Semi-automatic Text and Graphics Extraction of Manga using Eye Tracking Information

Christophe Rigaud, Thanh-Nam Le, J.-C. Burie, J.-M. Ogier
Laboratoire L3i,
Université de La Rochelle
Avenue Michel Crépeau, 17042 La Rochelle, France
{thanh_nam.le, christophe.rigaud, jcburie, jmogier}@univ-lr.fr

Shoya Ishimaru, Motoi Iwata, Koichi Kise
Osaka Prefecture University
1-1 Gakuencho, Nakaku, Sakai, Osaka 599-8531, Japan
ishimaru@m.cs.osakafu-u.ac.jp,
{iwata, kise}@cs.osakafu-u.ac.jp

***Abstract*—The popularity of storing, distributing and reading comic books electronically has made the task of comics analysis an interesting research problem. Different work have been carried out aiming at understanding their layout structure and the graphic content. However the results are still far from universally applicable, largely due to the huge variety in expression styles and page arrangement, especially in *manga* (Japanese comics). In this paper, we propose a comic image analysis approach using eye-tracking data recorded during manga reading sessions. As humans are extremely capable of interpreting the structured drawing content, and show different reading behaviors based on the nature of the content, their eye movements follow distinguishable patterns over text or graphic regions. Therefore, eye gaze data can add rich information to the understanding of the manga content. Experimental results show that the fixations and saccades indeed form consistent patterns among readers, and can be used for manga textual and graphical analysis.**

I. Introduction

Eye movements are generally considered by research communities as an important indication of human visual attention during their interaction with media such as images or videos. Early studies on recording and analyzing human eye movements date back to decades ago, and originally most attempts were to improve the accuracy of eye-tracking devices, and to invent less intrusive methods [1]. Nowadays, eye-tracking technique has reach a quite mature state where one can easily develop a system on his own, and thanks to its ubiquitous applications, a lot of manufacturers have developed a wide range of consumer products in industrial scale. These devices are getting more and more affordable and will probably be integrated to a large number of reading devices, in order to capture eye movements and provide additional functions.

Since eye-tracking data can be analyzed to provide many fascinating insights about reading activity, such as in [2]–[4], it can be applied to understand the behavior of readers on comic books reading. Manga (Japanese comic art) has long become popular world-wide and represent a cultural trait as well as an important industry. Manga, or comics in general, is the art of telling stories through a sequence of arranged pictures (panels). The panels normally consist of text and graphic elements: backgrounds, characters, speech balloons, visual effects, drawn texts, etc. The graphic is mainly in black and white and gray shades depicted by halftoning dots.

Fig. 1. Example of a heat map on average saccade path of 25 readers. Blue to orange hues (lower to higher temperature) correspond to rare to frequent paths taken by the readers. Manga content credits: [6].

Manga is often composed in a much more realistic and free style compared to Western comics, and is challenging to create. Besides immense drawing skill, it requires the experience to compose the layout of elements, to effectively convey the story and "*secure control of the readers attention and dictate the sequence in which the reader will follow the narrative*" [5]. Unlike motion picture, the story line is presented spatially rather than temporally. Each single panel is a frozen moment, carefully captured and set at its "right" place. Manga artists control the flow of reader's attention via placement of elements, to lead the reader smoothly through the pages (e.g. the speech balloons can be easily associated with its corresponding subjects and read in the correct order). The general rule to follow the pages in manga is to read from right to left, from top to bottom, line by line, following the "reverse-Z" shape. However in practice, it is not always the case. The configuration of the page, the size and the shape of the panels can be so different from a certain page to the next one, and some decisions are purely artistic choice, although there are certain conventions in manga design.

With such a complex and rather free-style layout arrangement, the available methods in analyzing manga page content still prove to be inefficient in various cases. Our work in this

paper is based on the premise that eye-tracking the readers yield much more semantic information than the "bottom-up" methods alone can provide. One contribution is that the additional information allows "online" analysis of manga in particular, which embraces top-down underlying semantic cues (Fig. 1). In this study, we perform eye-tracking on different readers on a set of manga images. All readers have some basic background of reading manga, so the flow from panel to panel is expected to be consistent. The eye-tracking patterns show that the first hint is to actually follow is the speech balloon arrangement, then the right-left, top-down rule. The viewers' attention is indeed guided to follow a certain flow, not only in pages but also inside each panel. This offers information to semi-automatically order panels or speech balloons. We also propose to highlight the benefits of eye gaze position for extracting elements from manga page images, as well as investigate the challenging problem of manga character retrieval by analyzing the fixations and saccades. Another contribution is the potential application of knowledge-driven analysis of manga content, where algorithms that can compute top-down semantic cues via eye gaze data, can interact with bottom-up information to infer new information, and to understand progressively the content of the page, or support manga artists in designing task.

In the next section, we review content extraction method related to manga image analysis along with eye gaze information. Section III details the proposed system from eye gaze movement acquisition to graphics retrieval. Section IV presents the experiments that we performed to validate the proposed method. Finally, Section V and VI discuss and conclude this work respectively.

II. Related Work

Manga analysis concerns several fields of research, from layout analysis to text extraction, graphics recognition and also text/graphics association. We first review eye gaze information for document analysis, and then address text and manga character extraction as they are the most important information in such images.

A. Eye gaze information

Eye gaze information is widely used for recognizing human's activity including cognitive tasks like reading [7] [8]. In the field of document analysis, eye gaze is one of the best features to investigate the content of document from the view of human's reading behavior. For example, important parts of document for many readers can be distinguished by reading-skimming detection proposed by Biedert *et al.* [9]. The total number of words and words per minute of a certain reader and the difficult to understand words can be also detected by using eye-tracking devices [2], [3]. Although lots of work investigate eye gaze with reading documents, there are few researches which involve manga images. Kunze *et al.* have proposed to classify five document types (novel, manga, magazine, newspaper, textbook) by using eye gaze from mobile eye tracker [4]. Our research follows the path yet we focus on more detail part in manga analysis. We recognize readers' eye gaze information on different types of content in a manga page (e.g. on text, comic character, and background) and use this information for the analysis of manga layout.

B. Manga image analysis

Manga image analysis is a quite recent and challenging field of research that attracts more and more attention according to the growing demand of the mobile manga reading market. Text and graphics analysis was studied for decades on different document images including newspaper, administrative documents, checks and graphical documents such as maps and engineering drawings [10]. More recently, manga image analysis has been investigated with the generalization of mobile devices for reading activities. Manga images have a very specific layout that guides the reader through the story. The early approaches developed toward layout analysis concerned the reading order of the panels [11], [12].

Text analysis in manga images opens up several interesting applications such as image compression [13] and content re-targeting to different mobile reading devices [14]. Despite the growing interest, text extraction from manga image remains a challenging problem because mangas are a mixture of text, graphics and graphic sounds with strong relationships [15]. The efficiency of combining text and speech balloon features to extract both of them has been highlighted recently [16]. The benefits of using eye-tracker for text extraction has already been demonstrated as in [2], [17]. Few work about comic character extraction have been published until now. Recently, we improved Sun's method to fit character retrieval for manga [18]. We demonstrated that few manually labeled data can make the detection more robust against various postures and facial expression [19]. We recently demonstrate that speech balloon and tail detection allow manga character region of interest extraction and character-to-speech association [20], [21].

III. Proposed approach

The proposed approach consists in eye gaze data acquisition, filtering and analysis. These three steps are detailed in the following subsections.

A. Eye gaze data processing

Human's eye movements are classified into fixations and saccades. A fixation is an attention on a single location of the vision, a saccade is a quick jump between two fixations [7]. We detect fixations and saccades from raw eye gaze data as shown in Fig. 2 and 3. The radius of each circle correspond to the fixation duration and the line between two circles represents a saccade. Noises of the eye tracking device are also filtered in this process. The fixation-saccade detecting algorithm is based on Buscher's approach [22]. Fig. 4 shows the process of the algorithm. First, a new minimum fixation is detected if 4 consecutive points are in a rectangle of Th_1 pixels (e.g. p_1, p_2, p_3 and p_4 in Fig. 4). Four consecutive points from 42Hz represents human's minimum fixation duration (80-100ms). If

Fig. 2. Non filtered raw eye gaze data from eye tracking device.

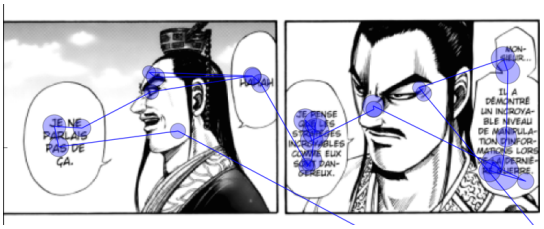


Fig. 3. Filtered output representing fixations (circles) and saccades (lines).

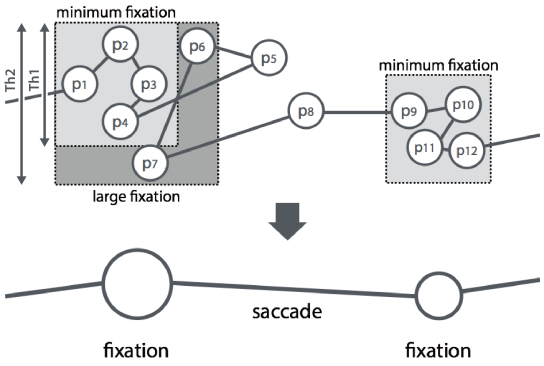
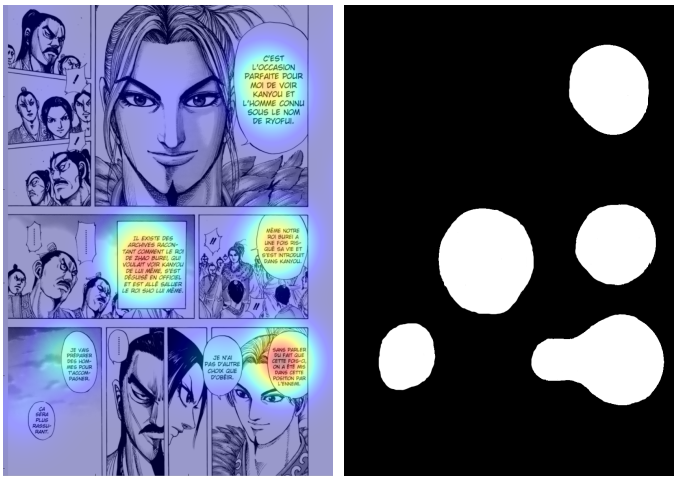


Fig. 4. Fixation and saccade detection process from eye gaze points. Two fixations and one saccade are computed from twelve non-filtered gaze points.

the length of a side of the rectangle including the next gaze point is smaller than Th_2 pixels, we include it and make the rectangle bigger. Yet if the position of gaze point is outside of rectangle with Th_2 pixels, we remove the point as a noise (e.g. p_5 and p_8 in Fig. 4). If 3 consecutive points are out of the rectangle (e.g. p_8, p_9 and p_{10} in Fig. 4), we start to detect the next minimum fixation from the first point which is outside the rectangle (e.g. p_8 in Fig. 4). Finally, we combine gaze points in each rectangle and store the center of mass as the coordinate of the fixation and duration between the time stamp of first to last point in the rectangle.

In order to enhance visualization of multiple reader eye gaze behaviors, we calculate the average of data of multiple readers and generate two kinds of heat map. The first heat map corresponds to the distribution of saccades. We draw saccade path as a line whose width is 30 pixel and overlay multiple readers' paths on one image. An example with 25 readers is presented in Fig. 1.



(a) Fixation heat map (single user) (b) Corresponding regions

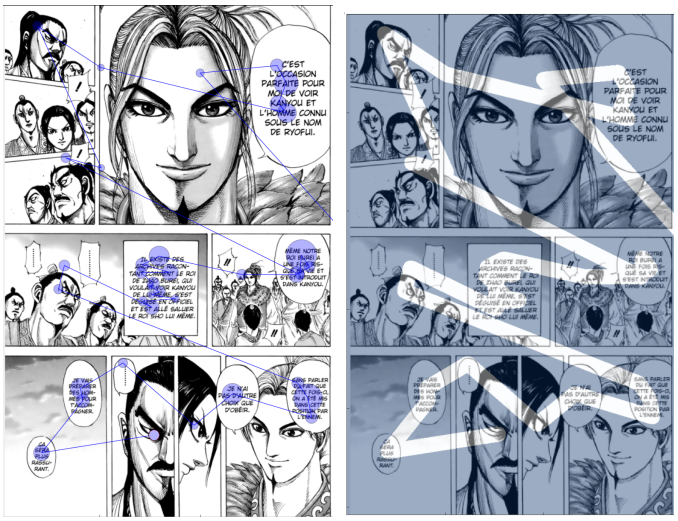
Fig. 5. Fixation heat map and corresponding region. Image credits: [6].

B. Fixation and saccade analysis

Filtered eye gaze data contain eye fixation positions and duration. We would like to use them for text ad "important" graphic region extractions (e.g. manga characters).

1) Text region extraction: For text regions, our brain requires much more effort to recognize, understand, and process the information, compare to understanding graphic regions in general (text in not natural). Therefore, fixation duration analysis is relevant for text region extraction, especially when the fixations are spatially close and the duration is significantly longer. We propose to combine position and duration of fixation points in order to extract text regions from the images. The combination is done by using a Gaussian distribution centered on each fixation. The distribution parameters are the fixation coordinate (μ) and a factor of duration as standard deviation (σ). The duration factor σ is computed from a distance in pixel N multiplied by the fixation duration. The distance N has been set to 4% of the image size which roughly correspond to the size of a text line into a speech balloon of manga (chosen by experiments). All the distributions are then summed and the important overlap between spatially close distribution makes them easily distinguishable on a heat map view for instance (Fig. 5a). Then we extract text regions using a connected component analysis on a binary segmentation of the distributions (Fig. 5b). The threshold used for the binary segmentation was fixed to 50% of the total distribution value range in order to extract the user's main fixation regions only. Note that the extracted text regions roughly correspond to speech balloon regions.

2) Graphics extraction: Graphics region can be decomposed into two categories: the foreground containing most of the focus attention of the story, and the background which is more for illustration purpose. In this paper we focus on foreground content extraction which is relevant in the context of manga understanding. As introduced in Section II-B, offline image analysis still quite weak at extracting graphics in



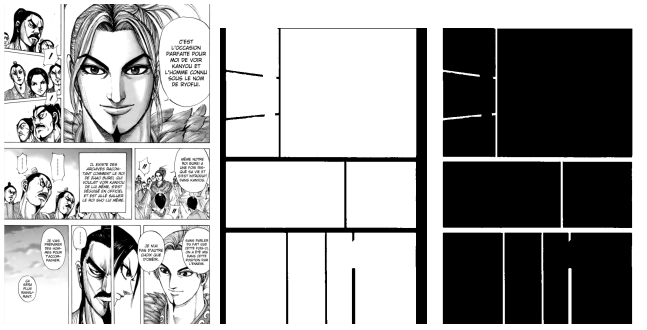
(a) Saccade segments (blue lines) (b) Corresponding regions

Fig. 6. Saccade heat map and corresponding region. Image credits: [6].

manga because their appearance change a lot within the pages, titles and even more between different titles from different artists (e.g. deformation, occlusion, perspective, scale change). Online analysis has the advantage of providing the watched regions encoded as saccades map (Fig. 6a).

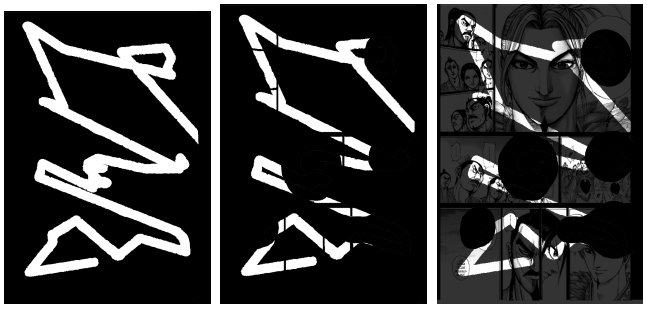
Saccades are deduced from jumps between to fixations as introduced in Section III-A. They do not correspond to well delimited graphics in the image but overlap part of the foreground regions in each panel. Sometimes, there is also short fixations on the graphics (Fig. 3). The human brain don't necessarily need a fixation to understand the "content". He can "catch" (or understand) the meaning of the picture also instantaneously. A saccade has the advantage to pass through important graphic regions chosen by the reader to understand the story but at the end it is only a 1D segment which is not easily convertible as a 2D region corresponding to a graphic region in the image. Note that we manually enlarged the saccade segment in order to obtain a 2D region representative of the main foreground elements that have been visualized by the reader (Fig. 6b). The enlargement is even more accurate when computed from several user's saccade paths (Section IV-B2). Nevertheless, saccade segments are not passing only through important foreground regions. They are also sometimes overlapping text and page background regions. We propose to remove these two information by combining the previously computed text regions (online) and the inverse of panel extraction results (offline).

Panel extraction algorithms aim to segment panel from image background (usually white). In our case, we are interested by extraction the background region which is simply the inverted output of panel extraction (Fig. 7). As introduced in Section II-B, several panel extractor algorithms have been proposed, we use our latest contribution but any of them could be used [20]. This panel extractor is based on dark connected component topology analysis (panels are not included in other



(a) Original image (b) Panel extraction (c) Background extraction

Fig. 7. Image background extraction process. Image credits: [6].



(a) Original saccade path (b) Balloon and background subtraction (c) Final region mask

Fig. 8. Graphics region extraction process. Image credits: [6].

elements).

Finally, from the original saccade path we remove text and background regions to isolate the most watched graphic regions (Fig.8).

IV. Experiments

In this section we introduce the experimental setup and then we evaluate the proposed text and graphic detection methods.

A. Experimental setup

The eye gaze data are recorded while the manga pages are being viewed on the screen by the participants, the setup is shown as in Fig. 9. We use *Tobii eyeX* [1], an inexpensive eye tracking device from *Tobii Technology*, at sampling rate of 42Hz. The device only requires a short calibration for each user, and allows them to move their head freely in front of the screen, although users are requested to try to keep a fixed distance to the screen to attain the best accuracy. The tracker operates based on corneal reflection technique: multiple infrared diodes, invisible to the human eyes, are used to create a reflection on the cornea, these reflections are then captured by the tracker's sensors and used to model the eyes in three dimensions and to calculate the gaze point. We asked 25 participants of both genders, from different ages (14 to 45 year old) and nationalities (French, Tunisian, Pakistani,

[1] http://www.tobii.com/eyex/

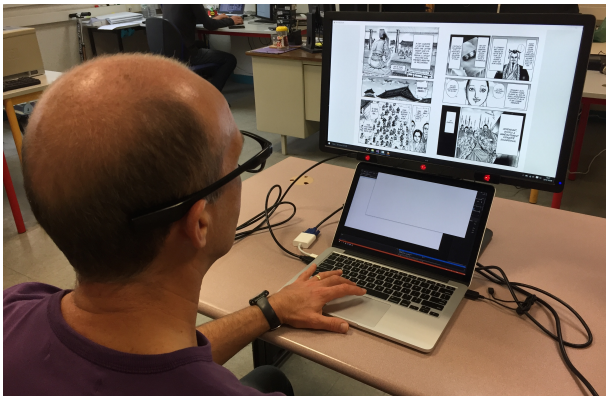


Fig. 9. Data recording setting with Tobii eyeX fixed under the top screen.

Cambodian, Colombian, Indonesian, Vietnamese, Japanese). All the participants were asked to read the same chapter from a famous Japanese manga [6]. We choose to let users stay on each image as long as they need because each participant has different speed of reading/understanding, or behavior, or what we call "manga literacy". The readers are guided to advance through the images by pressing on the keyboard. This may sometimes leads to outlier points where the eyes move out of the screen to seek the key, but those points can be filtered out quite easily. The image set is a whole chapter in the title, containing 20 pages in total, presented as double pages which resulted in 10 images in total. It contains 139 speech balloons and 124 main manga character instances. The resolution of each image is 1683x1200 pixels. Since the participants have different native languages, two versions were available for choosing (English and French) in order to provide the best natural reading experience. Fig.9 shows the overview of experiment setup. According to the image size, we set thresholds for fixation-saccade detection Th_1 and Th_2 as 50 and 80 pixels respectively. Note that the way of detecting the eyes, calculating the gaze position, the calibrating and the environment can all affect the accuracy of the eye tracker.

B. Evaluation

We compared the proposed online text and graphic detection methods to offline methods used for speech balloon and manga character extraction.Note that we performed the experiment, for each image, on the average results of all the participants in order to reduce the impact of participant variability such as gender, culture, age, unusual reading direction etc. (Fig. 1). The comparison between extraction regions and ground truth region were performed at object level such as the PASCAL VOC challenge [23]. In this challenge, the detection were assigned to ground truth objects and judged to be true or false positives by measuring bounding box overlap. To be considered a correct detection, the overlap ratio a_0 between the predicted bounding box B_p and the ground truth bounding box B_{gt} must exceed 0.5 (Formula 1). The predicted objects were considered true positives TP if $a_0 > 0.5$ or false positives FP (prediction errors).

$$a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{1}$$

Detected regions were assigned to ground truth objects according to maximum overlap criterion. Multiple detection of the same object in an image were considered false detection. The number of TP, FP and false negative (missed elements) FN was used to compute the recall R and the precision P of each of the methods using Formula 2 and 3. We also computed the F-measure F for each result.

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

1) Text regions: We evaluated the proposed text regions extractor to a state of the art offline method proposed by our team [16] which is able to extract speech balloon if they contain text-like elements (e.g. similar height, aligned, centered). Note that we did not compare to a text line extraction method because fixation and saccades analysis provides regions that are more similar to speech balloon than text lines. Speech balloon extraction performance are presented Table I.

TABLE I
SPEECH BALLOON SEGMENTATION PERFORMANCE IN PERCENT.

Method	R	P	F_1
Rigaud [16]	**74.10**	60.95	66.88
Proposed	68.84	**95.95**	**80.17**

The best recall concerns the offline method because the presented online method misses speech balloons that contain few text and thus require a very short fixation to be extracted compared to usual balloons. However, the proposed method has an excellent precision because it is based on user's eye movement which is highly reliable.

2) Graphics regions: We compared the proposed graphics extraction method to a state of the art method able to extract manga character faces [19]. We chose this method because it is the most similar to our proposal which also focus on character faces (highly viewed by the reader). Results are presented Table II for $a_0 = 0.2$ in order to accept more partial faces as correct.

TABLE II
CHARACTER FACE SEGMENTATION PERFORMANCE IN PERCENT.

Method	R	P	F_1
Iwata [19]	54.90	**72.72**	62.57
Proposed	**95.96**	66.11	**78.28**

The comparison between the offline method and the proposed method are quite interesting. Even through we relaxed the overlap criterion which means that the faces may only be detected at 20% of their total surface, the recall is very

good. This verifies that the eye movements are really reliable to localize main manga faces or character parts because it is passing over most of them. However, the precision of the proposed method is lower than the offline method because saccade paths are continuous and thus overlap other regions as well.

V. Discussion

The presented method relies on fixation and duration computed from eye gaze movements. Although a calibration step was performed for each reader before the recording, an offset may appear if the reader changes position while reading. We were not able to measure this offset which may shift fixation positions but not duration. This first work has been performed on a small set of manga images to validate the use of eye gaze information for online image analysis with Latin script. It could easily be extended to Franco-Belgium and American comics. The proposed method requires users to read the image which could be a constraint of the system. However, this could be done withing the editorial process where the author could invite some people to first read his work.

VI. Conclusion

In this paper we presented a semi-automatic extraction methods for text and graphics regions. The proposed method is based on eye gaze fixation and saccade analysis and works best when several readers are involved into the data acquisition. This is a preliminary work but it opens up several new ways of research. In the future will would like to link text to graphic regions in order to retrieve the associations between speech text and speaking characters. More generally, top-down knowledge-driven analysis could benefit from eye gaze data to interact more precisely with bottom-up information. Another use of this online method could be crowd-based groundtruthing.

Acknowledgment

This work was mainly supported by the University of La Rochelle (France) and the bilateral program Sakura between Campus France (PHC) and the Japan Society for the Promotion of Science (JSPS). It was also supported in part by JST CREST, JSPS Kakenhi, Grant Number 25240028 and 15K12172 from Japan. We are grateful to all authors and publishers of comics and manga from eBDtheque and Kingdom datasets for allowing us to use their works.

References

[1] A. O. Mohamed, M. P. Da Silva, and V. Courboulay, "A history of eye gaze tracking," 2007.

[2] K. Kunze, H. Kawaichi, K. Yoshimura, and K. Kise, "The wordometer–estimating the number of words read using document image retrieval and mobile eye tracking," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 25–29.

[3] A. Okoso, T. Toyama, K. Kunze, J. Folz, M. Liwicki, and K. Kise, "Towards extraction of subjective reading incomprehension: Analysis of eye gaze features," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2015, pp. 1325–1330.

[4] K. Kunze, Y. Utsumi, Y. Shiga, K. Kise, and A. Bulling, "I know what you are reading: recognition of document types using mobile eye tracking," in *Proceedings of 17th annual international symposium on International symposium on wearable computers*, 2013, pp. 113–116.

[5] W. Eisner, *Comics and Sequential Art: Principles and Practices from the Legendary Cartoonist (Will Eisner Instructional Books)*. WW Norton & Company, 2008.

[6] Y. Hara, *Kingdom chapter 175 - Riboku, Kanyou Bound*. Chiyoda, Tokyo: Shueisha Inc., 2009, vol. 17.

[7] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster, "Eye movement analysis for activity recognition," in *Proceedings of 11th International Conference on Ubiquitous Computing*, 2009, pp. 41–50.

[8] S. Ishimaru, K. Kunze, K. Tanaka, Y. Uema, K. Kise, and M. Inami, "Smart eyewear for interaction and activity recognition," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2015, pp. 307–310.

[9] R. Biedert, J. Hees, A. Dengel, and G. Buscher, "A robust realtime reading-skimming classifier," in *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 2012, pp. 123–130.

[10] G. Nagy, "Twenty years of document image analysis in pami," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 38–62, 2000.

[11] C. Guérin, "Ontologies and spatial relations applied to comic books reading." in *PhD Symposium of Knowledge Engineering and Knowledge Management (EKAW)*, Galway, Ireland, 2012.

[12] L. Li, Y. Wang, Z. Tang, and L. Gao, "Automatic comic page segmentation based on polygon detection," *Multimedia Tools Applications*, vol. 69, no. 1, pp. 171–197, 2014.

[13] C.-Y. Su, R.-I. Chang, and J.-C. Liu, "Recognizing text elements for svg comic compression and its novel applications," in *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*. Washington, DC, USA: IEEE Computer Society, 2011, pp. 1329–1333. [Online]. Available: http://dx.doi.org/10.1109/ICDAR.2011.267

[14] Y. Matsui, T. Yamasaki, and K. Aizawa, "Interactive Manga retargeting," in *ACM SIGGRAPH 2011 Posters on - SIGGRAPH '11*. New York, New York, USA: ACM Press, 2011, p. 1. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2037715.2037756

[15] A. Ghorbel, J.-M. Ogier, and N. Vincent, "Text extraction from comic books," in *Proceedings of the 11th IAPR International Workshop on Graphics Recognition (GREC)*, Nancy, France, 2015.

[16] C. Rigaud, J.-C. Burie, and J.-M. Ogier, "Text-independent speech balloon segmentation for comics and manga," in *Proceedings of the 11th IAPR International Workshop on Graphics Recognition (GREC)*, Nancy, France, 2015.

[17] A. Clavelli, D. Karatzas, J. Llads, M. Ferraro, and G. Boccignone, "Modelling task-dependent eye guidance to objects in pictures," *Cognitive Computation*, vol. 6, no. 3, pp. 558–584, 2014. [Online]. Available: http://dx.doi.org/10.1007/s12559-014-9262-3

[18] W. Sun and K. Kise, "Similar partial copy recognition for line drawings using concentric multi-region histograms of oriented gradients," in *Proceedings of the IAPR Conference on Machine Vision Applications*, ser. MVA2011, Nara, JAPAN, June 13-15, 2011.

[19] M. Iwata, A. Ito, and K. Kise, "A study to achieve manga character retrieval method for manga images," in *Proceedings of the 11th IAPR International Workshop on Document Analysis Systems (DAS2014)*, Apr. 2014, pp. 309–313.

[20] C. Rigaud, C. Gurin, D. Karatzas, J.-C. Burie, and J.-M. Ogier, "Knowledge-driven understanding of images in comic books," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 18, no. 3, pp. 199–221, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10032-015-0243-1

[21] C. Rigaud, N. Le Thanh, J.-C. Burie, J.-M. Ogier, M. Iwata, E. Imazu, and K. Koichi, "Speech balloon and speaker association for comics and manga understanding," in *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR). To be published.* IEEE, 2015.

[22] G. Buscher, A. Dengel, and L. van Elst, "Eye movements as implicit relevance feedback," in *CHI'08 extended abstracts on Human factors in computing systems*. ACM, 2008, pp. 2991–2996.

[23] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.