A Mixed Reality Head-Mounted Text Translation System Using Eye Gaze Input

Takumi Toyama, Daniel Sonntag, Andreas Dengel
German Research Center for AI (DFKI)
Trippstadter Str. 122, 67663 Kaiserslautern, Germany
firstname.lastname@dfki.de

Takahiro Matsuda, Masakazu Iwamura, Koichi Kise
Osaka Prefecture University
1-1 Gakuen-cho, 599-8531Osaka, Japan
{masa,kise}@cs.osakafu-u.ac.jp,
matsuda@m.cs.osakafu-u.ac.jp

ABSTRACT
Efficient text recognition has recently been a challenge for augmented reality systems. In this paper, we propose a system with the ability to provide translations to the user in real-time. We use eye gaze for more intuitive and efficient input for ubiquitous text reading and translation in head mounted displays (HMDs). The eyes can be used to indicate regions of interest in text documents and activate optical-character-recognition (OCR) and translation functions. Visual feedback and navigation help in the interaction process, and text snippets with translations from Japanese to English text snippets, are presented in a see-through HMD. We focus on travelers who go to Japan and need to read signs and propose two different gaze gestures for activating the OCR text reading and translation function. We evaluate which type of gesture suits our OCR scenario best. We also show that our gaze-based OCR method on the extracted gaze regions provide faster access times to information than traditional OCR approaches. Other benefits include that visual feedback of the extracted text region can be given in real-time, the Japanese to English translation can be presented in real-time, and the augmentation of the synchronized and calibrated HMD in this mixed reality application are presented at exact locations in the augmented user view to allow for dynamic text translation management in head-up display systems.

Author Keywords
Augmented Reality and Projection; Ubiquitous Computing, Smart Environments; Visualization; Mobile and embedded devices

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Input devices and strategies

INTRODUCTION
Optical Character Recognition (OCR) available today provides us with a great deal of opportunities to access digital resources by simply taking a picture of textual objects existing in the real world. For example, people nowadays can use the cameras integrated in their phones to translate texts written in a foreign language when they travel. However, the existing applications of these systems often require the user to click the shutter of the camera or to indicate the region of the objective texts by hand in order to obtain a preferable output. This paper presents an approach to use the persons eyes for more intuitive and efficient input for an OCR application. The user can use his/her eyes to indicate where the region of interest is and to activate the recognition. Visual feedback is presented in a see-through head-mounted display (HMD), not only for the provision of the recognition result but also for a navigation of the interaction process. The experiment results and the user study show the benefits of the proposed approach.

A profound growth of computer vision (CV) technologies observed today can change our everyday life drastically. People nowadays can access to a vast amount of information resources on the Internet, by simply taking a picture of texts in the environment. A number of applications, such as Google Goggles[1], make use of OCR technology in order to supply the user with digital information of textual objects in an Augmented Reality (AR) manner. It is true that the fusion of CV and AR has a great potential for enhancing people's everyday life.

OCR has an important role to play when it comes to automatic ubiquitous text translation services. But translations as a result of the OCR reading and translation process should only be provided upon the users request. We motivate the use of eye gaze instead of pressing a button or clicking a shutter to implement a direct manipulation intelligent user interface for the purpose of displaying text translations in a HMD. The important side effect is that the translations in the synchronized and calibrated HMD are presented at precise locations in the augmented user view. In OCR applications, we always have to inform the system about the interested text regions, mainly by cropping the text region manually. As an intelligent user interface, computers should understand eye gaze information in this context, and anticipate when the user wants translation help and provide it instantly where it is needed, as an augmentation in a HMD at a precise location relative to the real document. In this paper, we propose such a gaze based interactive OCR system, which extends the work in [5], by com-

[1]http://www.google.com/mobile/goggles/

IUI'14, February 24–27, 2014, Haifa, Israel.

http://dx.doi.org/10.1145/2557500.2557528

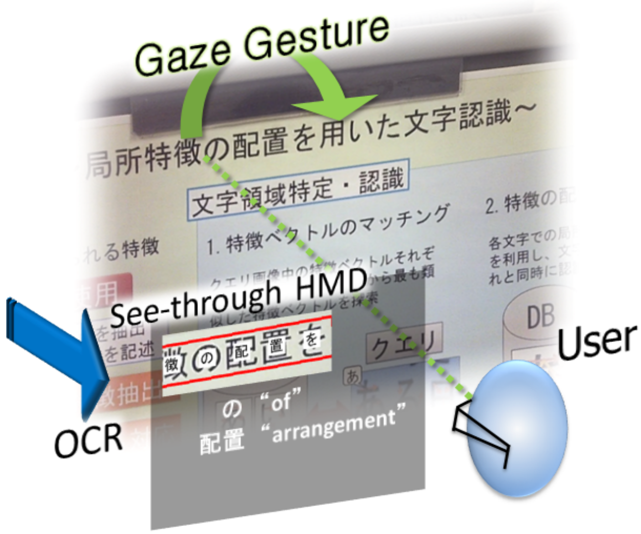

Figure 1. Proposed mixed reality translation application system. English translations of Japanese texts snippets the user is looking at in the real scene are presented in a see-through HMD.

bining the wearable eye tracking system with an AR system. The mixed reality translation application, which would be incomplete without both its real and virtual text components in Japanese and English, respectively, is shown in figure 1. We focus on a product for travelers who go to Japan or a foreign company and need to read signs and posters. One of the drawbacks of gaze interaction with a real scene environment is that the user cannot know instantly whether the computer correctly understands his input without any feedback. In our solution presented in this paper, the OCR reading feedback is presented as a red box superimposed on the source text of the real world, together with the translation result at the exact position (figure 1, grey box).

The paper is structured as follows: first, we presents related work; second, we present the technical approach of the translation system: the gaze gestures for a text in a real scene is discussed and gaze guided OCR techniques are presented. Then, the translation function and the visual feedback on the see-through HMD is presented. Third, we describe the experimental setup and the evaluation of the feasibility of the proposed approach and provide a conclusion. We focus on three contributions, the two methods for selecting/interacting with text in the ORC scenario, the design of the related translation capability, and an evaluation to test the selection methods, followed by a discussion whether these translation capabilities are useful.

BACKGROUND AND RELATED WORK

It is quite natural for humans to read a text with uniform eye movements, which means that we can leverage this phenomenon to create new input methods using OCR systems and head-mounted activation triggers, where the intelligence mainly rests in the head-mounted user interface.

By using human gaze as input, we can design an intuitive interface which uses eye gaze gestures as indicators of the text region of interest, which the user wants the system to recognize (trigger OCR) and translate in real-time. Research such as the idea presented in [10] suggests that eye gaze analysis would benefit text oriented interaction systems.

Additionally, there is another advantage of using gaze in this OCR context: current character recognition methods (even detection method) are still computationally costly, especially when they are applied to high resolution images. Since our method utilises gaze input, we can quickly obtain the region of interest and reduce the computational cost of OCR (see evaluation).

Additionally, it is essential to present information to the user in a fashion that allows clear viewing of the text without obstructing the main reading task. Research on text management systems such as the one recently developed in [7] clearly show the importance of managing information in the users field of view in an unobtrusive, yet viewable manner. We consequently follow in these footsteps in order to improve a users ability to read and translate on the fly without affecting the reading experience.

Eye tracking technology has been used in intelligent user interfaces (IUIs) more and more frequently in recent years. For example, the task of communicating with the user based on eye-gaze patterns [9] introduced an interesting idea: the possibility to sense users' interest based on eye-gaze patterns and manage computer information output accordingly. Following up on this essential observation, we bring eye gaze patterns into the context of translation help in mixed reality settings. Also motivated by previous findings showing the relevance of eye-gaze in multimodal conversational interfaces [8], we extended the passive eye gaze input idea to active eye gaze user input in the mixed reality realm.

Some previous works have contributed to the combination of an eye tracking device with a see-through HMD. A see-through HMD has better compatibility with an eye tracker, since a see-through HMD does not obstruct the user's view and affixing the HMD to the tracker can reduce or eliminate calibration errors. The user can then focus on an object in a real scene through the display. In [12], an augmented reality system for knowledge-intensive location-based expert work is presented. The multi-modal interaction system combines multiple on-body input and output devices (a speech-based dialogue system, a head-mounted augmented reality display, and a head-mounted eye- tracker) in a similar way, but focusses on speech recognition instead of eye gaze while reading to trigger a context-related function and its interpretation.

More recently, several studies and works have shown the benefits of utilizing human (eye) gaze in several HCI contexts, e.g., [6], [5], and [13] focus on the interaction using the objects in the real world and suggest many possibilities of gaze based interaction in a real world scenario. In [5], eye gaze is used to find relevant text passages in a scene and support a subsequent OCR process accordingly. Bringing this technology together to form a combined real-time translation-based interaction system is the unique contribution of the work presented here.

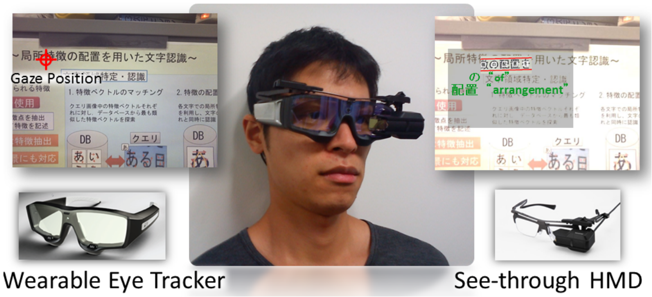

Figure 2. Technical HCI setup. The HMD screen size is limited and does not cover the whole user view (grey region).

Our attention recognition system (active gaze-based user input to trigger the translation function) is based on an eye control system such as [1]. If somebody's eyes are moving from left to right (or vice versa) and if the movement is repeated several times, the user is considered to be reading text. Using this methodology we provide two novel methods for interacting with text, and evaluate their potential for use with HMD systems.

TECHNICAL APPROACH

Figure 2 shows the technical HCI setup. We combine two different devices (head-mounted eye tracker and HMD) by mounting the HMD frame on the top of the wearable eye tracker frame; the HMD should be placed in front of the user's dominant eye which provides a more accurate eye gaze focus. The two scene images shown in figure 2 (left and right, the user is reading a poster) are captured by the scene camera of the eye tracker, which is integrated in the center of the eye tracker frame, and the eye tracker software computes the gaze position in the captured scene image (left). The see-through HMD translation output is show on the right hand side. We use SMIs Eye Tracking Glasses (ETG)[2] and Brothers Airscouter[3] in this work. However, the proposed approach is not limited to this hardware setting; it can be applied to any calibrated other eye tracking and HMD device combination. The resolution of the scene camera is 1280×960 and the gaze accuracy we can obtain via our calibration method is $0.5°$.

The usage procedure is as follows: in a first step to use the eye tracker, the user has to calibrate the system. The eye tracking calibration is done by looking at one reference point in the scene and by clicking a mouse on the reference point in the scene camera image (which is displayed in a computer screen editor that must be used before the real-time usage, similar to the method explained in [12]). After the calibration, the eye tracking server streams scene images with the exact gaze position. Our text-based gaze gesture recognition module receives them and processes the data accordingly, as explained next.

Text-based Gaze Gesture Recognition

By integrating eye gaze gesture recognition (as an extension to [2]), the system can detect the user's intention in a specific context. A typical smartphone-based AR application such as Wikitude SDK applications, www.wikitude.com/, presents additional location and orientation-specific information whenever the AR objects (for a travel guide for example) are in focus.

In our wearable computing HMD / eye-tracker scenario, we try to provide related the information of a more narrowed interaction context.

An intelligent head-mounted interface should filter out the irrelevant information and provides the information only when it is needed and about what is needed in the context of the text reading and translation scenario. Eye gaze input is a useful input method in such a scenario. On the other hand, using eye gaze input also contains an inherent problem, as known as *Midas touch problem* [3], when using it as a direct substitute for the mouse. It it not desirable for each eye move during normal visual perception to initiate a computer command.

Because of this, we employ a more complex eye gaze gesture recognition system, which recognizes rather complex gaze movement patterns, and we propose two new domain-specific eye gaze gestures for the text reading scenario in order to tackle these problems. First, eye gaze patterns observed during text reading are quite symptomatic compared to other gaze activities. Thus, we can utilize these representative and discriminatory gaze patterns as gestures for controlling the translation system and triggering the OCR system. Second, OCR on a high resolution image is still computationally costly. When we apply the OCR recognition (in combination with the text window detection) to the entire images obtained from the scene camera, it cannot be processed in real-time. By focusing on a particular region in a scene, we can reduce the computational cost required in the recognition process. Gaze gestures can easily indicate where text exists in the scene, because a gaze path can draw lines onto the text regions. Third, without a proper gaze gesture implementation, no text window detection and text snippet -based translation are possible in this head-mounted input context.

We propose two gaze gestures for an OCR text reading and translation scenario and investigate which type of gesture suits our OCR scenario best. Essentially, the first looks at the beginning and the end of the text line alternately and repeatedly (*gaze repetitive leap*), and the other is to move gaze from the beginning to the end gradually (*gaze scan*). These two types of gaze gestures can be divided into two groups based on the deliberateness and the complexity of the gesture. The gaze scan gesture is less complex and can occur less deliberately, whereas gaze repetitive leap is more complex and hardly occurs without intention. It can be that the less complex the gesture is, the more false recognitions (false positives) occur. However, if the gesture is too complex, it is also quite demanding for the user. Thus, we need to explore and evaluate the trade-off of the deliberateness and the complexity of the text-based gaze gesture.

The recognition algorithm of each gaze gesture is as follows: *Gaze Repetitive Leap (GRL)*: If a fixation that lasts one second is detected, it activates the recognition process and set the fixation as the start point. If the next detected fixation is right

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

[3] http://www.brother.co.uk/g3.cfm/s_page/888150

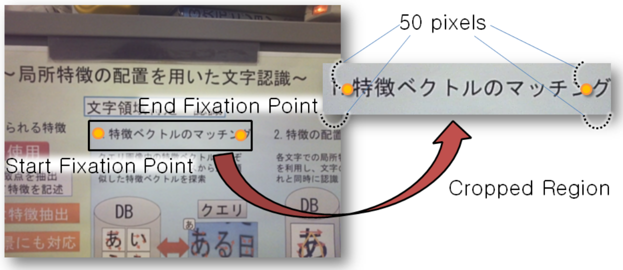

Figure 3. Text snippet extraction for translation guided by gaze gesture.

from the start fixation within ±30°, it is set as the end point fixation. If the third fixation is within d pixels from the start point, it continues the recognition; otherwise the recognition process is discarded. If the following fixations are within d from either the end point or the start point (switches alternately) and n of such fixations are detected in total, the gesture is recognized. (By changing n in the GRL gesture, we can increase the complexity of the gesture. If n is small, the gesture can also occur less deliberately.)

Gaze Scan (GS): If a fixation that lasts one second is detected, it activates the recognition process and set the fixation as the start point. If the next detected fixation is right from the start fixation within ±30°, it continues the recognition; otherwise the recognition process is discarded. If such a fixation (within ±30° to the right) is detected continuously more than three times, it is recognized as a gesture. The end point is determined when the fixation point swerved from ±30°.

Gaze Guided OCR

When the end of the gesture is recognized, the system extracts the text region indicated by the start fixation point and the end fixation point as shown in figure 3. Scene image tracking of the used KLT tracker [11] is running as a background process. It tracks reference points in scenes and computes the relative distances between the scene image frames. While tracking the image, we can estimate the position of the start point even at a later frame[4].

The cropping rectangle area is determined by a heuristic as follows: Left: 50 pixels left from the start point, Top: 50 pixels above from the higher point of either of the start point or the end point, Right: 50 pixels right from the end point, and Bottom: 50 pixels under from the lower point of either of the start point or the end point. The cropped text image is sent to the character recognition module.

Visual Feedback in HMD Screen

Once the start of the gaze gesture is recognized, a navigation image is presented in the HMD screen (figure 4). In this way, the user can ensure that the start of the gesture command is correctly recognized by the system. When the end of the gesture is recognized, the OCR module is triggered and returns the result of the text recognition of the given textual image (this step may include more complex natural language preprocessing steps).

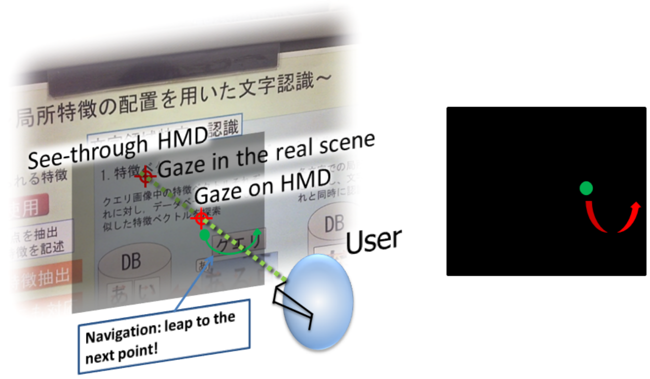

Figure 4. The interaction navigation is presented near the gaze position. In this example, GRL navigation is shown in the HMD.

The resulting selected text snippet, which also includes the cropped image, is visualized in the HMD screen (figure 1), so that the user can check if the correct region-of-interest is recognized; the visual feedback is given instantly as augmented reality (AR) boxes (figure 1, red, green and blue boxes). If no Japanese (Kanji) character is recognized in the given region, no boxes are presented in AR.

Hence, false gesture recognition is rejected based on the character recognition result (if gaze gesture is recognized mistakenly (false positive), and no text is read, it will be classified as gaze detection error (false positive).

By taking the relative position of the HMD in a scene image in the system into account, we can present the navigation near the gaze position as shown in figure 4. Dynamitic text management and intuitive positioning of the augmented translations in the users field of view to migrate user-centric text content is important because if the visualization is far from the user's focus, it cannot be perceived easily; and therefore, the see-through feature of the HMD plays an important role in this system. Another consideration is the position of the translated text throughout a users mobile environment ([7]).

EMBEDDED TRANSLATION SERVICE

Our Japanese-English translation system provides the user with a translated Japanese text snippet as a direct translation from the Japanese text snippet indicated by the eye gaze gesture figure 1).

For the test system presented in this paper, we implemented our own translation function. First, we built a Japanese-English dictionary using 10000 Japanese characters (Kanji)[5]. We selected 10000 common Japanese words from a common Japanese dictionary service and translated them into English using the Microsoft Bing Translator API[6]. The simple translation process, see *Translation Module* in the architecture

[4]To obtain the entire image from the beginning, one sometimes has to apply mosaicing; however, it is not integrated in this work because we focus on rather short texts and they can be captured by one image frame.

[5]Actually, it contains Kanji, Katakana, and Hiragana Japanese characters mixtures. For the sake of simplicity, we say Kanji.

[6]http://www.bing.com/translator

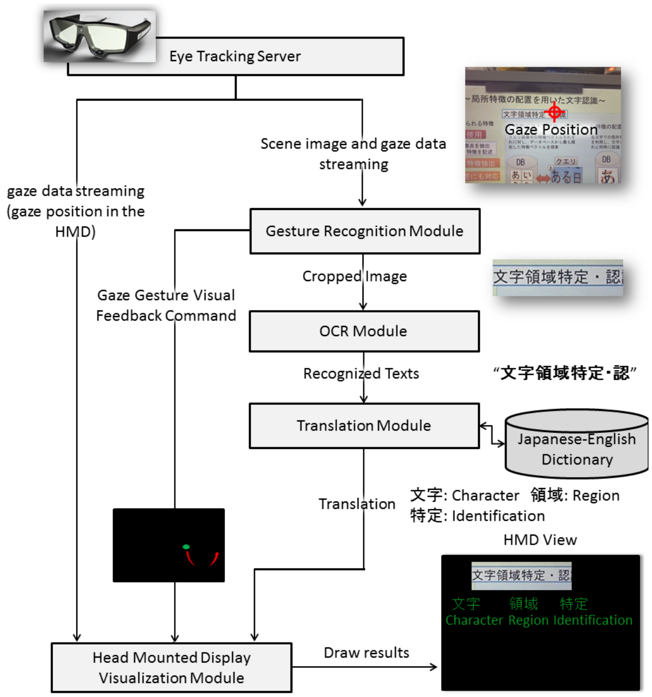

Figure 5. Architeture of embedded translation service

in figure 5, works as follows: 1) The recognized OCR text is preprocessed by a very shallow text processing pipeline for Japanese tokenization. In the evaluation, we used the OCR method presented in [4]. This method is specialized to Japanese characters; 2) The individual Kanji tokens (possibly compounds) are matched against an approximate Kanji index that uses a Levenshtein distance metric; 3) The token-by-token translation of the nearest Kanji compounds in the dictionary according to the distance metric are presented; and 4) If the Levenshtein distance exceeds a threshold, no translation is returned (we assume a bad OCR result). It is to be noted that complex linguistic-based translation pipelines (e.g., [14]) can easily be integrated into the architecture. We relied on a fast on-board solution for the translation integration test. There are also several online translation services available, but they are dependent on the speed of the wireless network and are blackbox processes. Their utility evaluation is beyond the reach of evaluating the real-time feature of the mixed reality head-mounted text translation system using eye gaze input.

EVALUATION

We provide an initial evaluation of the proposed gesture recognition approach; we sought to determine whether the two proposed eye gestures are adequate for triggering the eye gaze based translation function while reading the text.

Gesture Recognition

To test this, a user evaluation under realistic circumstances has been conducted including the following steps: First, we tested the gesture recognition without using any textual real-world image or document in order to evaluate the acceptability of the proposed gestures for the users, i.e., how well people can perform those gestures and how much the system can recognize them in general. The experiment included 10 participants, ranging from age 22 to 56, with an approximately even number of males and females. The participants were asked to perform the the two gestures in order to trigger the translation function. Two types of gesture navigation sheets (as shown in figure 6) were presented to the user and we asked them to perform each gesture (GRL and GS) on each sheet. In addition to the gesture type, three types of gesture length are prepared here (short, medium and long), to compare the recognition performance of each gesture with a different length. The overall result of each gesture is shown in table 1. For *GRL*, n was set to 4 and d was set to 50. In general, the results show that the longer the gesture length is, the less accurate the *GRL* gesture becomes; contrariwise, the longer the GS gesture is, the more accurate the recognition is. We hypothesize that this is because the user cannot find the end point easily when the distance of two reference points becomes greater; contrariwise, using the *GS gesture* it is easier to track a path (line) by eye gaze when it is longer.

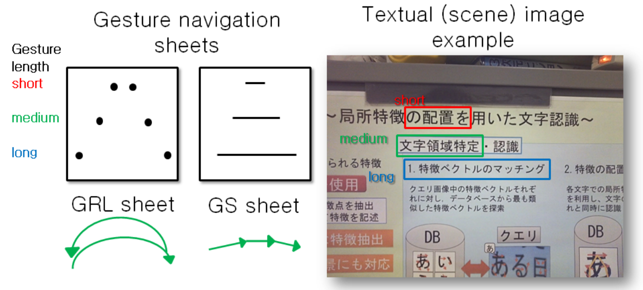

Figure 6. Gesture navigation sheets (left) and an example of actual scene image (right). The scene image contains a complex background.

Table 1. Gesture recognition accuracy.

	GRL gaze gesture		GS gaze gesture	
length	sheet	scene image	sheet	scene image
short	80%	83%	67%	77%
medium	60%	60%	80%	80%
long	60%	33%	87%	83%

In a second evaluation step, we tested the recognition on a textual (scene) image example of proper Japanese text of different lengths. An example image that we used is shown in figure 6 (right), and the result is also shown in table 1. Here, we asked the users to perform the gesture on proper text of length: short, medium, or long. Compared to the non-textual image, it can be said that it becomes even harder to find the end point with proper texts. The *GS* gesture slightly gets better with texts. Some users mentioned that to move eyes along a text line is easier than to move it along a normal straight line. Since these experiments were conducted without a training phase (for the users to try and learn these gestures over time), it would be interesting to see how the performance to use those two proposed gestures improves over time. Though some users had difficulties because they had to concentrate more than usual and felt stressed, these gestures were rather easy and intuitive to perform for the majority of users. We might switch the gesture depending on the length of text, since *GRL* suits short text length, while *GS* suits long text length. Furthermore, the visual feedback could reasonably navigate the user for the correct gesture throughout the experiment. Many users failed the gesture without the visual feedback in the HMD. This result shows the benefit of a see-through AR screen for visual feedback in the proposed system though the experiment we conducted is only adequate to show the basic effectiveness of our algorithm.

Gaze Guided OCR Time Performances

Similar to this image, most of the texts were cropped correctly if the gestures were correctly recognized. Some users had difficulties in gazing on a particular point if there is no characteristic point (reference point), which is sometimes the case when we focus on texts in a real scene. In a third part of the evaluation, we compared the processing time required for the OCR of an entire video image vs. the cropped video. The entire video image required $4.77sec$ on average, the cropped image OCR required only $0.19sec$ on average. This result shows that we can effectively reduce the computational cost by using the gaze gesture guided image cropping to present a translation result in real-time.

CONCLUSION

We proposed a mixed reality system with the ability to provide translations to the user in real-time. The eyes can be used to indicate regions of interest in text documents and activate optical-character-recognition (OCR) text recognition and translation functions. For this purpose, we proposed and evaluated two gaze-based image region selection and OCR activation functions. The experimental results showed the feasibility of the gaze gesture recognition and the utility of gaze gesture as an automatic indicator of the text region of interest. Image cropping guided by gaze also reduced the computational cost of the OCR in order to perform the whole triggering-OCR-translation processing workflow in real-time. Future work includes further user studies for a comprehensive evaluation of the approach in combination with other dynamic text management capabilities of see-through wearable display systems. The system can also be adapted for professional translators or learners of a foreign language, potentially for full translation of documents for the general public. Additionally, visitors in foreign countries can use the method to translate signs that may provide essential safety or navigation information. Lastly, industry workers will gain the ability to read warning labels or instructions on products which may not be translated to into their native language.

Acknowledgements

This research has been supported by the ERmed project, see http://www.dfki.de/RadSpeech/ERmed.

REFERENCES

1. J. M. Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498–504, 2003.
2. M. Hopmann, P. Salamin, N. Chauvin, F. Vexo, and D. Thalmann. Natural activation for gesture recognition systems. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 173–183, New York, NY, USA, 2011. ACM.
3. R. J. K. Jacob. Eye movement-based human-computer interaction techniques: Toward non-command interfaces. In *Advances in Human-Computer Interaction*, volume 4, pages 151–190. Ablex Publishing Co, 1993.
4. T. Kobayahsi, M. Iwamura, and K. Kise. An anytime algorithm for faster camera-based character recognition. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013.
5. T. Kobayashi, T. Toyamaya, F. Shafait, M. Iwamura, K. Kise, and A. Dengel. Recognizing words in scenes with a head-mounted eye-tracker. *IAPR International Workshop on Document Analysis Systems*, pages 333–338, 2012.
6. M. Kumar, A. Paepcke, and T. Winograd. Eyepoint: Practical pointing and selection using gaze and keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 421–430, New York, NY, USA, 2007. ACM.
7. J. Orlosky, K. Kiyokawa, and H. Takemura. Dynamic text management for see-through wearable and heads-up display systems. In *Proceedings of the International Conference on Intelligent user interfaces*, IUI '13, pages 363–370, New York, NY, USA, 2013. ACM.
8. Z. Prasov and J. Y. Chai. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, IUI '08, pages 20–29, New York, NY, USA, 2008. ACM.
9. P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 221–230, New York, NY, USA, 2005. ACM.
10. K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, pages 372–422, 1998.
11. J. Shi and C. Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 593–600, 1994.
12. D. Sonntag and T. Toyama. On-Body IE: A Head-Mounted Multimodal Augmented Reality System for Learning and Recalling Faces. In *9th International Conference on Intelligent Environments (IE)*, pages 151–156, 2013.
13. 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*, pages 91–98, 2012.
14. W. Wahlster, editor. *VERBMOBIL: Foundations of Speech-to-Speech Translation*. Springer, 2000.