Published in ICNGC2017b, The 3rd International Conference on Next Generation Computing, December 21~24, 2017, Kaohsiung, Taiwan.

A language assistant system for smart glasses

Shi Fan Zhang, Kin Hong Wong
The Dept. of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
Email:khwong.cse.cuhk.edu.hk

***Abstract*— In this work, we have developed a language assistant system for people using wearable smart glasses. The system assumes a user is wearing a camera and reading a page of text. When the user points to a word he needs language translation or dictionary lookup, our system will recognize the word using computer vison techniques and pass it on to an electronic dictionary for finding the meaning. It is useful for students and foreign travelers that instant language assistance is necessary. We have performed the study and produced a demonstration with satisfactory results.**

Keywords*— Optical character recognition, smart glasses, pose estimation

I. Introduction

In this paper, we present a prototype smart interface system which can offer language assistance when one is reading a document. As shown in *Fig. 1*, a user is holding a document of text that some words he/she doesn't understand. Using our system, the user only needs to use a pen or a finger to point to the word on the paper, and then a camera can recognize the word using an Optical Character Recognition (OCR) algorithm. The meaning of the word can be found by an electronic dictionary and the result can be displayed on the screen. It is assumed that the user is wearing a smart glass (such as the Google glass) or a camera on his/her head. And the display can be a computer screen or the display of the smart glass. During our research we found that the target paper printed with the text may not be perpendicular to the user's view. A pose rectification module is required to rectify the text geometrically so that the camera can see a non-geometrically distorted text. This scheme may increase the recognition rate of the OCR subsystem.

The major contribution of this work is to demonstrate the possibility of integrating various existing techniques in computer vison for building a language assistant for smart glass users. The product can be used by a traveler trying to understand foreign documents; or assist students to learn a foreign language.

The content of the paper is as follows. Section II discusses the background of the work. Section III reviews the theory and methodology of our approaches. Section IV shows the result of our system. And session V is the conclusions and discussions of the work.

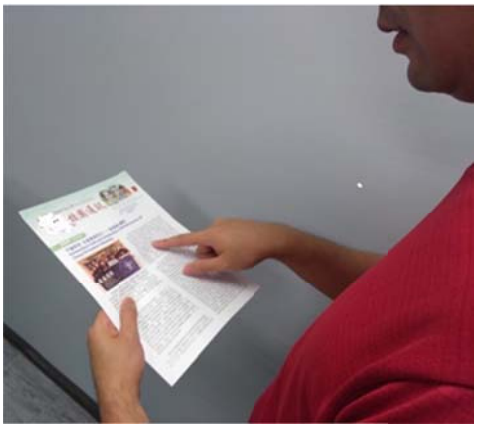

Fig. 1. Example of a person pointing to a word that he wants to translate.

II. Background

Smart cameras are becoming very popular in the field of human-computer interaction. For example, Looxcie [1] is a wearable camera, which targets for general-purpose hand free video recording. Much popular video footage taken by this camera for leisure can be found at YouTube (http://www.youtube.com/). Apart from recreational use, the Hong Kong Police Force is also running trials on "Body Worn Video Camera" for evidence capturing and recording [2]. Besides, there are many different products available for specific markets. These include the product from Recon Instrument [3] that is designed for athletes. Some gadgets even equipped with Global Positioning System (GPS), High Density camera, Head-up-display (HUD) and smartphone connectivity. Personal communication capabilities are also included in some systems, for example, Golden-i [4] features a wearable computer at work, and allows users to send/receive email, or browse webpages or files. The Google Glass is also an example that the Head-up-display and camera pair creates a futuristic look. Not only does it able to browse web-pages, web-map, e-mail etc., it may also handle face recognition and computer vision tasks. Some people are working in this direction and exploring new ways of using it [5].

III. Theory and Methodology

While reading a text, the user may come across some unknown words/characters in a foreign language. Our

system will help the user by translating the word to the language he understands or lookup the meaning of it in a dictionary. In our approach, first, the rectangular image of the paper with the text should be detected. However, since the paper that one is holding may not be perpendicular to the user's view, so the text may be geometrically distorted. To fix this problem, we need to recertify the image so the texts will appear as if they are perpendicular to the users. So that OCR rate can be improved. In our approach the four corners of the rectangular image of the paper are recognized first, and then affine transformation can be applied to rectify the image. Subsequently characters in the text can be recognized via OCR, under the condition that the text in the picture is adequately distinct. Our approach has the following steps.

- Locate the quadrangle which is the boundary of the paper. Since the 3D geometry of the paper is known (assume A4 size paper), so using a pose estimation technique we can find the pose of the paper. Another approach is to use affine transform to rectify the image to the normal pose.
- The next step is to use a finger pointing algorithm to find where the user wants the text to be located on the paper. So we should know which part of the image contains the text to be translated.
- Use the rectified image as the input of the OCR system.

A. *Quadrangle detection*

First we need to locate the boundary of the paper, which is a quadrangle. Line detection using Hough transform is used. In 2015, K. K. Lee [6] combines Randomized Hough Transform and Q-corners to facilitate multiple-quadrilateral detection. A similar idea can be applied to rectangle corner detection. Our approach is based on the method of Hough transform. First all lines inside the image are selected, our target is to find the four corners that constitute the quadrilateral of the paper. In our processing cycle, first we pick two detected lines and calculate their intersection point. If the intersection point is not out of the boundary, one score is voted for this point. We repeat the above for all line pairs, and then four points with the highest scores are the four points of the quadrangle.

Sometimes it is annoying that the points on edges are disturbing the result. We can remove them before our processing. Moreover, to make the result more satisfactory, the background can be neglected using background subtraction.

B. *Affine Transformation*

Three conditions under which Optical Character Recognition (OCR) has excellent performance:

- The text region is normal to the principal axis of the camera lens.
- The text is not skewed.
- The text is clear and distinct from the background.

Therefore, affine transformation is applied to rectify the frames so that the text lines are not geometrically distorted. OpenCV [7] provides several functions to rotate the input frames and are used in our project.

C. *Optical Character Recognition OCR*

We are using a popular tool called Tesseract for our OCR task. Tesseract [8] achieves high accuracy when dealing with printed English characters. Since our users may be Chinese, and we found that for the Chinese text, Tesseract performs poorly if all parameters are set to be default values. Often one Chinese character is split into two so it is obvious that Tesseract is difficult to recognize those characters with radicals (parts of a Chinese character). Therefore, to fix the problem, some parameters (for instance, set psm=6) are adjusted to inform Tesseract that the targets are of the same shape. An example of successfully Chinese text recognition is shown in *Fig. 2*.

Fig. 2. A page containing Chinese text is being processed by Tesseract

We performed the experiments using (1) default setting and (2) adjusted parameter setting; the results are shown in Table 1 and 2, respectively. From the results, it can be found that Tesseract recognizes Chinese characters precisely with adjusted parameters, regardless whether the characters are printed in standard format or some strokes are eroded to some extent.

It should be mentioned that Tesseract can be trained manually. Although the official dataset contains huge chunks of trained data, it may fail to completely recognize all the characters without mistake. However, related documents declare that official dataset cannot be revised. Manual training the Tesseract needs to start everything from scratch, which can be a hard task. So it may be the next target we will achieve in future.

Table 1: Tesseract recognizes Chinese characters with default parameters

	Number of all characters	Number of recognized characters	Overall accuracy
High quality characters	75	73	97.3%
Low quality characters	264	180	68.2%

Table 2: Tesseract recognizes Chinese characters with adjusted parameters

	Number of all characters	Number of recognized characters	Overall accuracy
High quality characters	75	74	98.7%
Low quality characters	264	262	99.2%

D. Finger location

In our system, if the user would like to look up a certain word of a text in the dictionary, he only needs to point to the word with his finger. And then the computer recognizes the character and searches it in the dictionary. We employ a tool similar to that described in [12]. Here the approach to locate the finger is briefly introduced. The feature of human hand includes:

- Five fingers.
- Each finger is a convex polygon
- The space between two adjacent fingers forms a hollow region.

These are useful information for finger detection. First, the background is removed, and then our system will perform image thresholding, contour and convex hull finding. The contour with maximum area (target contour) should form the image of our hand. If we calculate the distance between points on the contour and the convex hull, several local maxima can be found. They correspond to the hollow region between adjacent fingers. Thus, we can select the top point on the target contour and it is the fingertip of the finger. A small image area above the fingertip which contains the target word will be selected. Then, the OCR system can recognize the text which will be passed on to our dictionary to lookup the meaning.

E. Machine Translation for future development

Machine translation has much improved recently. Google even offers machine learning API on the Google Cloud Platform, including Google Cloud Translation. Hence, it is convenient to translate characters recognized by OCR into another language provided that the user's mobile device is connected to the Internet. Offline translation can also be achieved by traditional methods or more advanced ones such as those using neural network approaches. In the simplest case, it is straightforward just to construct a dictionary and do translation word-by-word. But the result is usually poor. Recently, the sequence-to-sequence (seq2seq) model has been applied to machine translation. As introduced by Cho et al. in 2014 [9], a basic sequence-to-sequence model consists of two recurrent neural networks, and the translation process is implemented using an encoder and decoder scheme. Recently, long short-term memory (LSTM) [10] approaches are popular and generate good machine translation results. In our current approach we are only using a dictionary we built and targeted for English vocabulary translation. We will study the machine translation methods for a whole sentence rather than isolated words. And hope in the future our system can translate the whole page of text rather than just isolated words.

F. Pose Estimation for future development

It happens regularly that you are reading a book that is being rotated, so the lines of text are not horizontal when viewing. M.L. Liu and K.H. Wong [11] introduced Pose Estimation algorithm using four corresponding points on an object. This four-point algorithm maps image projection of a 2-D plane back onto the 3-D model plane. The algorithm is capable of calculating the 3D rotation and translation of the paper with respective to the camera based on its image. This will be useful for rectifying the text from a distorted view back to the normal view, so the OCR being performed can be made more accurate. Another application of pose estimation is that if the user has a small mobile projector mounted on his/her head; the camera/projector system can project the correct text at any user specified position on the page. It can be used to display the translation of targeted word (English to Chinese or vice versa) next to the text the user is pointing to. To do this one needs to have the relative pose information between the camera/projector and the paper. This pose estimation process can be carried out by the four-point algorithm mentioned above. We will add this capability into our system in the future version.

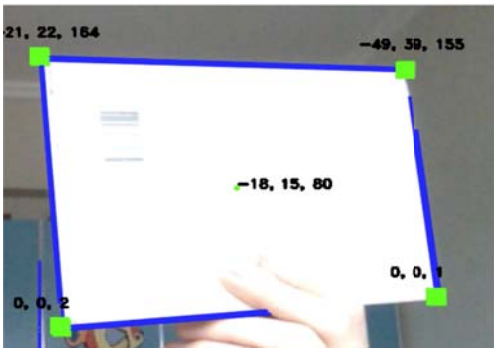

Fig. 3. Four points are found of a target paper, using the four point algorithm [11] the pose of the paper can also be found.

A screen shot of our four-point algorithm implementation is shown in *Fig.3*. The three numbers next to each corner point (green) represent the 3D position of that point, it shows we can obtain the 3D information of the paper object, hence the pose of the paper can be calculated. A video demonstration of our implementation can be found at

https://www.youtube.com/watch?v=diN565iTPdM

IV. Results and Demonstration

We integrated all modules of the system and we show that we can achieve what we proposed to do. That is, the user can point to a word, and the OCR module can look up its meaning in the dictionary and show the result on screen. As shown in *Fig.4*, the user points to a word (right window), since the camera view is left-right reversed, we reverse the image first and then send the selected image (middle window) to the OCR system for recognition. The result is the text of the target image. Then, the recognized text is shown on screen (left window). A video demonstration of this process running in real-time can be found at

https://www.youtube.com/watch?v=cepRpxxQbDs

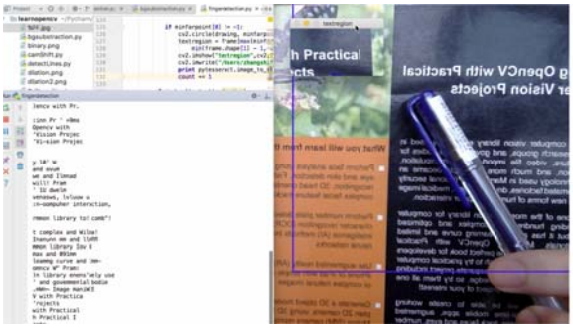

Fig. 4. A demonstrate showing the user is pointing to a word.

V. Conclusions and Discussions

From the result, it is shown that our idea of building a language assistant for smart glass users is feasible. The camera can detect the quadrangle of the paper. Affine transform is applied successful to rectify the text of the page. Then, a finger pointing detection system can be used to find the text to be translated. And an Optical Character Recognition (OCR) system translates the word into text. The meaning of the text can be then be found by an electronic dictionary. The whole process is in real-time and can be very useful for students or travelers where instant language translation of foreign words is necessary.

Acknowledgment

This work is supported by a direct grant (Project Code: 4055045) from the Faculty of Engineering of the Chinese University of Hong Kong.

References

[1] Boland, Justin, and Romulus Pereira. "Wireless headset camera lens." U.S. Patent No. D643,867. 23 Aug. 2011.

[2] Hong Kong Police Force, Body Worn Video Camera Field Trial, http://www.police.gov.hk/ppp_en/11_useful_info/bwvc.html, accessed on 16 Nov 2017.

[3] Recon Instrument, https://www.reconinstruments.com/, accessed on 15 Nov. 2017.

[4] Golden-I, http://www.kopin.com/offerings/headset-solutions/default.aspx accessed on 15 Nov. 2017.

[5] McNaney, Rísin, et al. "Exploring the acceptability of google glass as an everyday assistive device for people with parkinson's." Proceedings of the 32nd annual ACM conference on Human factors in computing systems. ACM, 2014.

[6] Lee, Kai Ki, Ying Kin Yu, and Kin Hong Wong. "Multiple quadrilateral detection for projector-camera system applications." Industrial Electronics and Applications (ICIEA), 2015 IEEE 10th Conference on. IEEE, 2015.

[7] Tuohy, Shane, et al. "Distance determination for an automobile environment using inverse perspective mapping in OpenCV." (2010): 100-105.

[8] Smith, Ray. "An overview of the Tesseract OCR engine." Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. Vol. 2. IEEE, 2007.

[9] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[10] Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM." (1999): 850-855.

[11] Liu M.L. and Wong K.H., "Pose Estimation Using Four Corresponding Points", Pattern Recognition Letters, Volume 20, Number 1 January 1999, pp. 69-74

[12] Gurav, Ruchi Manish, and Premanand K. Kadbe. "Real time finger tracking and contour detection for gesture recognition using OpenCV." Industrial Instrumentation and Control (ICIC), 2015 International Conference on. IEEE, 2015.