AUTOMATIC DETECTION AND TRANSLATION OF TEXT FROM NATURAL SCENES

Jie Yang, Xilin Chen, Jing Zhang, Ying Zhang, Alex Waibel

Interactive Systems Laboratory
Carnegie-Mellon University
Pittsburgh, PA 15213

ABSTRACT

Large amounts of information are embedded in natural scenes. Signs are good examples of natural objects with high information content. In this paper, we discuss problems in automatic detection and translation of text from natural scenes. We describe the challenges of automatic text detection and propose methods to address these challenges. We extend example based machine translation technology for sign translation and present a prototype system for Chinese sign translation. This system is capable of capturing images, automatically detecting and recognizing text, and translating the text into English. The translation can be displayed on a palm size PDA, or synthesized as a voice output message over the earphones.

1. INTRODUCTION

Signs are everywhere in our lives. They suggest the presence of a fact, condition, or quality. They make our lives easier when we are familiar with them, but sometimes they pose problems or even danger. For example, a tourist or soldier might not be able to understand a sign in a foreign country that specifies military warnings or hazards. In this research, we are interested in signs that contain text, and have direct influence upon a tourist from a different country or culture. These signs include street and company names, bulletin boards, announcements, advertisements, and warning notices, and others. At the Interactive Systems Laboratory at Carnegie Mellon University, we are developing technologies for automatically detecting, recognizing, and translating signs [3,15,16]. Sign translation, in conjunction with spoken language translation, can help international tourists to overcome language barriers. This research is a part of our efforts in developing a tourist assistant system [14].

A successful sign translation system relies on three key technologies: text detection, optical character recognition (OCR), and language translation. At current stage of the research, we focus our efforts on automatic text detection and sign translation while taking advantage of existing OCR technologies.

Automatic detection of signs from natural scenes is a challenging problem because they usually embedded in the environment. The task is related to text detection and recognition from video images, or so called Video OCR. Compared to video OCR tasks, sign detection takes place in a more dynamic environment. The user's movement can cause unstable input images. Non-professional equipment can make the video input poorer than that of other video OCR tasks, such as detecting captions in broadcast news programs. Sign detection must also be implemented in real time using limited resources.

Sign translation has some features which makes it different from a traditional language translation task. In general, the text used in the sign is short and concise. This characteristic of signs means that lexical mismatch and structural mismatch become more serious problems. Furthermore, sign translation usually requires context or environment information because sign designers assume a human reader would use such information in understanding signs.

We address these challenges at both technology and system levels. We have developed algorithms for automatic sign detection and applied example-based machine translation (EBMT) [1] technology to sign translation. We have applied the technology to Chinese sign translation. The prototype system can recognize Chinese sign input from a camera, and translate the signs into English or voice stream. The system can run on multiple platforms including palm size PDAs (Personal Digital Assistants).

2. SYSTEM DESIGN

An automatic sign translation system utilizes a camera to capture the image with signs, detects signs in the image, recognizes signs, and translates results of sign recognition into a target language. Such a system relies on technologies of sign detection, OCR, and machine translation. It is also constrained by available resources such as hardware and software.

Our goal is to design a system that requires the minimum modification for working on different platforms and environments. In order to achieve such flexibility, we modularize the system into three modules: capture module, interactive module, and recognition and translation module. These modules work in client/server mode and are independent of each other. They can be on the same machine or different machines. Figure 1 shows the system architecture.

The capture module handles video input. It is hardware dependent. We are currently working on windows platform. The module supports both video for window and DirectX interface. We have tested the module using many different video and digital (DV) cameras through different input channels such as PCI card, PCMCIA card, USB port, and Pocket camera for a PDA.

The video stream or picture is inputted to the recognition and translation module for processing. The recognition and translation module is a key part of the system. The module first performs sign detection and only focuses on the text areas. The sign extraction results are also fed back to a user for potential. These sign regions are further processed and fed into the OCR engine, which recognizes the contents of the sign areas in the original language. Then, the recognition results are sent to the translation module to obtain an interpretation in target language. Under the Interlingual [4] framework, both the source and target

languages can be expanded extensively. Such extension makes a system work for multiple languages.

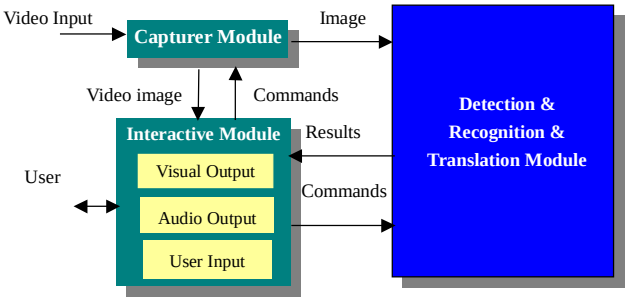

Figure 1. System architecture.

Interactive module provides an interface between a user and the system. A user-friendly interface is important for a user-centered system. It provides necessary information to a user through an appropriate modality. It also allows a user to interact with the system if needed. In our current system, the interface provides the recognition/translation results and allows a user to select sign regions manually or confirm automatically detected text regions. For example, a user can select a sign that he/she is interested to be translated directly if multiple signs have been detected; or in some unfavorable cases the automatic sign extraction algorithms may fail, but a user can still select a sign manually by circling the areas using pointing devices. This function also allows a user to obtain a translation for any part of a sign by selecting the parts where he/she is interested in.

We have successfully demonstrated the prototype system using a HP Jornada Pocket PC with a HP pocket camera as shown in Figure 2. The camera can be simply plugged into the CF type I slot of the pocket PC. It can capture a 640x480 resolution image with a swivel lens that can be rotated 180 degrees. The system runs on Windows CE environment. We developed the system using Microsoft embedded development tools on a desktop computer and then download to the pocket PC.

Figure 2. Automatic sign translation using a PDA.

3. TEXT DETECTION

The previous research in text detection falls into three different areas: (1) automatic detection of text areas from general backgrounds [2, 5, 7, 8, 9, 10, 13, 17]; (2) document input from a video camera under a controlled environment [11]; and (3) enhancement of text areas for character recognition [8, 9, 10, 12, 13]. Although text with a limited scope can be successfully detected using existing technologies, such as in high quality printed documents, it is difficult to detect signs with varying size, embedded in the real world, and captured using an unconstrained video camera. Fully automatic extraction of signs is a challenging problem because signs are usually embedded in the environment.

In order to deal with dynamic environment of sign detection, we have developed algorithms to handle automatic text detection and image deformation. We propose a new adaptive algorithm [3]. The algorithm embeds the adaptive strategy in a hierarchical structure, with different emphases at each layer. The first layer detects possible sign regions. We utilize a multi-resolution approach to compensate these variations and eliminate noises in the edge detection algorithm; i.e., we apply edge detection algorithm using different scale parameters, and then fuse the results from different resolutions. The second layer of the framework performs adaptive search. The adaptive search strategy is constrained by two factors: initial candidates detected by the first layer and layout of the signs. More specifically, the search starts from the initial candidate but the search *directions* and *acceptance criterion* are determined by taking the *layout* of signs into account.

While most signs in western languages are in the horizontal direction, Chinese signs in both horizontal and vertical directions are commonly used. One reason might be that Chinese language is rather character based than word based. Some special signs are designed in specific shapes for aesthetics reasons. We will ignore these layouts at this stage. In addition to directions, we can use shape and color criteria to discriminate different signs. Using these heuristics, we designed searching strategy and criterion under the constraints above, which we called the *syntax* of sign layout. In fact, it plays the similar role as syntax in language understanding when parsing sentences, except that it is used to discriminate different layouts to assist the adaptive searching of sign regions.

The objective of layout analysis is to align characters in an optimal way, so characters belong to the same sign will be aligned together. Chinese text layout has some unique features. A major contribution of the new framework lies in its ability to refine the detection results using local color and layout information. We are considering incorporating more information to this framework to further enhance the detection rate. The detailed algorithm can be found in [3].

Figure 3 is the screen shot of a sign detection example. The sign with Arabic and English text was download from the Internet. The left upper window shows the original sign and detection results marked by black rectangles. The lower windows illustrate the detection process: the left window is the initial detection and the right one is the final decision. Figure 4 shows system robustness against image deformation. The sign was taken from a high resolution camera and printed out on a paper. The system captures the sign from a non-favorite angle. The system can still successfully detect the sign. This demonstrates that the detection framework provides considerable flexibility to allow the detection of slanted signs and signs with non-uniform character sizes.

Figure 3. An example of automatic text detection.

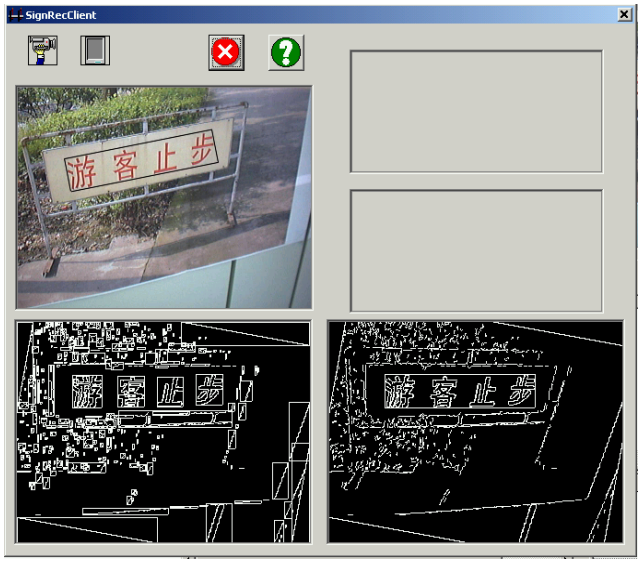

Figure 4. An example of deformed text detection.

We have evaluated the prototype system for sign detection and translation. We have built up a database containing over 2000 Chinese signs taken from China and Singapore. In our database, signs were taken by a high resolution digital camera and printed out on papers. During the test, the signs are caught by a video camera and detected in real time. We have tested the robustness of the detection algorithms by changing conditions, such as image resolution, camera view angle, and lighting conditions. The algorithms worked fairly well for low resolution images (e.g., from 320 x 240 to 80 x 60). The algorithms can also handle signs with significant slant, various character size and lighting conditions. We tested the sign detection algorithm using 50 randomly selected signs from our sign database. Table 1 gives the automatic sign detection results. By properly selecting parameters, we can control the ratio of miss detection and false alarms. Presently, such parameters are selected according to user's preferences, i.e. acceptability of different types of errors from users' point of view.

Table 1. Results of automatic detection on 50 Chinese signs

Detection without missing characters	*False alarms*	*Detection with missing characters*
45	*8*	*5*

4. SIGN TRANSLATION

Sign translation is different from a traditional language translation task in some aspects. The lexical requirement of a sign translation system is different from an ordinary machine translation (MT) system, since signs are often filled with abbreviations, idioms, and names, which do not usually appear in formal languages. The nature of a sign requires it be short and concise. The lexical mismatch and structural mismatch problems become more severe in sign translation because shorter words/phrases are more likely to be ambiguous due to insufficient information from the text to resolve the ambiguities. Furthermore, sign translation is sensitive to the domain of the sign: lexical in different domains has different meaning. However, domain identification is difficult because the signs are concise and provide few contexts. For structural matching, the system needs to handle ungrammatical language usage, which is common in signs.

Moreover, imperfect sign recognition makes sign translation more difficult. Though in many cases human being can correctly "guess" the correct meaning using context knowledge even with erroneous input, for MT systems, it is still a difficult problem. In summary, with the challenges mentioned above, sign translation is not a trivial problem that can be readily solved using the existing MT technology.

In the existing MT techniques, the knowledge based MT system works well with grammatical sentences, but it requires a great amount of human effort to construct its knowledge base, and it is difficult for such a system to handle ungrammatical text which appears frequently in signs.

On the other hand, Statistical MT and Example Based Machine Translation (EBMT) [1] enhanced with domain detection is more appropriate to a sign translation task. This is a data-driven approach. What EBMT needs are a bilingual corpus and a bilingual dictionary where the latter can be constructed statistically from the corpus. Matched from the corpus, EBMT can give the same style of translations as the corpus. Our translation system is based on this approach. In addition, we can use a database search method to deal with names, phrases, and symbols related to tourists.

We start with the EBMT software. The system is used as a shallow system that can function using nothing more than sentence-aligned plain text and a bilingual dictionary. Given sufficient parallel texts, the dictionary can be extracted statistically from the corpus. In the translation process, the system looks up all matching phrases in the source-language and performs a word-level alignment on the entries containing matches to determine a (usually partial) translation. Portions of the input for which there are no matches in the corpus do not generate a translation.

We tested the EBMT based method using 50 randomly selected signs from our database, assuming perfect sign recognition in the test. We first tested the system using a

Chinese-English dictionary from the Linguistic Data Consortium (LDC), and a statistical dictionary built from the HKLC (Hong Kong Legal Code) corpus. As a result, we only obtained about 30% reasonable translations. We then trained the system with a small corpus of 670 pairs of bilingual sentences [6], the accuracy is improved from 30% to 52% on 50 test signs. It is encouraging that the improvement in EBMT translations is obtained without requiring any additional knowledge resources. We will further evaluate the improvement of the translation quality, when we combine words into larger chunks on both sides of the corpus.

Figure 5 illustrates the error analysis of the translation result. It is interesting to note that 40% of errors come from mis-segmentation of Chinese words. Obviously, there is a significant room for improvement in word segmentation. The improvements in proper name and domain detection can also enhance the accuracy of the system significantly.

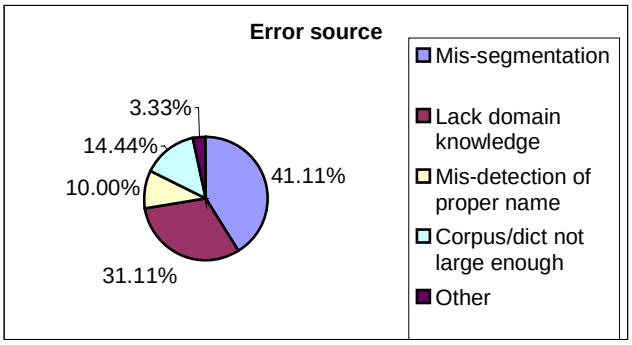

Figure 5. Error analysis of the translation experiment.

5. CONCLUSIONS

We have presented a system for automatic sign detection and translation. Automatic sign detection and translation can assist international tourists to overcome language barriers, and help visually handicapped users increase their environmental awareness. We have proposed a framework for automatic detection of signs from natural scenes. This framework considers critical challenges in sign extraction and can extract signs robustly under different conditions (image resolution, camera view angle, and lighting). We have extended EBMT technologies to Chinese sign translation and demonstrated its effectiveness and efficiency. There is still a room to improve the sign detection and translation methods. For example, it is possible to eliminate false sign detections by combining sign detection with OCR. The confidence of the sign detection can be improved by incorporating the OCR engine in an early stage. We are also interested in enhancing translation quality even with an imperfect OCR system.

ACKNOWLEDGEMENTS

We would like to thank Dr. Jiang Gao for his contribution to this project in the early stage. We would like to thank Dr. Ralf Brown and Dr. Robert Frederking for providing initial EBMT software. We would also like to thank other members in the Interactive Systems Labs for their inspiring discussions and support. This research is partially supported by DARPA under TIDES project.

REFERENCES

[1] Brown, R.D., Example-based machine translation in the pangloss system. *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 169-174, 1996.

[2] Cui, Y. and Huang, Q., Character Extraction of License Plates from Video. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 502-507, 1997.

[3] Gao, J. and Yang, J., "An Adaptive Algorithm for Text Detection from Natural Scenes," Proceedings of Computer Vision and Pattern Recognition (CVPR 2001).

[4] Hutchins, John W., Machine Translation: Past, Present, Future, Ellis Horwood Limited, England, 1986.

[5] Jain, A.K. and Yu, B., Automatic text location in images and video frames. *Pattern Recognition*, vol. 31, no. 12, pp. 2055-2076, 1998.

[6] Kubler, Cornelius C., "Read Chinese Signs". Published by Chheng & Tsui Company, 1993.

[7] Li, H. and Doermann, D., Automatic Identification of Text in Digital Video Key Frames, *Proceedings of IEEE International Conference of Pattern Recognition*, pp. 129-132, 1998.

[8] Lienhart, R., Automatic Text Recognition for Video Indexing, *Proceedings of ACM Multimedia 96*, pp. 11-20, 1996.

[9] Ohya, J., Shio, A., and Akamatsu, A., Recognition of characters in scene images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 214-220, 1994.

[10] Sato, T., Kanade, T., Hughes, E.K., and Smith, M.A., Video OCR for digital news archives. *IEEE Int. Workshop on Content-Based Access of Image and Video Database*, 1998.

[11] Taylor, M.J., Zappala, A., Newman, W.M., and Dance, C.R., Documents through cameras, *Image and Vision Computing*, vol. 17, no. 11, pp. 831-844, 1999.

[12] Watanabe, Y., Okada, Y., Kim, Y.B., and Takeda, T., Translation camera, *Proceedings Fourteenth International Conference on Pattern Recognition*, pp. 613-617, 1998.

[13] Wu, V., Manmatha, R., and Riseman, E.M., Textfinder: an automatic system to detect and recognize text in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 11, pp. 1224-1229, 1999.

[14] Yang, J., Yang, W., Denecke, M., and Waibel, A., Smart sight: a tourist assistant system. *Proceedings of Third International Symposium on Wearable Computers*, pp. 73-78. 1999.

[15] Yang, J., Gao, J., Zhang, Y., Waibel, A., Towards Automatic Sign Translation, Proceedings of Human Language Technology 2001.

[16] Yang, J., Gao, J., Zhang, Y., Chen, X., Waibel, A., An automatic sign recognition and translation system, Proceedings of Perceptual User Interface Workshop (PUI2001).

[17] Zhong Y., Karu, K., and Jain, A.K., Locating Text in Complex Color Images, Pattern Recognition, vol. 28, no. 10, pp. 1523-1536, 1995.