The Reading-life Log – Technologies to Recognize Texts That We Read

Takashi Kimura*, Rong Huang*, Seiichi Uchida*, Masakazu Iwamura†, Shinichiro Omachi‡ and Koichi Kise†
*Kyushu University, Fukuoka, Japan
†Osaka Prefecture University, Osaka, Japan
‡Tohoku University, Sendai, Japan

***Abstract*—Reading life log is a type of techniques to automatically and unconsciously record people's reading intentions, interests and habits. Besides, it can also serve as various assistants in our daily life. In this paper, a reading-life log system is implemented by a head-mounted and unobtrusive video camera with a high resolution and a high shutter speed. We utilize DP matching, and propose a text-based frame mosaicing method to integrate multiple frames in a clip. The developed system is tested in the various environments indoor and outdoor. The experimental results show that our system can provide reliable outputs with respect to the most correct responses. The infrequent misregistration between lines also indicates the feasibility and validity of the text-based frame mosaicing.**

I. Introduction

Texts expressing the explicit information surround us, and reading is one of the most essential activities in our daily life. For example, we read books, documents, posters, menus, goods packages, web pages, video captions, signboards, etc., for getting various types of knowledge and information from them. As one can imagine, we will be in a tight corner without those texts. On the other hand, nowadays, the ubiquitous mini electronic devices with high resolution and high shutter-speed enable us record the texts appearing in our daily life. Since the text collection is conducted by first-person vision, it is possible to analyze the personal intentions, interests and habits.

This paper develops a practical reading-life log system, which enables us acquire the text information automatically by using OCR technologies. Reading-life log, as the terminology itself implies, is a type of life-log systems, which try to capture various types of information automatically and unconsciously from some sensor. Our reading-life log is specialized for acquiring text information from a head-mounted and unobtrusive video camera. The developed device can work under various realistic environments such as book, magazine, newspaper, signboards in the wild world, LCD (Liquid Crystal Display) screen.

II. Overview of Reading-Life Log

Reading-life log has initiated a lot of interests in the document analysis community due to the additional attention information. Campbell *et al.* [1] aimed at tracking the texts which an user was reading, and further analyzed his/her interests or goals by exploring the eye movement information. Kienzle *et al.* [2] invented a nonparametric visual saliency model based on the human eye movement data. Bulling *et al.* [3] utilized a wearable electrooculography to detect the reading activity under a variety of indoor and outdoor situations. Buscher *et al.* [4] developed a precise user-oriented document classifier by introducing the personalized attention information. Xu *et al.* [5] devised a document summarization algorithm by taking the time consumption estimation of reading each word as a clue. Judd *et al.* [6] considered the top-down image semantics and collected eye tracking data for learning a saliency model. Nyström *et al.* [7] proposed a velocity-based algorithm which could robustly identify fixations, saccades and glissades in the eye tracking data for event detection. Loboda *et al.* [8] suggested using eye movement information to infer the word relevance. Carlos *et al.* [9] designed a mobile head-mounted, and employed the Maximal Stable Extremal Regions (MSERs) algorithm for scene text segmentation. Duggan *et al.* [10] investigated the eye movement data, and verified the fact that readers have the ability to explicitly catch the most important information in which they were interested. Dimigen *et al.* [11] sampled electroencephalogram (EEG) and eye movement signals when subjects were conducting text reading task, and demonstrated the feasibility of the combination of EEG and eye tracking data. Buscher *et al.* [12] revealed the relations between eye movement measures and user-perceived relevance for personal information retrieval application. To distinguish the reading or skimming behavior in the eye tracking data, Biedert *et al.* [13] extracted average forward speed and angularity features to train a classifier. A fresh work submitted by Kai *et al.* [14] attempted to measure brain electrical activity by an off-the-shelf EEG device for distinguishing genres of documents.

Through the above survey, we find that most existing works focus on the eye tracking data while seldom employ the practical recognized data to refine the final results. The proposed reading-life log system fulfills text-based mosaicing with two-dimensional alignment of recognized data at two consecutive frames. A detailed system outline is given in the next section.

III. System Outline

The proposed reading-life log system is simply realized by a head-mounted and unobtrusive video camera connected to a laptop via USB 3.0. Figure 1 shows a realization of the device. This commercial video camera is characterized by a high resolution and high shutter-speed. The high resolution is necessary to sample the target text lines with high quality. For our implementation of the reading-life log system, at least, so-called "HDTV resolution" is necessary to capture a part of a document containing multiple text lines. Moreover, since the subject who equips the video camera may introduce the unwanted motion blur which seriously degrades the subsequent text recognition performance, the high shutter-speed is also necessary to avoid this type of performance attenuation.

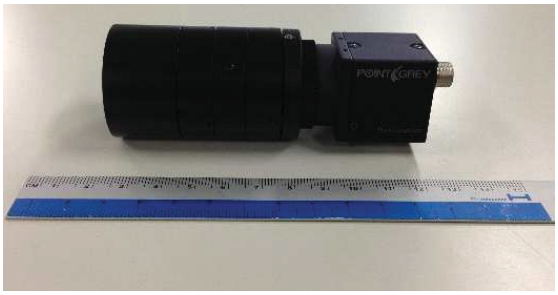


Fig. 1: The appearance of the video camera.

The camera should be equipped aiming at the target text lines, namely the direction in parallel with the sight line, in order to acquire what the user reads. In this sense, our reading-life log implementation coincides with the "first-person vision" systems so that the user activity can be reflected through the video images from a head-mounted camera. There exist numerous research activities on scene text detection and recognition like [15–19], which generally target on all characters appearing in the wild world. Different from them, the proposed system exactly focuses on the specific regions in which an user is interested. The "first-person vision" fashion not only performs practically in the sense of well gearing towards the user's attention, but also eases the subsequent detection and recognition task by providing the prior knowledge of location.

At the early stage of the authors' trial, a glass-type mobile eye-tracker was used for acquiring "gazed" texts. Unfortunately, the trial revealed that even a state-of-the-art eye-tracker cannot locate the gazing point with an enough accuracy. Specifically, it was very difficult to track target text lines along with eye movement. Furthermore, even through we can locate the gazing point perfectly, it is still problematic to exactly understand the "reading" point since gazing point and reading point might be different. Consequently, we resort to the simple head-mounted video camera which can be equipped physically aiming at the target text lines without the need of eye movement estimation. Most existing eye movement based systems, as surveyed above, definitively rely on an additional eye-tracker assembly which renders the appearance of the device obtrusive.

Video frames captured by the head-mounted camera are fed to an Optical Character Recognition (OCR) system prepared in the laptop for extracting text information. Then, the recognized results will be stored in a storage of the laptop. See details in section V. Note that instead of OCR, we can also exploit a document image retrieval technique to get the text information under the condition that we can specify the key target texts. More specifically, instead of directly recognizing the video frames, the document image retrieval technique will search a database to get its exact text information.

IV. Applications

The proposed reading-life log system can serve as a journey assistant to bridge the language barriers. Besides, the visually handicapped can use our system as a navigation assistant. Moreover, as a manage assistant, the designed system helps us locate or recall the contents we have read somewhere. More generally, the portability allows us to collect the text information appearing in our daily life as a recorder assistant.

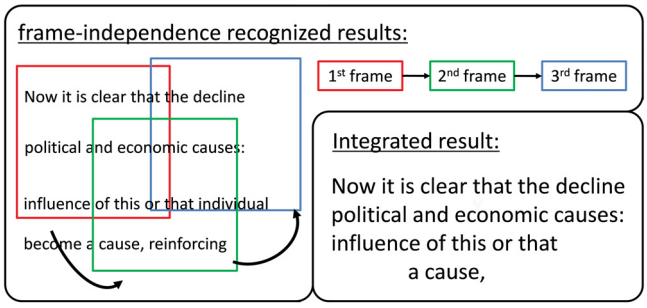


Fig. 2: The diagram of text-based frame mosaicing.

V. Technologies

A. Recognizing texts in a video frame

The video camera working under the shutter-speed at 1/2000[sec] acquires the target text lines. Each frame is then processed by smoothing and Otsu's method. The binary frames are fed into a commercial OCR one by one. For the aim of low power consumption, in the current trial, we just employ the straightforward binarization operator as the preprocessing stage of OCR. In addition, the specialized technology to segment the scene text is out of our scope since it is feasible and tractable to further integrate the existing methods [15–19] with our system.

B. Text-based frame mosaicing

The raw outputs of OCR are far from actual use since frame-independence recognized results may only cover a partial target text lines. We propose a text-based frame mosaicing to integrate all frame-independence recognized results into completed text information (see the diagram in Fig.2). Although it is a natural idea to integrate multiple frames in the reading-life log research, the proposed mosaicing approach effectively leverages the recognized texts, which differs from the conventional image-based mosaicing methods with the prohibitive computational complexity.

The frame mosaicing is fulfilled by the Dynamic Programming (DP) matching in an iteration way. Let P denote a mosaiced frame until the previous frame and S represent currently processing frame. Our problem is to merge P and S for creating a new mosaiced P. At first, P and S are initialized as the first and second frame, respectively. Each row in P and S is assigned by a unique label as illustrated in Fig.3. The line registration between P and S is boiled down to an optimization problem solved by DP matching iteratively. For $\forall\ p_i$ and s_j, where $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$, we compute the edit distance $d(i, j)$ which measures the difference between two strings, and the cumulative distance $D(i, j)$ as follows.

$$D(i,j) = d(i,j) + \min \begin{cases} D(i-1, j-1) \\ D(i-2, j-1) + \alpha \\ D(i-1, j-2) + \alpha \end{cases}$$

where α is the penalty controlling the slope constraint between two steps. Then, we backtrack from (I, J) along each (i, j), and determine the line registration between P and S.

Note that we compute the edit distance $d(i, j)$ by a similar DP matching process. For $\forall\ p_i$ and s_j, let $p_i = p_{i,1}, \cdots, p_{i,m}, \cdots, p_{i,M}$ and $s_i = s_{j,1}, \cdots, s_{j,n}, \cdots, s_{j,N}$ denote the specific elements of two text lines, respectively. A single element $p_{i,m}$ or $s_{j,n}$ corresponds to a recognized

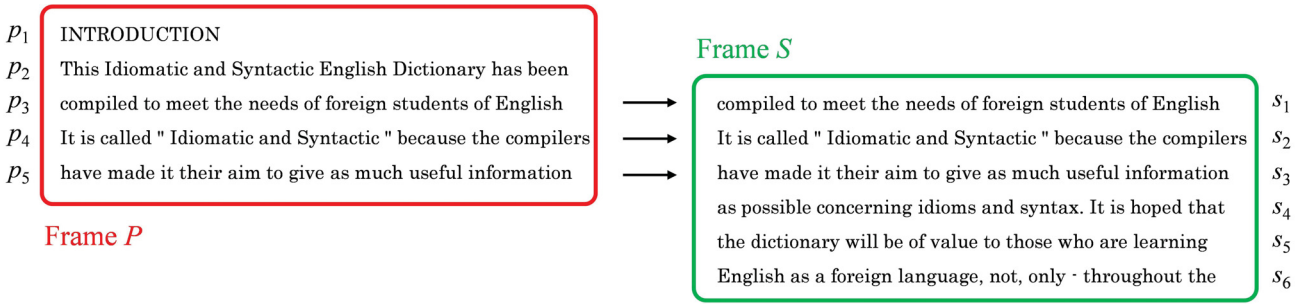


Fig. 3: The illustration of row label assignment for the consecutive frames P and S.

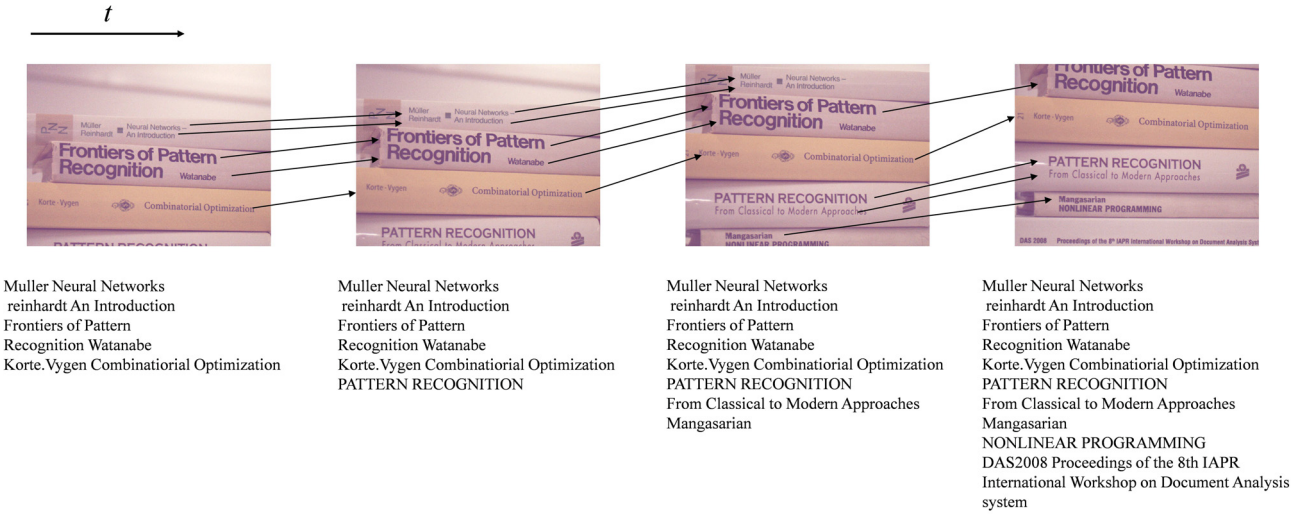


Fig. 4: The real process of the text-based frame mosicing.

character. The calculation process can be formulated below.

$$g(m,n) = \min \begin{cases} g(m-1,) + \beta \\ g(m-1,n-1) + \delta(p_{i,m}, s_{j,n}) \\ g(m,n-1) + \beta \end{cases} \quad (1)$$

Here, $\delta(p_{i,m}, s_{j,n})$ is the function to check whether $p_{i,m}$ and $s_{j,n}$ are identical. Specifically, if $p_{i,m} = s_{j,n}$ holds, $\delta(p_{i,m}, s_{j,n}) = 0$, otherwise $\delta(p_{i,m}, s_{j,n}) = 1$. Similarly, β is the penalty, which is always fixed at one. The edit distance $d(i,j)$ is obtained as $g(M,N)$.

In addition, the actual algorithm is a bit more complicated to be a "starting-point and ending-point free" DP matching to accommodate the fact that the first several lines in P, namely p_1, p_2,$\cdots$, and last several lines in S, namely s_J, s_{J-1}, s_{J-2},$\cdots$, usually do not appear in the S and P, respectively. This happens when the camera moves and captures the new lines from top to bottom. Similarly, for the left-to-right case, the edit distance calculation of Eq. (1) should be organized in the "starting-point and ending-point free" mode.

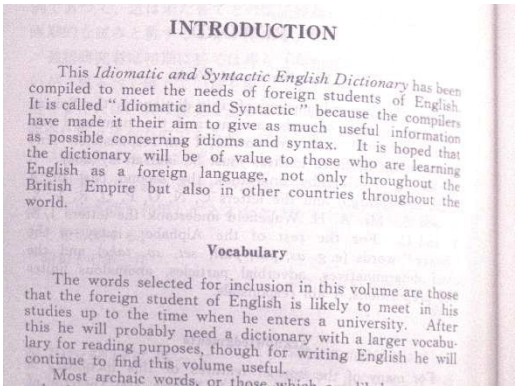
In Fig.4, we display a real process for illustrating the text-based frame mosaicing. The arrows between frames indicate the results of line registration.

VI. Feasibility Test

In this section, we conducted experiments to confirm the feasibility of the proposed system under various environments and targets. The images (a) of Figs.5∼9 displayed the target text lines which appeared in outdoor and indoor signboards, printed-document, LCD screen, and book/magazine cover. For each video, we manually prepared its groundtruth as shown in the left column of Figs.5∼9 (b).

To clearly exhibit the performance, the correct recognized and integrated results were printed in black, the false results in red. Blue characters stood for the missing detections while orange ones represented a misregistration between lines. See the detailed results in the right column of Figs.5∼9 (b). As expected, apart from the outdoor video frame, our system provided reliable recognized and integrated results. For the outdoor environment, the scene text might severely suffer from the non-uniform lighting or specularities so that the global binarization method failed to make a correct segmentation. The specialized methods [15–19] can be further employed in our system to cope with the target text lines appearing in the natural scene. Remarkably, it was infrequent for the misregistration between lines (only occurred in Fig.5) throughout the experiments, which demonstrated that the proposed DP matching based approach was feasible to integrate multiple frames. Moreover, the results with high accuracy given in Fig.9 implied that our system was available for the LCD screen. Furthermore, this example also proved the robustness of our method in the sense that it could work under partial occlusion situation. This was because benefiting from the frame mosaicing stage, the hidden part would be recovered

from other frames.



INTRODUCTION

This *Idiomatic and Syntactic English Dictionary* has been compiled to meet the needs of foreign students of English. It is called "Idiomatic and Syntactic" because the compilers have made it their aim to give as much useful information as possible concerning idioms and syntax. It is hoped that the dictionary will be of value to those who are learning English as a foreign language, not, only throughout the British Empire but also in other countries throughout the world.

Vocabulary

The words selected for inclusion in this volume are those that the foreign student of English is likely to meet in his studies up to the time when he enters a university. After this he will probably need a dictionary with a larger vocabulary for reading purposes, though for writing English he will continue to find this volume useful.

Most archaic words, or those which are likely to occur only in purely scientific and technical contexts, have been excluded. Colloquial and slang words and expressions have been included if they are of the sort likely to be found in books (e. g. modern fiction and drama) read by- students, Foreign words and Latin words and phrases of common occurrence in English have also been included. ;

Definitions

Definitions have been made as simple as possible. Where Definitions in easy, common words was not practicable or satisfactory, pictures and diagrams have been supplied. A lobster, in the Concise Oxford Dictionary is defined as "a large marine stalk-eyed ten-footed long-tailed edible crustacean with large claws formed by the first pair of feet, , bluish-black before and scarlet after boiling ; it's flesh as food ". The foreign student of English, if he is a beginner, is likely to be puzzled by certain words in this definition ("stalk-eyed " and " crustacean", for example). The C. O. D. was not written specially for him. The ordinary user of the C.O.D. is likely to be a person who knows quite well what a lobster is and who refers to this word in a dictionary only when he needs exact and scientific information of the kind given in the above admirably concise and complete definition. the foreign student usually needs only to identify the new

(a) Original frame.

(b) Groundtruth and results

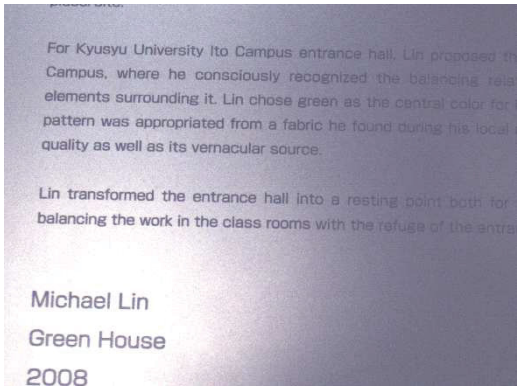
Fig. 5: Printed-document.

VII. Conclusion

In this paper, we propose a reading-life log system implemented by a head-mounted and unobtrusive video camera, which can serve as various assistants in our day life. Compared with the eye tracker-based system, our trial practically performs a text-based frame mosaicing procedure by DP matching. In the feasibility tests, we conduct the experiments by taking various environments as targets. The proposed system can give reasonable responses for most target text lines. There seldom exists the misregistration between lines, indicating the text-based mosaicing can effectively integrate multiple frames. Since the performance in the natural scene is still dubious, our future work is to combine with the existing scene text detection method under a low power consumption and real-time mode.

Acknowledgment

The authors thank the support of Creation of Human-Harmonized Information Technology for Convivial Society (CREST), which is a project organized by Japan Science and Technology Agency (JST).



(a) Original frame.

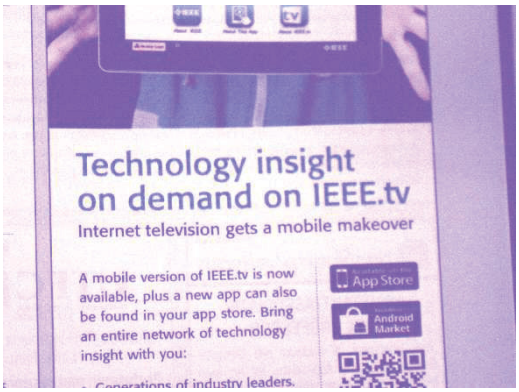
(b) Groundtruth and results

Fig. 6: Indoor signboard.

(a) Original frame.

(b) Groundtruth and results

Fig. 7: Outdoor signboard.

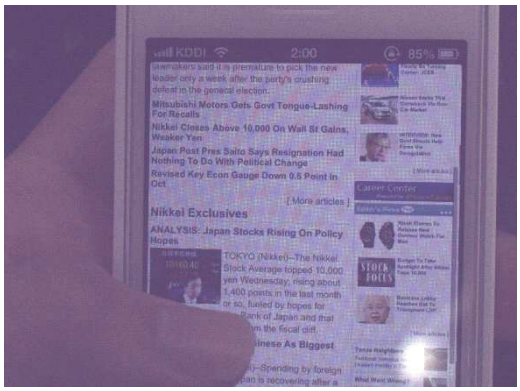


(a) Original frame.

Technology insight on demand on IEEE.tv Internet television gets a mobile makeover A mobile version of IEEE.tv is now available, plus a new app can also be found in your app store. Bring an entire network of technology insight with you: Generations of industry leaders. The newest innovations. Trends shaping our future. Access award-winning programs about the what, who, and how of technology today. Go mobile or get the app. www.ieee.tv IEEE	Technology insight on demand on IEEE.tv Internet television gets a mobile makeover A mobile version of IEEE.tv is now available, plus a new app can also ue Touna in your app store. Dring an entire network of technology insight with you: Generations of industry leaders. The newest innovations. Trends shaping our future. Access award-winning programs about the what, who, and how of technology today. Go mobile or get the app. vvmvJeGe.tv mvw.ieee.tv

(b) Groundtruth and results

Fig. 8: Magazine cover.



(a) Original frame.

SE Asians Replacing Chinese As Biggest Spenders In Japan TOKYO (Nikkei)-Spending by foreign tourists in Japan is recovering after a prolonged sluggish spell caused by worsening Japan-China ties. Hosono Not Running For DPJ Leadership TOKYO (Nikkei)-Goshi Hosono, the Democratic Party of Japan's policy chief, has decided not to run in the party's Dec. 22 leadership election, The Nikkei learned on Wednesday. Toyota Eyeing Output Of 9.9mn Units In '13 Generic-drug Makers Focusing On Cancer Drugs M'bishi Dives Into Shrimp Farming In Thailand Aeon To Take Over Warner Bros. Cineplex Venture [More articles] (Channel JAPAN) Nikkei and TBS Broadcasting for Asia 4th NIKKEI.com Seminar Free access for NIKKEI.com at JAL Lounge in	SE Asians Replacing Chinese As Biggest Spenders In Japan ' TOKYO (Nikkei)-Spending by foreign J.;- I tourists in Japan is recovering after a .-J prolonged sluggish spell caused by worsening Japan-China ties. Hosono Not Running For DPJ Leadership TOKYO (Nikkei)-Goshi Hosono, the Democratic Party of Japan's policy chief, has decided not to run in the party's Dec. 22 leadership election. The Nikkei learned on Wednesday. Toyota Eyeing Output Of 9.9mn Units In '13 Generic-Drug Makers Focusing On Cancer Drugs M'bishi Dives Into Shrimp Farming In Thailand Aeon To Take Over Warner Bros. Cineplex Venture [More articles] (Channel JAPAN) Nikkei and TBS Broadcasting for Asia 4m NIKKEI.com Seminar Free access for NIKKEI.com at JAL Lounge in

(b) Groundtruth and results

Fig. 9: LCD screen.

REFERENCES

[1] C. Campbell, P. Maglio, A Robust Algorithm for Reading Detection, *ACM Workshop on Perceptive User Interfaces*, 2001.

[2] W.Kienzle, F. Wichmann, B. Schölkopf and M. Franz, A Nonparametric Approach to Bottom-Up Visual Saliency, *NIPS*, 2006.

[3] A. Bulling, J. Ward, H. Gellersen and G. Tröster, Roubst Recognition of Reading Activity in Transit Using Wearable Electrooculography, *International Conference on Pervasive Computing*, 2008

[4] G. Buscher and A. Dengel, Attention-Based Document Classifier Learning, *DAS*, 2008.

[5] S. Xu, H. Jiang and F. Lau, User-Oriented Document Summarization through Vision-Based Eye-Tracking, *International Conference on Intelligent User Interfaces*, 2009

[6] T. Judd, K. Ehinger, F. Durand and A. Torralba, Learning to Predict Where Humans Look, *ICCV*, 2009.

[7] M. Nyström and K. Holmqvist, An Adaptive Algorithm for Fixation, Saccade, and Glissade Detection in Eyetracking Data, *Behavior Reserach Methods*, vol.42, no.1, pp.188-204, 2010.

[8] T. Loboda, P. Brusilovsky and J. Brunstein, Inferring Word Relevance from Eye-Movements of Readers, *International Conference on Intelligent User Interfaces*, 2011.

[9] M. Carlos, L. Karel and M. Majid, A Head-Mounted Device for Recognizing Text in Natural Scenes, *CBDAR*, 2011.

[10] G. Duggan and S. Payne, Skim Reading by Satisficing: Evidence form Eye Tracking, *SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[11] O. Dimigen, W. Sommer, A. Hohlfeld, A. Jacobs and R. Kliegl, Coregistration of Eye Movements and EEG in Natural Reading: Analyses and Review, *Journal of Experimental Psychology: General*, vol.140, no.4, pp.552-572, 2011.

[12] G. Buscher, A. Dengel, R. Biedert and L. Elst, Attentive Documents: Eye Tracking as Implicit Feedback for Information Retrieval and Beyond, *ACM Trans. on Interactive Intelligent Systems*, vol.1, no.2, article 9, 2012.

[13] R. Biedert, J. Hees, A. Dengel and G. Buscher, A Robust Realtime Reading-Skimming Classifier, *Symposium on Eye Tracking Research and Applications*, 2012.

[14] K. Kai etc., to be submitted to ICDAR 2013.

[15] K. Jung, K. Kim and A. Jain, Text Information Extraction in Images and Video: A Survey, *Pattern Recognition*, vol.37, no.5, pp.977-997, 2004.

[16] J. Liang, D. Doermann and H. Li, Camera-Based Analysis of Text and Documents: A Survey, *IJDAR*, vol.7, no.2, pp.84-104, 2005.

[17] C. Yi, Y. Tian, Text String Detection From Natural Scenes by Structure-Based Partition and Grouping, *IEEE Trans. on Image Processing*, vol.20, no.9, pp.2594-2605, 2011.

[18] Y. Pan, X. Hou and C. Liu, A Hybrid Approach to Detect and Localize Texts in Natural Scene Images, *IEEE Trans. on Image Processing*, vol.20, no.3, pp.800-813, 2011.

[19] L. Neumann and J. Matas, Real-Time Scene Text Localization and Recognition, *CVPR*, 2012.