# Towards A Real-Time System for Finding and Reading Signs for Visually Impaired Users

Huiying Shen and James M. Coughlan

The Smith-Kettlewell Eye Research Institute, San Francisco, CA
hshen@ski.org, coughlan@ski.org

**Abstract.** Printed text is a ubiquitous form of information that is inaccessible to many blind and visually impaired people unless it is represented in a non-visual form such as Braille. OCR (optical character recognition) systems have been used by blind and visually impaired persons for some time to read documents such as books and bills; recently this technology has been packaged in a portable device, such as the smartphone-based kReader Mobile (from K–NFB Reading Technology, Inc.), which allows the user to photograph a document such as a restaurant menu and hear the text read aloud. However, while this kind of OCR system is useful for reading documents at close range (which may still require the user to take a few photographs, waiting a few seconds each time to hear the results, to take one that is correctly centered), it is not intended for signs. (Indeed, the KNFB manual, see knfbreader.com/upgrades_mobile.php , lists "posted signs such as signs on transit vehicles and signs in shop windows" in the "What the Reader Cannot Do" subsection.) Signs provide valuable location-specific information that is useful for wayfinding, but are usually viewed from a distance and are difficult or impossible to find without adequate vision and rapid feedback.

We describe a prototype smartphone system that finds printed text in cluttered scenes, segments out the text from video images acquired by the smartphone for processing by OCR, and reads aloud the text read by OCR using TTS (text-to-speech). Our system detects and reads aloud text from video images, and thereby provides *real-time feedback* (in contrast with systems such as the kReader Mobile) that helps the user find text with minimal prior knowledge about its location. We have designed a novel audio-tactile user interface that helps the user hold the smartphone level and assists him/her with locating any text of interest and approaching it, if necessary, for a clearer image. Preliminary experiments with two blind users demonstrate the feasibility of the approach, which represents the first real-time sign reading system we are aware of that has been expressly designed for blind and visually impaired users.

**Keywords:** visual impairment, blindness, assistive technology, OCR, smartphone, informational signs

## 1 Introduction and Related Work

OCR is designed to process images that consist almost entirely of text, with very little non-text clutter, such as would be obtained from a picture (e.g., acquired by a flat-bed image scanner) of a single page of a book. A growing body of research [2] has focused on the complementary problem of finding text in cluttered images, such as are encountered by a person searching for a sign, so that the text can be isolated in each image in order to be processed effectively by OCR. Some research [4] has specifically tackled the added challenge of finding and reading text on a portable device, and smartphone apps such as Word Lens (http://questvisual.com/) have been developed, which are able to find and read scene text at several video frames per second, but are intended for use by people with normal vision.

A comparatively small amount of work has addressed the specific problem of finding and reading signs or other non-document text for blind or visually impaired people. Yi and Tian [6] have focused on computer vision algorithms for finding text in complex backgrounds (e.g., found in typical indoor and outdoor urban scenes), training their algorithms on an image dataset collected by ten blind users, but have not yet addressed the formidable user interface issues posed by a full system that helps a visually impaired user find text and have it read aloud to him/her. The "Smart Telescope" SBIR project from Blindsight Corporation (www.blindsight.com) is a novel system to help a person with low vision find and read text by automatically detecting text regions in a scene acquired by a wearable camera and presenting the regions one at a time to the user, using a head-mounted display that zooms into the text to enable him/her to read it. Finally, [3] reports studies with three blind users of a real-time computer vision-based smartphone system for locating special "color marker" signs, describing the strategies employed by the users to find each marker, walk towards it and touch it. While color markers are specially designed for ease of detection by the system, and are therefore much easier to find and read than the kinds of text signs considered in our application, the search strategies adopted by the users underscore the challenges of finding any kind of sign with a camera-based system.

## 2 Finding Text in Images and Performing OCR

The foundation of our prototype system is a processing pipeline that includes a computer vision algorithm for finding text in images, followed by a standard OCR package run on the text regions identified by this algorithm.

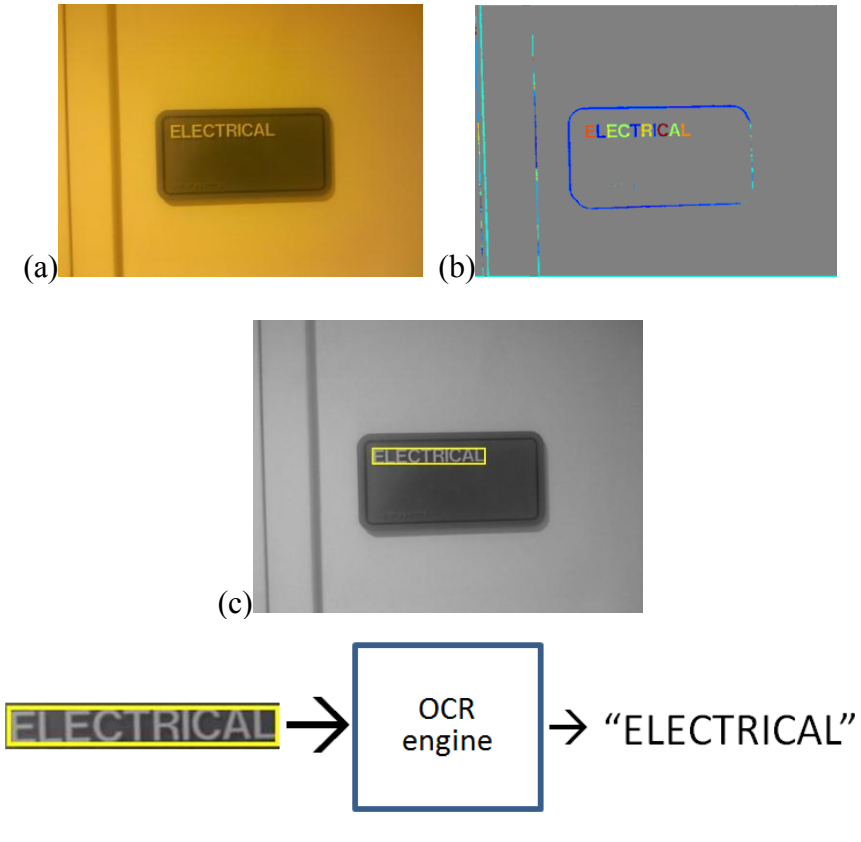


**Fig. 1.** Main stages of text detection and recognition. (a) Sample indoor image taken by smartphone. (b) “Blob” regions detected in image (each blob is given a separate, random color for visibility). (c) Detected text region drawn as a “text box” (in yellow). (d) When this text region is input to OCR it is correctly read as “ELECTRICAL.”

The text detection algorithm, which builds on previous work by the authors [5], processes a video frame (which has 640x480 resolution, see Fig. 1a) and converts it to grayscale for subsequent processing. “Blob”-like structures in the image (Fig. 1b) are detected in the image, one blob typically being extracted for each character of text (in addition to many other blobs corresponding to non-text clutter in the image). Blobs whose shape and/or size are incompatible with that of text characters are removed, and the remaining blobs are searched for groups of consistently sized ones that are aligned in a way that is consistent with a horizontal word or line of text. This procedure is applied to the image at both polarities (for detecting light text on a dark background and vice versa), yielding blob groups that are classified as text groups, which form candidate text regions demarcated by rectangles (Fig. 1c), referred to as “text boxes.”

Each text box forms a cropped portion of the image that is sent to the Tesseract OCR engine (http://code.google.com/p/tesseract-ocr/), an open source OCR package that runs in real time on the smartphone (Fig. 1d). Some OCR output contains errors, either because it results from a false positive text box (i.e., it is reported incorrectly as a text region), or because the text box is valid but OCR is unable to process it correctly. To reduce the number of spurious or incorrect OCR output strings to communicate to the user, we apply a simple filtering procedure to discard strings with unlikely characters or character combinations.

## 3 System and User Interface

Our software was programmed in C++ and implemented on an LG-P990 Android smartphone processing video frames using the smartphone's camera. After processing each video image frame as described above, we read aloud each text string using TTS. If more than one text string is detected in an image, the text strings are read aloud in the following order: from the top of the image to the bottom of the image, and from left to right among strings that are at roughly the same height in the image.

Depending on the complexity of the images and amount of text contained in them, the processing proceeds at a rate as high as one or two frames per second (for simpler images with small amounts of texture). After experimentation we chose a TTS setting that allows all text to be read aloud, before processing the next frame. The advantage of this setting is that scenes with longer strings of text are less likely to be cut off, but at the cost of sometimes delaying the processing of a new frame for a few or more seconds.

The philosophy behind our user interface is that some errors are inevitable with any OCR system, especially one based on a handheld camera; the simplest way to overcome the errors is for the user to obtain multiple readings of each text sign over time and arrive at a consensus among the readings. Specifically, spurious readings (e.g., due to false positives from background clutter) can be ignored because of their inconsistency over time; minor reading errors (e.g., a few misread characters in a word) can often be "repaired" by waiting for a correct reading (which is more likely to be read consistently, and usually makes more sense to the user in a given context, than an incorrect reading) or inferring the most likely word that gives rise to multiple misreadings.

To improve the basic TTS user interface, we introduced three novel functions. First, we implemented a tilt detection function (similar to that in [1]), using the smartphone accelerometer to sense the direction of gravity, which allows the user to point the camera arbitrarily above or below the horizon and to the left or to the right, but issues a vibration warning if the camera is rotated clockwise or counterclockwise about its line of sight. This maximizes the chances that text appears roughly horizontal in the image (as required for successful detection). Second, any text string that originates

from a text box that is close to the border of the image is read aloud in a low pitch, to warn the user that important text may be cut off at the edge of the image. (For instance, a "No smoking permitted" sign may be detected as "smoking permitted" if the first word falls outside of the image.) Finally, any text string corresponding to text that is sufficiently small in the image is read aloud in a high pitch, which warns the user that such text may be incorrectly recognized (and that the user should approach closer if possible to get a more reliable reading).

## 4 Experimental Results and User Testing

We explained the purpose and operation of the system to two completely blind volunteer subjects. Particular emphasis was placed on the importance of moving the camera slowly to avoid motion blur, ensuring the camera lens was not covered (e.g., by the user's fingers), and thoroughly sweeping the desired target region to accommodate the camera's limited field of view. After a brief training session with a handheld sign, we took the subjects to a conference room in which ten text signs were posted along two adjoining walls. The signs were high contrast (black and white), of varying font, font size and polarity (i.e., dark text on light background or vice versa), and were placed at approximately chest level; they contained the type of text that might be expected in an office building, such as "Room 590" or "Main entrance." The subjects were told to search both walls for an unknown number of signs, standing a few meters away from the signs (i.e., out of reach), and to tell the experimenter the content of each sign detected.


**Fig. 2.** Scene from experiment shows signs posted on wall with blind volunteer holding system.

The first subject took under six minutes to search for the signs, and reported six of them perfectly correctly. Of the remaining four signs, two were completely missed,

the sign labeled “Dr. Samuels” elicited a TTS response of “Samuels” (which was audible to the experimenter but not the subject) and the “Meeting in Session” sign gave rise to the words “Meeting” and “section” (though they were not uttered together). The second subject searched for the signs in about the same length of time, but only reported three of them perfectly, in part because he moved the camera quickly while searching for them. The pattern of errors he encountered among the other seven signs is telling: for instance, the sign labeled “Exam Room 150” was detected and read aloud correctly, but he was unable to understand the word “exam” (perhaps because there was no context to prepare him for it); and he reported “D L Samuels meeting in session” as a sign, which is an incorrect combination of two signs, “Dr. Samuels” (in which the system misread “Dr.”) and “Meeting in Session.” Of the three special user interface functions we devised, the tilt sensor appeared to be most consistently useful to the subjects, while the situations requiring the use of the low/high pitch signals were less common.

While the results show that the system needs to be improved substantially before it becomes practical, the study provides proof of concept of the approach and provides insight into the most important problems to be addressed. First, the main challenge in using the system was finding text in an unknown location, which required the user to patiently scan large areas. Slow processing speeds (especially on images of high-texture regions), combined with motion blur (exacerbated by low lighting conditions where the experiment was conducted), forced the user to scan slowly. False positive text detections created a significant amount of spurious TTS responses, which further slowed down the process. Somewhat surprisingly, even when the system functioned perfectly, the TTS output was not always interpreted correctly by the user. Finally, the simple procedure we used for deciding the order in which to announce multiple text lines was helpful, but did not address the need to announce the contents of each sign separately from the others. We discuss possible solutions to these problems, which we are currently implementing, in the next section.

## 5 Conclusion

We have demonstrated a novel smartphone system to find and read aloud text signs for blind and visually impaired users. A prototype system has been implemented on the Android smartphone, which includes special user interface features to help guide the search for text. We have conducted preliminary experiments with blind volunteers to test the system, demonstrating its feasibility.

We are planning several future improvements and extensions to the system. First and foremost, speed and accuracy improvements to the text detection algorithm will make the system faster and create fewer false positive readings; a faster algorithm may also permit processing of higher resolution video images, which would enable signs to be detected from farther away. The ability to detect text that is poorly resolved (because of small size or motion blur) would also permit text detection in some cases when the

text is not clear enough to be read. A more efficient user interface might then signal the presence of text with a brief audio tone, help the user center and/or approach the text and then have it read aloud. Multiple text lines will be clustered into distinct sign regions, which will help both with centering of signs and intelligibility of the TTS output, and the user will be able to hear the TTS output repeated for any given sign upon request. Eventually we envision a system that analyzes an entire scene as an image panorama (i.e., mosaic), acquired by panning the camera back and forth, which is able to seamlessly read lines of text that extend beyond the borders of any individual image frame.

**Acknowledgments** The authors acknowledge support by the National Institutes of Health from grants No. 1 R21 EY021643-01, 1 R01 EY018210-01A1 and 1 R01 EY018890-01, and by the Department of Education, NIDRR grant number H133E110004. We would like to thank Dr. Ender Tekin and Dr. Vidya Murali for many helpful conversations about the paper.

## 6 References

1. V. Ivanchenko, J. Coughlan and H. Shen. "Real-Time Walk Light Detection with a Mobile Phone." 12th International Conference on Computers Helping People with Special Needs (ICCHP '10). Vienna, Austria. July 2010.
2. J. Liang, D. Doermann and H. Li. "Camera-based analysis of text and documents: a survey." International Journal on Document Analysis and Recognition, 7:83–200, 2005.
3. R. Manduchi, S. Kurniawan and H. Bagherinia. 2010. "Blind Guidance Using Mobile Computer Vision: A Usability Study." ACM SIGACCESS Conference on Computers and Accessibility (ASSETS).
4. M. Pilu and S. Pollard. "A light-weight text image processing method for handheld embedded cameras." British Machine Vision Conference, 2002.
5. P. Sanketi, H. Shen and J. Coughlan. "Localizing Blurry and Low-Resolution Text in Natural Images." 2011 IEEE Workshop on Applications of Computer Vision (WACV 2011). Kona, Hawaii. Jan. 2011.
6. C. Yi and Y. Tian. "Assistive Text Reading from Complex Background for Blind Persons." The 4th International Workshop on Camera-Based Document Analysis and Recognition (CBDAR), 2011.