auditory stimuli, much of our information diet is still based on printed text. Aside from newspapers, magazines, books and pamphlets, we browse smartphone, tablet and laptop screens. When we drive or walk we cannot avoid looking at scene text. Far more text passes before our eyes than we can remember or even assimilate. Can we preserve it all for recall at will?

In 1945 Vannevar Bush anticipated that: *The camera hound of the future wears on his forehead a lump a little larger than a walnut. It takes pictures 3 millimeters square, later to be projected or enlarged, which after all involves only a factor of 10 beyond present practice. … .Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified. … The entire material of the Britannica in reduced microfilm form would go on a sheet eight and one-half by eleven inches. …* [1].The technology to accomplish far more than this is now within reach due to the confluence of camera-based OCR, document image analysis, wearable electronics, information retrieval, web science, cognitive computing, and speech recognition.

Vannevar Bush's memex (*memory extender*) was an optical device. Inspired by the progress of digital technology, contemporary researchers are reviving the notion of keeping track of what we read or see [2, 3, 4, 5].

The development of a *Lifetime Reader* requires no technological leaps, yet it is more than assembling off-the-shelf components and software. It needs two building blocks: (1) a wearable sensor assembly that autonomously accumulates visible text and audible speech, and (2) a standard platform (smartphone, tablet or laptop) for selective retrieval and presentation of the collected corpus of text. In the following sections we outline a plausible design, make some relevant back-of-envelope calculations, offer pointers to the prospective component technologies, and raise concomitant ethical issues.

Wearables: camera, mic, and microprocessor

The sensor assembly can be far simpler than virtual reality headsets like Google Glass [6] or Hololens [7] because no display is needed and all of the processing required for retrieval will take place on a standard platform. What is needed is a camera light enough for constant wear, and capable of collecting images at roughly the same rate as a human can without head motion [8]. This translates to a 60° field of view (narrower than that of most current body cameras) and 5K × 4K RGB pixels with autofocus from 25 cm to infinity. Larger targets, like maps and unfolded newspapers, can be mosaicked with software developed for copying a large document on a small scanner [9, 10].

Most wearable cameras are video cams, but several can take 16 Megapixel still pictures [11]. Some have built in GPS, gyroscopic stabilization and microphones. However, behind-the-ear and spectacle-mounted cameras don't yet have quite enough pixels, they are still too heavy for constant wear, and they require frequent recharge [12]. Smallest and lightest are borescope and endoscope cameras, but they are typically integrated with light sources and a fat cable. Because of their application to critical diagnostics, we can expect rapid further increase in their capabilities and decrease in their size.

Because no video is required, the necessary resolution can be attained with a sensor module that is no more burdensome than spectacles or hearing aids. Image acquisition at 1 frame per second (fps) will be fast enough. One second is about the time required by a human to decide whether a text in view is worth reading. It is also enough to recognize an expected street sign or a familiar advertisement. It is, of course, far too short for attentive reading, which may require several minutes per page and thereby affords ample time for consolidating text-image data acquired at 1 fps.

A mic can add useful functionality. Tiny microphones are common in hearing aids and body cameras. Assistive hearing devices and voice recognition software already attempt to differentiate intelligible speech from noise. Speech and text recognition have much in common; OCR with a speech recognition toolkit was demonstrated in [13]. Aside from recording personal conversations and speech on radio, television and computers, spoken input could let the user provide brief optional annotations of the text input.

In principle, video continuously collected over a whole lifetime can also be retained, as famously suggested in [14]. Much current research addresses tasks like face, scene and action recognition and health and safety monitoring from personal video. These endeavors raise, however, entirely different technical and ethical issues than textual information. We consider here only printed, rendered, and spoken text. Nevertheless, temporal and spatial tagging of each image, as is common in wearable cameras, would be a definite advantage. This could be contributed either by location hardware integrated into the sensor assembly, or by a wireless link to the GPS on some other mobile device worn or carried by the user.

The microprocessor integrated with the sensor should identify when there is text in the field of view and compress and temporarily store the text images. Fast image and video compression algorithms are available, but speed is not essential because the compression need not be done in real time. Recently acquired images can be filtered and compressed when there is no text in the field of view. Furthermore, the camera can be kept in a low-resolution surveillance mode except when it sees readable text. More selective filtering of redundant images, mosaicking, layout analysis, character recognition, indexing and tagging will be done on the mobile or desktop host platform (or, at the user's choice, in a cloud). In contrast, most current research on camera-based OCR addresses real-time output [15], as required, for example, for translating posted signs [16].

Memory considerations

How much wearable storage is necessary? On clean text images, a compression ratio of 40:1 is readily achievable with. JBIG-2, DjVu, or newer methods [17]. Even keen readers will have text in view during at most half their waking hours and will often dwell on the same text for several seconds. Therefore an average compression ratio of 100:1 seems conservative. The raw image data rate is 20 × 3 MB per second (the factor of 3 is for RGB pixels). Compressed hundredfold, this is only $20 \times 10^6 \times 3 \times 8$ hours × 3600 seconds / 100 ≅ 17 GB per day, well within the 64 GB capacity of available flash drives.

Quasi-continuous wireless uploading to a mobile platform would require only Bluetooth. Uploading to a stationary platform daily or weekly may need a faster broadband link. Like many implanted medical devices, the sensor module could automatically upload nightly the data collected during the day.

We can estimate the overhead of image vs. text storage and of the relentless pace of image collection. GZIP, Lempel-Ziv or other dictionary-based methods yield a five-fold compression on normal prose [18]. Reading or listening at 300 words per minute for eight hours a day would accumulate only 144,000 words or ~300 KB per day after text compression. (We assume throughout two-byte Unicode character representation even though current estimates of the entropy of English text are below 2 bits per character.) 300 KB per day is ~0.002% of the image storage! Note, however, that merely looking at text pages of 1200 words at one page per second, as opposed to reading it, will raise the 300 word/minute rate by a factor of 240, to ~0.5% of the image volume.

We can expect significant differences in text exposure according to age, education, employment, and perhaps even gender. This increases the difficulty of preparing suitable data for experimentation. A possible start would be a mixture drawn from recent competitions on robust camera and smartphone based reading and the benchmark data sets of the International Association for Pattern Recognition Technical Group on Reading (IAPR TC11).

Host computer

The host computer will cull unreadable and repetitive images that were not filtered out by the camera computer, mosaic some frames, perform layout analysis to determine reading order, and then recognize (OCR) and index the text for eventual retrieval. Its only outputs are a display for minimally formatted text and an audio channel (for example, to listen to passages from a long-ago-read book while driving or exercising). Already available language translation and privacy/security (encryption) features can be added at small cost.

Often-noted differences between scanned and camera captured text are the possibility of severe geometric—affine and perspective – distortion, and contrast variations due to uncontrolled illumination. Although dozens of binarization and skew detection/removal methods are available, the extent of distortion in camera captured text, especially scene text, requires affine-invariant methods similar to those used in computer vision [19]. However, the most important, i.e., purposively read, text will be subject only to modest distortion because most people prefer to read in good light, and tend to keep what they read (hardcopy or display) horizontal and perpendicular to their (and the camera's) line of sight.

Because of the variety and unpredictability of the input stream, such as single and multi-column text, bureaucratic forms, comic books, email, posts on social networks, blogs with advertising pop-ups, and scene text, layout analysis will be more demanding than required for the relatively uniform input streams of commercial and historical document digitization. Some adaptation may be possible because of the relative consistency of individual reading, browsing and travel.

Trainable camera-based character recognition was first demonstrated, letter by letter, with the Mark I perceptron in 1959 [20]. It took over thirty years until camera-captured snippets of printed pages could be OCR'd [21]. Soon thereafter entire page images were rectified and mosaicked [22, 23, 24]. An excellent survey of early work on camera-based text analysis is [25]. A proposed alternative approach matches fragments of documents at the image level for retrieval of the entire document from a database [26, 27, 28].

Although comfortable, ubiquitous and uninterrupted text acquisition and transcription of the collected data does present some new problems, none seem insurmountable. The really difficult puzzle is retrieving vaguely or inaccurately remembered material that one may have browsed in the distant past.

Research on a universal personal filing system began more than 30 years ago [29]. For retrieval of non-annotated material, we could perhaps adopt and adapt browser technologies which are now well beyond simple keyword search. Unlike the web, this collection never has to be re-indexed, because one cannot un-read something. Initially there won't be any PageRank, but some cross-linkages can be automatically constructed using temporal or spatial proximity. The system can, of course, construct a complete and accurate profile of its single user. Ontological tools developed for the semantic web may also play a useful role in personal collections.

Another set of query tools is available from the library side. These started out with *Author* and *Subject* catalogs, bibliographies and concordances, but now incorporate all the tools of information retrieval like pattern matching on compressed text, inverted indices, vector-space models, perfect hashing, signature files, elaborate text tagging, fuzzy clustering, latent semantic indexing, graph algorithms, and relevance feedback.

Three factors facilitate retrieval from a personal collection. The first advantage over web search is that there is not that much data to be indexed compared to the World Wide Web. Even if one started in grade school and lived to be a hundred, the final volume would be only 300 KB of text per day × 365 days × 100 years ≅ 10 GB. This is much less than one millionth of the estimated size of World Wide Web even if we include text only seen but not read. The second advantage is that the list of top-ranking items displayed in response to a query will already seem familiar, so we can parse it quickly to find the page, passage or phrase that we sought. (This is why some of us hang on to our obsolete but well-thumbed textbooks.) Finally, we won't be bothered by OCR errors because we are all used to

fractured and misspelled prose and because we are unlikely to disseminate *verbatim* what we retrieve.

Underlying research problems

The technology—miniature high-resolution camera and adequate and affordable computing and storage capacity—is almost here. It would seem far simpler than what has already been demonstrated for self-driving cars and autonomous drones. The natural language processing aspects are well within the state of the art. Nevertheless a number of interesting and interdependent algorithmic problems require further research, and some ethical issues require thoughtful consideration. While each of the items listed below could be the subject of a full paper, here we can only hope to call attention to them.

Image acquisition

- Text detection in spatial context, at home, at work, in local venues, in transit, abroad
- Mosaicking required by head and body motion
- Lazy compression of sparse-text images
- Long-lasting or self-charging power supply
- Optional Hands-free (via mic) annotation
- Optional visible (gestural) annotation, e.g. by tracing a phrase on a printed page or computer screen with a designated finger

Text-image analysis

- Perspective-invariant recognition instead of rectification
- Reading-order (without gaze tracking)
- Duplicate detection from consecutive frames and after (possibly lengthy) interruptions
- Retention policy for undecipherable and unindexable fragments of text, and for near-duplicates
- Adaptation to predictable reading material like the daily newspaper, magazines, the remaining volumes of the Jack Aubrey series, IJDAR, Python v2.7.6 documentation

Information retrieval

Retrieval strategies that mesh with our own mental recall

- Personalization: scripts and languages— reading speed—reading postures—computer display settings—work, leisure, shopping and napping habits
- Selective, topic-, time-, or location-specific summarization
- Logging queries, responses, and user reactions for improving the system even as one's own memory deteriorates

Ethical and legal issues

- Security and privacy: what do these mean over a lifetime?
- What is the legal difference between deliberately acquired information, as with a smartphone or camera, and autonomously acquired information?
- Where must the owner of a Lifetime Reader not look (and record)? [30]
- What responsibility does delayed discovery of a crime entail (for instance, reading an airplane seat neighbor's laptop screen that one glanced at two years ago)?
- What are the social and marketing implications of lifetime text logging? [31]

Disclosure

The author discussed some of the above ideas in the final section of an invited historical review of interactive document image analysis. A draft submitted over a year ago remains in the journal editors' hands without a final title or publication schedule. Because of the timeliness of the topic, the notion of an autonomous lifetime reader is revised and expanded here with additional references.

Author biography

George Nagy received his BEng (1959) and MEng (1960) from McGill and his PhD in electrical engineering from Cornell (1962). He worked at IBM T.J. Watson Research Center until 1972, at UNL until 1985, and at RPI since then. He conducted research on statistical pattern recognition, OCR, document image analysis, green interaction, remote sensing and computational geometry.

References

[1] Vannevar Bush, "As We May Think," The Atlantic, July 1945.

[2] H. Fujisawa, H. Sako, Y. Okada, and S-W. Lee, "Information Capturing Camera and Developmental Issues," Proc. Int. Conf. Document Analysis and Recognition (ICDAR), pp. 205–208, 1999.

[3] T. Kimura, R. Huang, S. Uchida, M. Iwamura, S. Omachi, K. Kise: "The Reading-Life Log — Technologies to Recognize Texts That We Read," Proc. Int. Conf. Document Analysis and Recognition (ICDAR) pp. 91–95, 2013.

[4] K. Kunze, K. Masai, M. Inami, Ö. Sacakli, M. Liwicki, A. Dengel, S. Ishimaru, K. Kise: "Quantifying reading habits: counting how many words you read," Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015, Osaka, Japan, pp. 87–96, 2015.

[5] M. Matsubara, J. Folz, T. Toyama, M. Liwicki, A. Dengel, K. Kise: "Extraction of read text using a wearable eye tracker for automatic video annotation," Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers, UbiComp/ISWC Adjunct 2015, Osaka, Japan, pp. 849–854, 2015.

[6] T. Toyama, A. Dengel, W. Suzuki, K. Kise, "Wearable Reading Assist System: Augmented Reality Document Combining Document Retrieval and Eye Tracking," Proc. Int. Conf. Document Analysis and Recognition (ICDAR), 2013 pp. 30–34, 2013.

[7] E.E. Sabelman, R. Lam, "The real-life dangers of augmented reality," IEEE Spectrum 52, 7, pp, 48–53, July 2015.

[8] A. Betancourt, P. Morerio, C. S. Regazzoni; M. Rauterberg, "The Evolution of First Person Vision Methods: A Survey," IEEE Transactions on Circuits and Systems for Video Technology, 25, 5, 2015.

[9] J. Cullen and J.J. Hull, ''Oversize document copying system,'' IAPR Workshop on Document Analysis Systems, Malvern, PA, 1996.

[10] 10 M. Pilu and B. Pollard, "Method and apparatus for scanning oversized documents," Patent US 6975434 B1, Dec 13, 2005.

[11] 2016 Best Wearable Cameras Reviews: http://wearable-cameras-review.toptenreviews.com/, 2016.

[12] I. Frazer, "Got a Bad Memory? This Company Has You Covered," Wall St Daily, Aug. 29, 2015.

[13] P. Schone, A.Cannaday, S.Stewart, R; Day, J. Schone, “Automatic Transcription of Historical Newsprint by Leveraging the Kaldi Speech Recognition Toolkit,” Proc. Document Recognition and Retrieval, IS&T Electronic Imaging, 1-10(10), February 2016.

[14] J. Gemmell, R. Lueder and G. Bell, "The MyLifeBits lifetime store" in Transactions on Multimedia Computing, Communications and Applications, ACM , pp. 1–5, 2002.

[15] T. Kobayashi, M. Iwamura, T. Matsuda, K. Kise, “An Anytime Algorithm for Camera-Based Character Recognition,” Proc. Int. Conf. Document Analysis and Recognition (ICDAR), pp. 1140–1144, 2013.

[16] M. Mantha and J K. Chaithanya, “Vision based Traffic Panel Text Information and Sign Retrieval,” Int. J. Current Engineering and Technology, 5, 4, Aug 2015.

[17] E. Ageennko, P. Franti, “Lossless compression of large binary images in digital spatial libraries,” Computers & Graphics. 24, 1, pp. 91–98, 2000.

[18] N. J. Larsson, A. Moffat, “Off-line dictionary-based compression,” Proceedings of the IEEE, 88, 11, pp. 1722–1732, Nov. 2000.

[19] T. Nakai, K. Kise, M. Iwamura, “Camera-based document image retrieval as voting for partial signatures of projective invariants,” Proc. Int. Conf. Document Analysis and Recognition (ICDAR), pp. 379–383, 2005.

[20] G. Nagy, "Neural Networks - Then and Now," IEEE Transactions on Neural Networks, 2, 2, pp. 316–318, March 1991.

[21] M. Koga, R. Mine, T, Kameyama, T. Takahashi, M. Yamazaki, T. Yamaguchi, "Camera-based Kanji OCR for mobile-phones: practical issues," Proc. Int. Conf. Document Analysis and Recognition (ICDAR), 2, 29, pp. 635–639, 2005.

[22] T Kasar and AG Ramakrishnan, “CCD: Connected component descriptor for robust mosaicing of camera-captured document images,” IAPR Workshop on Document Analysis Systems, pp. 480–486, 2008.

[23] J. Liang, D. DeMenthon, D. Doermann, “Geometric Rectification of Camera-Captured Document Images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 30, 4, pp. 591–605, April 2008.

[24] J. Liang, D. DeMenthon, D. Doermann, “Mosaicing of camera-captured document images,” Computer Vision and Image Understanding,” 113, 4, pp. 572–579, April 2009.

[25] J. Liang, D. Doermann, H. Li, “Camera-based analysis of text and documents: a survey,” International Journal of Document Analysis and Recognition 7, 2, pp. 84–104, 2005.

[26] J. Moraleda and J.J. Hull, “Toward Massive Scalability in Image Matching,” IAPR Int. Conf. on Pattern Recognition (ICPR), Istanbul, Turkey, pp. 3424-3427, Aug. 23–26, 2010.

[27] S. Ahmed, K. Kise, M. Iwamura, M. Liwicki, A. Dengel, “Automatic Ground Truth Generation of Camera Captured Documents Using Document Image Retrieval.” Proc. Int. Conf. Document Analysis and Recognition (ICDAR), pp. 528–532, 2013.

[28] T. Nakai, K. Kise, M. Iwamura, “Camera-based document image retrieval as voting for partial signatures of projective invariants,” Proc. Int. Conf. Document Analysis and Recognition (ICDAR), pp. 379–383, 2005.

[29] H. Fujisawa, A. Hatakeyama, and J. Higashino, “A Personal Universal Filing System Based on the Concept-Relation Model,” Proc. 1st Int. Conf. Expert Database Systems, Charleston, SC, pp. 31–44, 1986.

[30] T.M.Mok, F. Cornish, J. Tarr, “Too Much Information: Visual Research Ethics in the Age of Wearable Cameras,” Integrative Psychological and Behavioral Science, 49, 2, pp 309–322, June 2015.

[31] A. R. Doherty, N. Caprani, C. Ó Conaire, V. Kalnikaite, C. Gurrin, A. F. Smeaton, N. E. O’Connor, “Passively recognising human activities through lifelogging,” Computers in Human Behavior 27, 5, pp. 1948–1958, September 2011.