2013 12th International Conference on Document Analysis and Recognition

Wearable Reading Assist System: Augmented Reality Document Combining Document Retrieval and Eye Tracking

Takumi Toyama and Andreas Dengel
DFKI GmbH
Kaiserslautern, Germany
Email: Takumi.Toyama@dfki.de
Andreas.Dengel@dfki.de

Wakana Suzuki and Koichi Kise
Osaka Prefecture University
Osaka, Japan
Email: wakana@m.cs.osakafu-u.ac.jp
kise@cs.osakafu-u.ac.jp

***Abstract*—We present a new system that assists people's reading activity by combining a wearable eye tracker, a see-through head mounted display, and an image based document retrieval engine. An image based document retrieval engine is used for identification of the reading document, whereas an eye tracker is used to detect which part of the document the reader is currently reading. The reader can refer to the glossary of the latest viewed key word by looking at the see-through head mounted display. This novel document reading assist application, which is the integration of a document retrieval system into an everyday reading scenario for the first time, enriches people's reading life. In this paper, we i) investigate the performance of the state-of-the-art image based document retrieval method using a wearable camera, ii) propose a method for identification of the word the reader is attendant, and iii) conduct pilot studies for evaluation of the system in this reading context. The results show the potential of a document retrieval system in combination with a gaze based user-oriented system.**

I. Introduction

The advance of recent computer technologies evolves the way of people's reading life. Today, there are plenty of options for readers which form they use when they read a document, not only traditional paper based printed documents, but also digital documents on a computer display, on a tablet PC, etc. The powerful features of such digital form of document are its reusability and linkability. We can easily quote lines by copy and paste on, browse other documents from embedded links. The fact that the usefulness of these various functionalities of digital documents is recognized by many people shows the potential of various types of interaction with documents. Therefore, to explore the potential of new types of document interaction is profoundly important in the context of document analysis. We propose a new human document interaction augmented reality system by combining an eye tracking technology and a document retrieval technology. This system monitors real-time gaze data of the reader and presents feedback information of the document in a head mounted display with respect to the user attention. That is, when the reader reaches to a particular part (word) in a document, the system detect where it is and display the supportive information such as a translation, a glossary, a picture of the article, etc. The proposed system is the integration of a document retrieval system into a wearable computing system in an everyday reading context for the first time.

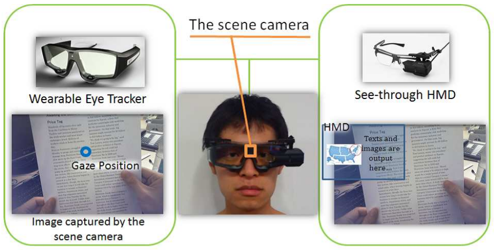

The scene camera
Wearable Eye Tracker
Gaze Position
Image captured by the scene camera
See-through HMD
HMD
Texts and images are output here...

Fig. 1. We combine a wearable eye tracker and a see-through head mounted display. The scene image is captured by the scene camera of the eye tracker (integrated into the center of the frame).

Human attention on document reading is one of the main topics in the eye tracking society over several years [1], [2], [3]. Following the studies in this domain, a number of gaze interaction applications have been developed. For example, in [4], the authors present a framework for developing such a gaze interaction application for documents on a computer display. Although these traditional gaze interaction systems typically rely on a stationary (desk mounted) eye tracker, eye tracking devices available today, which have become small, light-weight and wearable, as shown in Figure 1 open up opportunities to extend the scenario to more ubiquitous scenes. Our approach is based on those previous works but extends the form of a document, not only digital one but also a printed document by using a document retrieval engine and a see-through head mounted display (HMD).

The image of the apparatus used in this system is shown in Figure 1. The scene camera used for the image based document retrieval is integrated into the center of the eye tracker glasses. The see-through HMD allows the user to see the scene in front through the display unobtrusively. Visual feedback of the reading document is presented in this HMD. Furthermore, by calibrating the gaze on the HMD, the user can also control the system using gaze input on the HMD.

The core process of this system is comprised by the gaze analysis process and the image based document retrieval process, as shown in Figure 2. The image based document retrieval module enables us to apply the system universally to printed documents as same as digital documents on a

1520-5363/13 $26.00 © 2013 IEEE
DOI 10.1109/ICDAR.2013.15

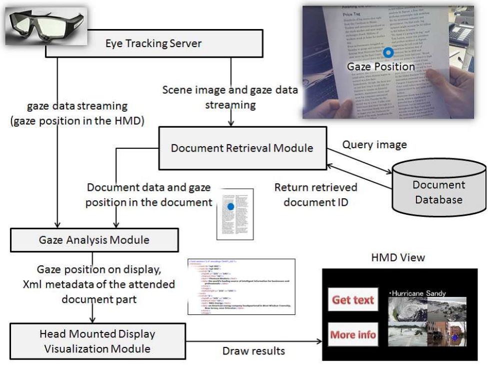

Eye Tracking Server
Gaze Position
Scene image and gaze data streaming
gaze data streaming (gaze position in the HMD)
Document Retrieval Module
Query image
Document Database
Document data and gaze position in the document
Return retrieved document ID
Gaze Analysis Module
Gaze position on display, Xml metadata of the attended document part
HMD View
Get text
More info
Hurricane Sandy
Head Mounted Display Visualization Module
Draw results

Fig. 2. The workflow image.

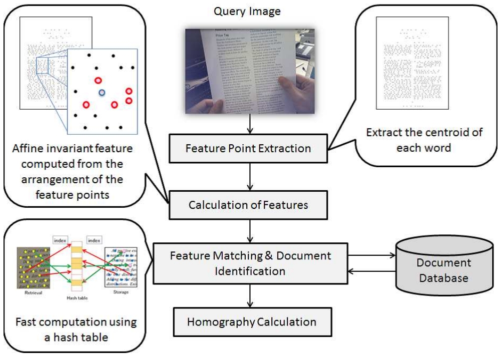

Query Image
Affine invariant feature computed from the arrangement of the feature points
Feature Point Extraction
Extract the centroid of each word
Calculation of Features
index
index
Retrieval
Hash table
Storage
Fast computation using a hash table
Feature Matching & Document Identification
Document Database
Homography Calculation

Fig. 3. An overview of the document retrieval (LLAH) process.

computer display. We would like to emphasize this point because regardless of rapid growth of digital forms of a document, the use of printed, handheld documents still has a large share. In this paper, we evaluate the feasibility of the state-of-the-art document retrieval approach for this wearable system. We also compare the performance with a document image displayed in a screen and with a document printout.

In addition, we propose a method to identify the attended word in the document by the reader during reading using gaze information. As a result of document retrieval, we obtain an identity of the reading document image (page) and its perspective transformation in the camera image. By mapping the gaze position in the image to the original reference document image, upon which part of the document the user is attended is detected. Then the system detects the attended *key word*, which is likely referred to other documents by many users, such as a proper noun, an uncommon term etc. and visualizes the information about the word when the user looks at the display. Note that, the visualization does not appear in the HMD immediately when the attention on the word is detected, because it could be obtrusive for the user. Instead, the user can trigger the information provision by looking at the display.

The contributions of this work are characterized as follows. Firstly, we evaluate the performance and feasibility of the image based document retrieval system in this context using a wearable camera. Second, we propose a method to detect the word attended by the reader using a wearable eye tracker. Lastly, we investigate the benefits of our proposed attention based assist system to the users.

II. Related Works

Several human document interaction applications that integrate image based document retrieval have been presented. For example, HOTPAPER [5] provides a platform to interact with handheld documents using a mobile phone. PaperCraft [6] realizes a system that can recognize user gesture commands on a paper printouts and the user can manipulate digital documents. We would like to extend such a reading assist system concerning the user attention on the document.

In [7], Bulling et al. proposed a method for recognizing reading activity in multiple scenes using body-worn sensors. We can also apply our system in ubiquitous scenarios by incorporating user's gaze activity recognition and realizing a context-aware system. Furthermore, gaze interaction in an augmented reality scenario is also a trend in recent years. Lee et al. presented a system that combined a wearable eye tracker and a see-through HMD [8]. By using gaze, the user can select a marker in the real world and control the system in the HMD interface. In [9] and [10], user gaze is used for triggering the augmented reality information provision about real scene objects or texts. Our method for visualization and attention analysis are based on these previous works.

III. Proposed System

The workflow image of the system is shown in Figure 2. The eye tracking server streams online scene video images and gaze data. The document retrieval module retrieves the currently read document from the database, which is built in advance. Then, it computes the gaze position on the document image. The gaze analysis module checks whether the reader is looking at the document or the HMD. If attention on a particular key word is detected, it extracts xml metadata of the attended word and send it to the HMD visualization module. The visualization module presents the result in the HMD, if the user is looking at the HMD.

A. Eye Tracking

Since we use an off-the-shelf eye tracking product, the full detail of the eye tracking method is not available. However, basically the eye tracking is processed using images from two infrared eye cameras and one scene camera. Each eye is illuminated by six infrared lighting sources and the system tracks the changes of these six infrared light reflections. In order to use this eye tracker, the wearer has to calibrate the system. The calibration process requires the wearer to look at one (or three, if necessary) point(s) in a real scene. The process does not take long for most of users, however it sometimes takes very long for some other users, when they have difficulties to process it once, thus they have to recalibrate it several times until accurate calibration is obtained.

B. Document Retrieval

We adopt an image based document retrieval method proposed in [11]. This method, called LLAH (Locally Likely Arrangement Hashing) is robust to perspective distortion of an image and scale-invariant. An overview of the document retrieval method in shown in Figure 3. When a scene image is given from the camera, the image is blurred by a Gaussian kernel and thresholded adaptively into a binary image in order to detect the centroid of each word region. By changing the size of the Gaussian kernel, we can adjust the optimal image blur for document retrieval, i.e. the distance to the document can be adjusted to some extent. From the arrangements of the detected centroids, affine invariant feature vectors are calculated. The recognition process is done by matching the extracted features to the features previously stored in the database. A hashing technique is used for fast computing. In [12], it is reported that this method can be extended to handle the database of documents up to 10 million pages.

By matching the features between the scene image and the retrieved database image, we also calculate the homography between them. Based on this homography, the gaze on the scene image can be mapped to the gaze on the retrieved document image. The gaze on the document indicates at which word the reader is currently gazing.

C. Gaze Analysis

The gaze analysis module receives the gaze data from the eye tracking server and the result from the document retrieval module. It checks whether the user gaze is on the HMD, and if so, it sends the XML metadata for the latest attended word.

1) Gaze on Document: As a result from the document retrieval, the gaze position of the currently read document is obtained. Here, on which word the gaze is located is referred to the document metadata file which contains the position data for each word. However, it is still very challenging to check it on every single word level, since the accuracy of the gaze position obtained by the eye tracker varies among the users, especially for those who have difficulties for processing proper calibration. Hence, instead of checking the gaze for every single word, we check if the reader's attention is close to a *key word*, which is likely to be referred to other documents by several users, such as, person's name, name of places, uncommon terms, etc. These key words typically do not appear often in each document, but they are very important when reading documents and thus can be seen as an augmented glossary. The system stores the latest attended key word in a stack for displaying information.

2) Gaze on HMD: Since the HMD is mounted on either side of user perspective, as shown in Figure 1, the user can look at the display spontaneously when he/she needs information. Though the display can be partly perceived during reading even if it is not in focus, the transparent display does not disturb the user's reading. The system detects when the user activity transits from reading the document to watching the display, by checking if the user gaze is on the display or not. Therefore, we need to calibrate the HMD. In the HMD calibration process, four individual dots are presented in the HMD and the user has to click the position of each dot in a calibration window, as shown in Figure 4. Then, the system calculates the homography between the scene image and the HMD, so that the gaze position in the scene image can be mapped to the gaze position on the HMD.

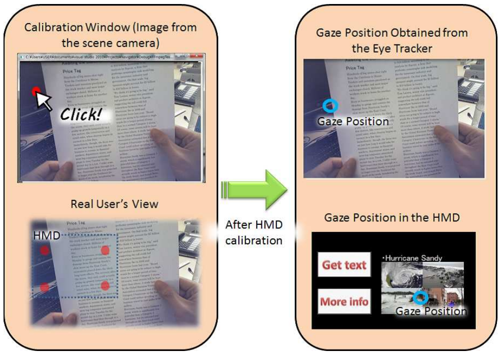


Fig. 4. Calibration of HMD. The user has to click a computer mouse on four points in the window corresponding to the dots in the HMD view.

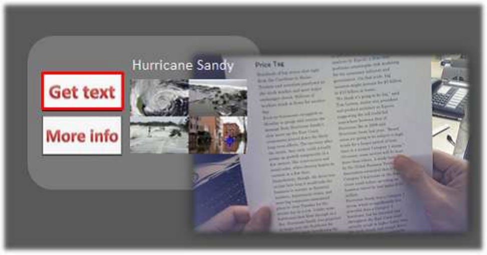


Fig. 5. A sample view of visual feedback. The red bounding rectangle indicates the user is currently selecting the "Get Text" button.

D. Visual Feedback in HMD

When the user's activity transition to watching the HMD is detected, the gaze analysis module sends the XML metadata of the latest attended key word to the display visualization module. Then, the visual feedback is presented to the user in the HMD. A sample view of visual feedback is shown in Figure 5[1].

The user can select a button by gazing. We employ a dwell time (approx. two seconds[2]) approach used in [9] in order to avoid unintended interaction, so called the *Midas-touch* problem. When a button is selected by the user, the visualization module switches the visual feedback accordingly.

IV. Experiments

We conducted three studies in order to evaluate the whole system. We used SMI Eye Tracking Glasses (ETG)[3] for the eye tracker and Brother Airscouter for the HMD. The temporal resolution of the eye tracker is 30 Hz (binocular) and gaze position accuracy is $0.5°$ over all distances. The resolution of the scene camera is 1280×960.

[1] In order to provide the user with a option, on which eye front the HMD is mounted (left or right), the visualization can be flipped depending on the HMD position.

[2] The dwell time can be tuned to each user.

[3] http://www.eyetracking-glasses.com

TABLE I. THE DOCUMENT RETRIEVAL ACCURACY FOR EACH DISTANCE OF THE CAMERA TO THE PRINTOUT.

Distance [cm]	15.0	20.0	25.0	30.0	35.0	40.0	45.0	50.0
Accuracy (Gaussian kernel size 3×3) [%]	99.41	100.0	100.0	100.0	100.0	98.78	74.37	23.36
Accuracy (Gaussian kernel size 7×7) [%]	100.0	100.0	100.0	99.59	47.65	0.0	0.0	0.0

TABLE II. THE DOCUMENT RETRIEVAL ACCURACY FOR EACH ANGLE OF THE CAMERA TO THE PRINTOUT.

Angle [°]	45	60	75	90
Accuracy [%]	18.44	100.0	100.0	100.0

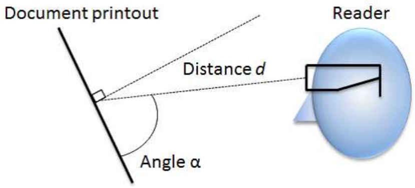


Fig. 6. Distance d and angle α to the document.

A. Document Retrieval Performance Using a Wearable Camera

Firstly, we evaluated the performance of the state-of-the-art image based document retrieval method and investigated the feasibility of the system in this reading context. Table I and II show the results of the document retrieval using the scene camera of ETG, when the distance and the angle to the document (A4 printout, single column) are changed as shown in Figure 6, respectively. In Table I, the results when the size of Gaussian kernel is changed are also shown. We built a database of 10 document images[4] (pages). The accuracy is calculated as the ratio of the number of correctly retrieved document images to the number of retrieval processes for 30 seconds (one document retrieval process takes less than 40 msec., i.e. faster than the scene camera capturing speed with 25 fps). From these results, we can observe this method works quite well when the distance from the document to the camera is ranged from 15 cm to 40 cm and the angle is ranged from 60° to 120°. The performance drops significantly when the distance or the angle is not in these ranges. These results indicate that the system allows the user to move the head position quite freely.

We then also asked 10 persons to wear the eye tracker and read one page of printed document naturally, i.e., they read the document as they usually do. The retrieval results are shown in Table III (with Gaussian kernel size 7×7). The system worked perfectly for almost all persons. We also asked the persons to read documents (PDFs) displayed on a computer screen (Samsung SyncMaster 24 inches) with the same size as the document printout. As shown in the table, this method can also deal with displayed digital documents. In addition to these results, we also confirmed these results change very little with a different brightness value of the screen. The results show the feasibility of the state-of-the-art document retrieval method in this reading context using a wearable camera for a handheld document printout as well as a digital document displayed on a computer screen.

B. Attended Key Word Detection

Next, we investigated the accuracy of the attended word detection. In this study, we asked 13 persons to participate. They were given one page of a document (A4 printout with two columns), which was generated from the text of The New York Times online article on October 30, 2012, titled ”Awaiting the Storms Price Tag”. We selected seven key words (*the Carolinas and Maine, Hurricane Sandy, the Northeast, Eqecat, Hurricane Ike, Hurricane Irene* and *Category*) from 377 words. We asked the test persons to read out the texts and checked if the spoken word matches the detected one for each key word. The histograms of recall and precision rates of entire test persons are shown in Figure 7. When the test persons had problems in the calibration of the eye tracker, the performance drops drastically (0 - 20%). However, for a couple of users the system worked quite well (more than 80% precision and 100% recall rate). For many persons, the precision rates were worse than the recall rates. This usually happened when two key words located nearby (in two consecutive lines) as shown in Figure 8, the other key word is presented mistakenly. However, if two key words have one line in between, the detection mostly succeeded, i.e., the system can detect attended key words with approx. 0.4 cm (one line) accuracy, which actually corresponds to the accuracy of ETG (0.4 cm is approx. $40(cm) \times tan(0.5^\circ)$). From this result, we can observe performance gaps among test persons. Including the previous experiment, we could infer that the attended key word detection approach reasonably performs well depending on the eye tracking calibration. As long as the calibration is done properly, the attended key word is detected correctly.

C. Gaze Interaction

We also conducted a study for gaze interaction on the HMD. Here, we investigated i) if the user activity transition from reading the text to watching the display is detected and ii) if the gaze input for selecting buttons works properly. Similar to the previous studies, the test persons were given one page of a document. The user read the document but in the given document, there is a line requesting the user to look at the HMD. Then, the user has to look at the HMD and select buttons in the HMD using the gaze. We checked if the system properly presents the result in the HMD when the user looks at it and if the gaze button controls are done successfully. In transition detection, the recall rate was 100% and the precision rate was 44% in average. This means that the results were successfully presented every time the user looked at the display, however it also presented even if they did not look at the display. The average of the accuracy of gaze button control was 81%. Again, we could observe significant performance gaps among persons in this study as well. Some persons achieved perfect results (100%), whereas some other persons had no results (gaze position was not obtained because of the error of calibration). In this display interaction, the problem of the calibration was not only with the eye tracking, but also with the HMD calibration for a couple of users (who

[4] This method is however able to extend the database size even larger than 10 million images without significant performance loss [12],

TABLE III. THE DOCUMENT RETRIEVAL ACCURACY FOR DIFFERENT PERSON AND DIFFERENT DOCUMENT FORMS.

Test person	A	B	C	D	E	F	G	H	I	J
Distance [cm]	35.0	40.0	40.0	30.0	30.0	35.0	30.0	35.0	40.0	35.0
Angle [°]	60	60	60	65	65	50	80	50	65	60
Accuracy (with a handheld printout) [%]	100.0	100.0	100.0	100.0	100.0	97.31	100.0	100.0	99.80	100.0
Accuracy (with a computer screen) [%]	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.87	100.0	100.0

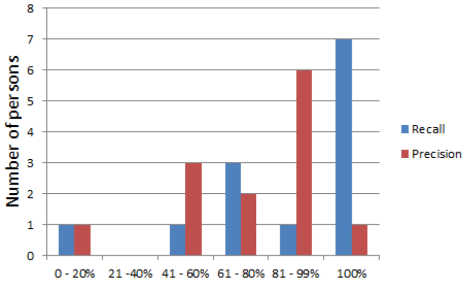
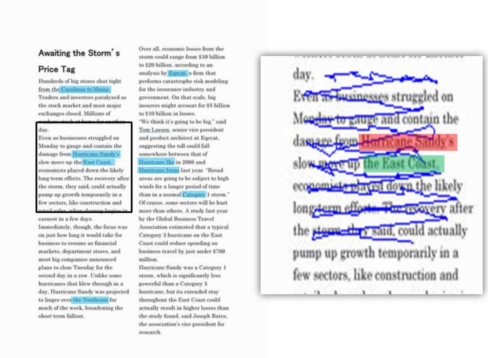


Fig. 7. The histograms of recall and precision rates of attended word detection for entire test persons.



Awaiting the Storm's Price Tag

Hundreds of big stores shut tight from the Carolinas to Maine. Traders and investors paralyzed as the stock market and most major exchanges closed. Millions of workers stuck at home for another day.

Even as businesses struggled on Monday to gauge and contain the damage from Hurricane Sandy's slow move up the East Coast, economists played down the likely long-term effects. The recovery after the storm, they said, could actually pump up growth temporarily in a few sectors, like construction and retail sales, when cleanup begins in earnest in a few days.

Immediately, though, the focus was on just how long it would take for business to resume as financial markets, department stores, and most big companies announced plans to close Tuesday for the second day in a row. Unlike some hurricanes that blew through in a day, Hurricane Sandy was projected to linger over the Northeast for much of the week, broadening the short-term fallout.

Over all, economic losses from the storm could range from $10 billion to $20 billion, according to an analysis by Eqecat, a firm that performs catastrophe risk modeling for the insurance industry and government. On that scale, big insurers might account for $5 billion to $10 billion in losses.

"We think it's going to be big," said Tom Larsen, senior vice president and product architect at Eqecat, suggesting the toll could fall somewhere between that of Hurricane Ike in 2008 and Hurricane Irene last year. "Broad areas are going to be subject to high winds for a longer period of time than in a normal Category 1 storm."

Of course, some sectors will be hurt more than others. A study last year by the Global Business Travel Association estimated that a typical Category 3 hurricane on the East Coast could reduce spending on business travel by just under $700 million.

Hurricane Sandy was a Category 1 storm, which is significantly less powerful than a Category 3 hurricane, but its extended stay throughout the East Coast could actually result in higher losses than the study found, said Joseph Bates, the association's vice president for research.

day.
Even as businesses struggled on
Monday to gauge and contain the
damage from Hurricane Sandy's
slow move up the East Coast,
economists played down the likely
long-term effects. The recovery after
the storm, they said, could actually
pump up growth temporarily in a
few sectors, like construction and

Fig. 8. Left: The page of document used in the study. Blue rectangles show the positions of key words in this document. Right: A gaze path image obtained from one of the users. As shown in this image, two words closely located are hard to distinguish on which word the user is attended.

also had difficulties of the eye tracker calibration).

D. Discussion

We asked the test persons (13 persons) a couple of questions for evaluating the whole system after the tests. In summary, they had positive impressions about this system. For example, regarding the question about the benefit of this type of interaction system: "Would you appreciate additional information when you read a document?", more than 77% of the users agreed. This result shows the potential of human document interaction with augmented reality in this reading context, even though this study is rather small and still an artificial environment. However, some of the participants disliked the hardware constraints that they have to wear two glasses (the HMD and the eye tracker, even more when they had their own optical glasses) and reported they sometimes felt stress during the calibration, especially when they had to repeat it. We still need to tackle with these challenges in order to realize a more useful application.

V. CONCLUSION

We presented a system that assists people's reading activity by combining a wearable eye tracker, a see-through head mounted display and an image based document retrieval engine. Furthermore, we showed the feasibility of the state-of-the-art image based document retrieval method using a wearable camera and proposed a method for detecting an attended word in a document during reading. The results from the pilot studies showed the real potential of the future of this assist system in a reading context. In future, we would like to improve the calibration performance and increase the resolution of attended word detection, not only with key words, but with arbitrary words. In addition, a thorough study for the usability of the system and interface design (HMD) is required.

ACKNOWLEDGMENT

This work was supported in part by JST, CREST as well as JSPS KAKENHI Grant Number 25240028.

REFERENCES

[1] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychological Bulletin*, pp. 372–422, 1998.

[2] G. Buscher, A. Dengel, R. Biedert, and L. V. Elst, "Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond," *ACM Trans. Interact. Intell. Syst.*, vol. 1, no. 2, pp. 9:1–9:30, Jan. 2012.

[3] F. Alt, A. S. Shirazi, A. Schmidt, and J. Mennenöh, "Increasing the user's attention on the web: using implicit interaction based on gaze behavior to tailor content," in *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, ser. NordiCHI '12, 2012, pp. 544–553.

[4] R. Biedert, G. Buscher, S. Schwarz, J. Hees, and A. Dengel, "Text 2.0," in *Proc. of the 28th of the International Conference on Human Factors in Computing Systems*, 2010, pp. 4003–4008.

[5] B. Erol, E. Antúnez, and J. J. Hull, "Hotpaper: multimedia interaction with paper using mobile phones," in *Proceedings of the 16th ACM international conference on Multimedia*, 2008, pp. 399–408.

[6] C. Liao, F. Guimbretière, K. Hinckley, and J. Hollan, "Papiercraft: A gesture-based command system for interactive paper," *ACM Trans. Comput.-Hum. Interact.*, vol. 14, no. 4, pp. 18:1–18:27, Jan. 2008.

[7] A. Bulling, J. A. Ward, and H. Gellersen, "Multimodal recognition of reading activity in transit using body-worn sensors," *ACM Trans. Appl. Percept.*, vol. 9, no. 1, pp. 2:1–2:21, Mar. 2012.

[8] J.-Y. Lee, S.-H. Lee, H.-M. Park, S.-K. Lee, J.-S. Choi, and J.-S. Kwon, "Design and implementation of a wearable ar annotation system using gaze interaction," in *Consumer Electronics (ICCE), 2010 Digest of Technical Papers International Conference on*, jan. 2010, pp. 185 –186.

[9] T. Toyama, T. Kieninger, F. Shafait, and A. Dengel, "Gaze guided object recognition using a head-mounted eye tracker," in *Proc. of the Symposium on Eye Tracking Research and Applications*, 2012, pp. 91–98.

[10] T. Kobayashi, T. Toyamaya, F. Shafait, M. Iwamura, K. Kise, and A. Dengel, "Recognizing words in scenes with a head-mounted eye-tracker," *Document Analysis Systems, IAPR International Workshop on*, vol. 0, pp. 333–338, 2012.

[11] T. Nakai, K. Kise, and M. Iwamura, "Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval," in *Lecture Notes in Computer Science (7th International Workshop DAS2006)*, vol. 3872, Feb. 2006, pp. 541–552.

[12] K. Takeda, K. Kise, and M. Iwamura, "Real-time document image retrieval for a 10 million pages database with a memory efficient and stability improved llah," *2011 International Conference on Document Analysis and Recognition*, pp. 1054–1058, Sep. 2011.