Towards Automatic Sign Translation

Jie Yang, Jiang Gao, Ying Zhang, Alex Waibel
Interactive Systems Laboratory
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{yang+,jgao,joy,waibel}@cs.cmu.edu

ABSTRACT

Signs are everywhere in our lives. They make our lives easier when we are familiar with them. But sometimes they also pose problems. For example, a tourist might not be able to understand signs in a foreign country. In this paper, we present our efforts towards automatic sign translation. We discuss methods for automatic sign detection. We describe sign translation using example based machine translation technology. We use a user-centered approach in developing an automatic sign translation system. The approach takes advantage of human intelligence in selecting an area of interest and domain for translation if needed. A user can determine which sign is to be translated if multiple signs have been detected within the image. The selected part of the image is then processed, recognized, and translated. We have developed a prototype system that can recognize Chinese signs input from a video camera which is a common gadget for a tourist, and translate them into English text or voice stream.

Keywords

Sign, sign detection, sign recognition, sign translation.

1. INTRODUCTION

Languages play an important role in human communication. We communicate with people and information systems through diverse media in increasingly varied environments. One of those media is a sign. A sign is something that suggests the presence of a fact, condition, or quality. Signs are everywhere in our lives. They make our lives easier when we are familiar with them. But sometimes they also pose problems. For example, a tourist might not be able to understand signs in a foreign country. Unfamiliar language and environment make it difficult for international tourists to read signs, take a taxi, order food, and understand the comments of passersby.

At the Interactive Systems Lab of Carnegie Mellon University, we are developing technologies for tourist applications [12]. The systems are equipped with a unique combination of sensors and software. The hardware includes computers, GPS receivers, lapel microphones and earphones, video cameras and head-mounted displays. This combination enables a multimodal interface to take advantage of speech and gesture inputs to provide assistance for tourists. The software supports natural language processing, speech recognition, machine translation, handwriting recognition and multimodal fusion. A vision module is trained to locate and read written language, is able to adapt to new environments, and is able to interpret intentions offered by the user, such as a spoken clarification or pointing gesture.

In this paper, we present our efforts towards automatic sign translation. A system capable of sign detection and translation would benefit three types of individuals: tourists, the visually handicapped and military intelligence. Sign translation, in conjunction with spoken language translation, can help international tourists to overcome these barriers. Automatic sign recognition can help us to increase environmental awareness by effectively increasing our field of vision. It can also help blind people to extract information. A successful sign translation system relies on three key technologies: sign extraction, optical character recognition (OCR), and language translation. Although much research has been directed to automatic speech recognition, handwriting recognition, OCR, speech and text translation, little attention has been paid to automatic sign recognition and translation in the past. Our current research is focused on automatic sign detection and translation while taking advantage of OCR technology available. We have developed robust automatic sign detection algorithms. We have applied Example Based Machine Translation (EBMT) technology [1] in sign translation.

Fully automatic extraction of signs from the environment is a challenging problem because signs are usually embedded in the environment. Sign translation has some special problems compared to a traditional language translation task. They can be location dependent. The same text on different signs can be treated differently. For example, it is not necessary to translate the meanings for names, such as street names or company names, in most cases. In the system development, we use a user-centered approach. The approach takes advantage of human intelligence in selecting an area of interest and domain for translation if needed. For example, a user can determine which sign is to be translated if multiple signs have been detected within the image. The selected part of the image is then processed, recognized, and translated, with the translation displayed on a hand-held wearable display, or a head mounted display, or synthesized as a voice output message over the earphones. By focusing only on the information of interest and providing domain knowledge, the approach provides a flexible method for sign translation. It can enhance the robustness of sign recognition and translation, and speed up the recognition and translation process. We have developed a prototype system that can recognize Chinese sign input from a video camera which is a common gadget for a tourist, and translate the signs into English text or voice stream.

The organization of this paper is as follows: Section 2 describes challenges in sign recognition and translation. Section 3 discusses methods for sign detection. Section 4 addresses the application of EBMT technology into sign translation. Section 5 introduces a prototype system for Chinese sign translation. Section 6 gives experimental results. Section 7concludes the paper.

2. PROBLEM DESCRIPTION

A sign can be a displayed structure bearing letters or symbols, used to identify or advertise a place of business. It can also be a posted notice bearing a designation, direction, or command. Figure 1 and Figure 2 illustrate two examples of signs. Figure 1 shows a Russian sign completely embedded in the background. Figure 2 is a sign that contains German text with no verb and article. In this research, we are interested in translating signs that have direct influence upon a tourist from a different country or culture. These signs, at least, include the following categories:

- Names: street, building, company, etc.
- Information: designation, direction, safety advisory, warning, notice, etc.
- Commercial: announcement, advertisement, etc.
- Traffic: warning, limitation, etc.
- Conventional symbol: especially those are confusable to a foreign tourist, e.g., some symbols are not international.

Fully automatic extraction of signs from the environment is a challenging problem because signs are usually embedded in the environment. The related work includes video OCR and automatic text detection. Video OCR is used to capture text in the video images and recognize the text. Many video images contain text contents. Such text can come from computer-generated text that is overlaid on the imagery (e.g., captions in broadcast news programs) or text that appears as a part of the video scene itself (e.g., a sign outside a place of business, or a post). Location and recognition of text in video imagery is challenging due to low resolution of characters and complexity of background. Research in video OCR has mainly focused on locating the text in the image and preprocessing the text area for OCR [4][6][7][9][10]. Applications of the research include automatically identifying the contents of video imagery for video index [7][9], and capturing documents from paper source during reading and writing [10]. Compared to other video OCR tasks, sign extraction 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. In addition, sign extraction has to be implemented in real time using limited resources.

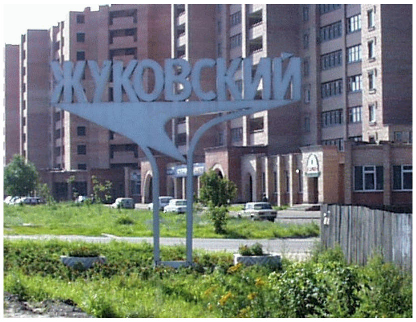


Figure 1 A sign embedded in the background

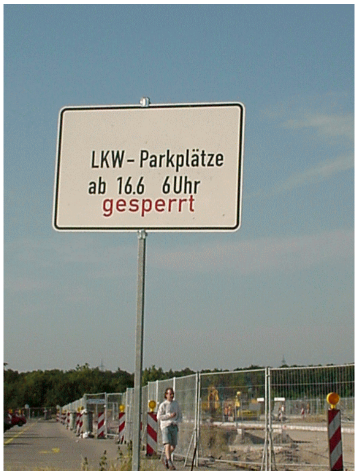


Figure 2 A German sign

Sign translation requires sign recognition. A straightforward idea is to use advanced OCR technology. Although OCR technology works well in many applications, it requires some improvements before it can be applied to sign recognition. At current stage of the research, we will focus our research on sign detection and translation while taking advantage of state-of-the-art OCR technologies.

Sign translation has some special problems compared to a traditional language translation task. The function of signs lead to the characteristic of the text used in the sign: it has to 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 and insufficient information from the text to resolve the ambiguities which are related to the environment of the sign.

We assume that a tourist has a video camera to capture signs into a wearable or portable computer. The procedure of sign translation is as follows: capturing the image with signs, detecting signs in the image, recognizing signs, and translating results of sign recognition into target language.

3. AUTOMATIC SIGN DETECTION

Fully automatic extraction of signs from the environment is very difficult, because signs are usually embedded in the environment. There are many challenges in sign detection, such as variation, motion and occlusion. We have no control in font, size, orientation, and position of sign texts. Originating in 3-D space, text on signs in scene images can be distorted by slant, tilt, and shape of objects on which they are found [8]. In addition to the horizontal left-to-right orientation, other orientations include vertical, circularly wrapped around another object, slanted, sometimes with the characters tapering (as in a sign angled away from the camera), and even mixed orientations within the same text area (as would be found on text on a T-shirt or wrinkled sign). Unlike other text detection and video OCR tasks, sign extraction is in a more dynamic environment. The user's movement can cause unstable input images. Furthermore, the quality of the video input is poorer than that of other video OCR tasks, such as detecting captions in broadcast news programs, because of low quality of equipment. Moreover, sign detection has to be real-time using a limited resource. Though automatic sign detection is a difficult task, it is crucial for a sign translation system.

We use a hierarchical approach to address these challenges. We detect signs at three different levels. At the first level, the system performs coarse detection by extracting features from edges, textures, colors/intensities. The system emphasizes robust detection at this level and tries to effectively deal with the different conditions such as lighting, noise, and low resolution. A multi-resolution detection algorithm is used to compensate different lighting and low contrasts. The algorithm provides hypotheses of sign regions for a variety of scenes with large variations in both lighting condition and contrast. At the second level, the system refines the initial detection by employing various adaptive algorithms. The system focuses on each detected area and makes elaborate analysis to guarantee reliable and complete detection. In most cases, the adaptive algorithms can lead to finding the regions without missing any sign region. At the third level, the system performs layout analysis based on the outcome from the previous levels. The design and layout of signs are language and culture dependent. For example, many Asia languages, such as Chinese and Japanese, have two types of layout: the horizontal and the vertical. The system provides considerable flexibility to allow the detection of slanted signs and signs with non-uniform character sizes.

4. SIGN TRANSLATION

Sign translation has some special problems compared to a traditional language translation task. Sign translation depends not only on domain but also on functionality of the sign. The same text on different signs can be treated differently. In general, the text used in the sign is short and concise. For example, the average length of each sign in our Chinese sign database is 6.02 Chinese characters. The lexical mismatch and structural mismatch problems become more severe for sign translation because shorter words/phrases are more likely to be ambiguous and there isn't sufficient information from the text to resolve the ambiguities which are related to the environment of the sign. For example, in order to make signs short, abbreviations are widely used in signs, e.g., 寄研所 (/ji yan suo/) is the abbreviation for 寄生虫研究所 , (/ji sheng chong yan jiu suo/ institute of parasites), such abbreviations are difficult, if not impossible, even for a human to understand without knowledge of the context of the sign. Since designers of signs always assume that readers can use the information from other sources to understand the meaning of the sign, they tend to use short words. e.g. in sign 慢行 (/man xing/, drive slowly), the word 行 (/xing/, walk, drive) is ambiguous, it can mean 行走 (/xing zou/ "move of human," walk) or 行驶 "move of a car," drive). The human reader can understand the meaning if he knows it is a traffic sign for cars, but without this information, MT system cannot select the correct translation for this word. Another problem in sign is structural mismatch. Although this is one of the basic problems for all MT systems, it is more serious in sign translation: some grammatical functions are omitted to make signs concise. Examples include: (1) the subject "we" is omitted in 礼貌待客 (/li mao dai ke/, treat customers politely); (2) the sentence is reordered to emphasize the topic: rather than saying 请将包装纸投入垃圾箱 (/qing jiang bao zhuang zhi tou ru la ji xiang/, please throw wrapping paper into the garbage can), using 包装纸请投入垃圾箱 (/bao zhuang zhi qing tou ru la ji xiang/, wrapping paper, please throw it into the garbage can) to highlight the "wrapping paper." With these special features, sign translation is not a trivial problem of just using existing MT technologies to translate the text recognized by OCR module.

Although a knowledge-based MT system works well with grammatical sentences, 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 that appears frequently in signs.

We can use a database search method to deal with names, phrases, and symbols related to tourists. Names are usually location dependent, but they can be easily obtained from many information sources such as maps and phone books. Phrases and symbols related to tourists are relative fixed for a certain country. The database of phrases and symbols is relatively stable once it is built

We propose to apply Generalized Example Based Machine Translation (GEBMT) [1][2] enhanced with domain detection to a sign translation task. This is a data-driven approach. What EBMT needs are a set of bilingual corpora each for one domain and a bilingual dictionary where the latter can be constructed statistically from the corpora. Matched from the corpus, EBMT can give the same style of translations as the corpus. The domain detection can be achieved from other sources. For example, shape/color of the sign and semantics of the text can be used to choose the domain of the sign.

We will start with the EBMT software [1]. The system will be used as a shallow system that can function using nothing more than sentence-aligned plain text and a bilingual dictionary; and given sufficient parallel text, the dictionary can be extracted statistically from the corpus. In a translation process, the system looks up all matching phrases in the source-language half of the parallel corpus 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. Because the EBMT system does not generate translations for 100% of its input text, a bilingual dictionary and phrasal glossary are used to fill any gaps. Selection of the "best" translation is guided by a trigram model of the target language and a chart table [3].

5. A PROTOTYPE SYSTEM

We have developed a prototype system for Chinese sign recognition and translation. Figure 3 shows the architecture of the prototype system. A user can interactively involve sign recognition and translation process when needed. For example, a user can select the area of interest, or indicate that the sign is a street name. The system works as follows. The system captures the sign in a natural background using a video camera. The system then automatically detects or interactively selects the sign region. The system performs sign recognition and translation within the detected/selected region. It first preprocesses the selected region, binarizes the image to get text or symbol, and feeds the binary image into the sign recognizer. OCR software from a third party is used for text recognition. The recognized text is then translated into English. The output of the translation is fed to the user by display on screen or synthesized speech. Festival, a general purpose multi-lingual text-to-speech (TTS) system is used for speech synthesis.

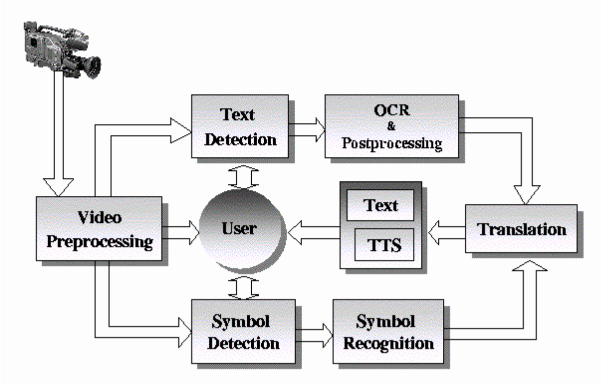


Figure 3 Architecture of the prototype system

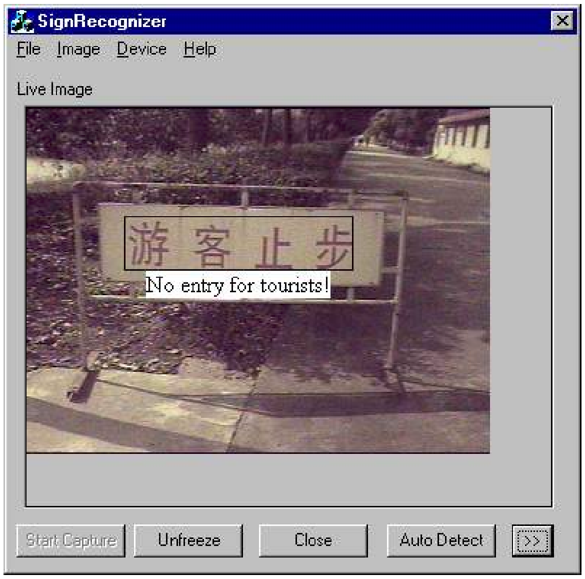


Figure 4 The interface of the prototype system

An efficient user interface is important to a user-centered system. Use of interaction is not only necessary for an interactive system, but also useful for an automatic system. A user can select a sign from multiple detected signs for translation, and get involved when automatic sign detection is wrong. Figure 4 is the interface of the system. The window of the interface displays the image from a video camera. The translation result is overlaid on the location of the sign. A user can select the sign text using pen or mouse anywhere in the window.

6. EXPERIMENTAL RESULTS

We have evaluated the prototype system for automatic sign detection and translation. We have built a Chinese sign database with about 800 images taken from China and Singapore. We have tested the automatic detection module using 50 images randomly selected from the database. Table 1 shows the test result of automatic sign detection. Figure 5 and Figure 6 show examples of automatic sign detection with white rectangles indicating the sign regions. Figure 5 shows correct detection after layout analysis. Figure 6 illustrates a result with a false detection (Note the small detection box below and to the left of the larger detection).

Table 1 Test Results of Automatic Detection on 50 Chinese Signs

Detection without missing characters	Detection with false alarm	Detection with missing characters
43	12	5

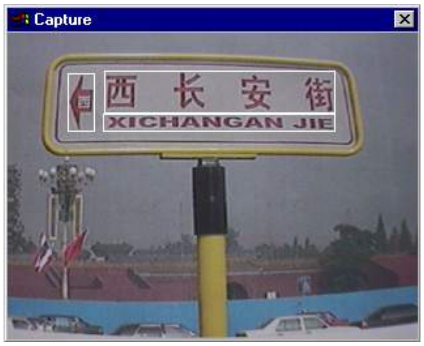


Figure 5 An example of automatic sign detection

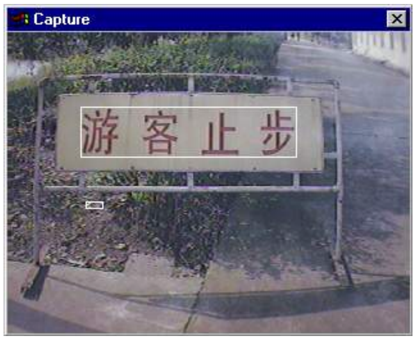
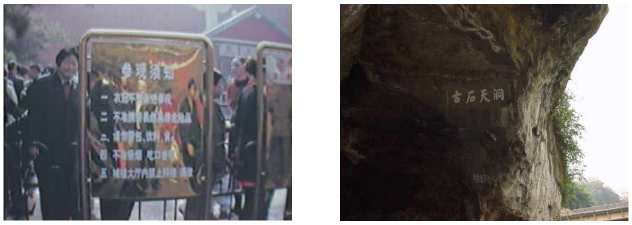


Figure 6 An example of false detection

Figure 7 illustrates two difficult examples of sign detection. The text in Figure 7(a) is easily confused with the reflective background. The sign in Figure 7(b) is embedded in the background



(a) (b)

Figure 7 Difficult examples of sign detection

We have also tested the EBMT based method. We assume perfect sign recognition in our test. We randomly selected 50 signs from our database. We first tested the system includes a Chinese-English dictionary from the Linguistic Data Consortium, and a statistical dictionary built from the HKLC (Hong Kong Legal Code) corpus. As a result, we only got about 30% reasonable translations. We then trained with a small corpus of 670 pairs of bilingual sentences [7], The accuracy is improved from 30% to 52% on 50 test signs. Some examples of errors are illustrated below:

Mis-segmentaion:

Chinese with wrong segmentation:
[各种] [车辆] [请] [绕] [行]
/ge zhong che liang qing rao xing/
Translation from MT:
All vehicles are please wind profession
Correct segmentation:
[各种] [车辆] [请] [绕行]

Translation if segmentation is correct:
All vehicles please use detour

Lack-domain information:

Chinese with segmentation:
[请勿] [动手]

/qing wu dong shou/
Please don't touch it
Translation from MT:
Please do not get to work

Domain knowledge needed to translate 动手: "start to work" in domain such as work plan and "don't touch" in domains like tourism, exhibition etc.

Proper Name:

Chinese with segmentation:
[北京] [同仁] [医院]

/bei jing tong ren yi yuan/

Beijing Tongren Hospital

Translation from MT:

Beijing similar humane hospital

同仁 is translated to the meaning of each character because it is not identified as a proper name which then should only be represented by its pronunciation.

Figure 8 illustrates error analysis of the translation module. It is interesting to note that 40% of errors come from mis-segmentation of words. There is a big room for improvement in proper word segmentation. In addition, we can take advantage of the contextual information provided by the OCR module to further improve the translation quality.

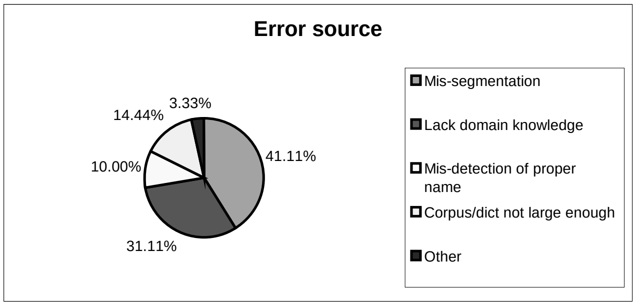


Figure 8 Error analysis of the translation module

7. CONCLUSION

We have reported progress on automatic sign translation in this paper. Sign translation, in conjunction with spoken language translation, can help international tourists to overcome language barriers. A successful sign translation system relies on three key technologies: sign extraction, OCR, and language translation. We have developed algorithms for robust sign detection. We have applied EBMT technology for sign translation. We have employed a user-centered approach in developing an automatic sign translation system. The approach takes advantage of human intelligence in selecting an area of interest and domain for translation if needed. We have developed a prototype system that can recognize Chinese signs input from a video camera which is a common gadget for a tourist, and translate them into English text or voice stream.

ACKNOWLEDGMENTS

We would like to thank Dr. Ralf Brown and Dr. Robert Frederking for providing initial EBMT software and William Kunz for developing the interface for the prototype system. 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] R.D. Brown. Example-based machine translation in the pangloss system. Proceedings of the 16th International Conference on Computational Linguistics, pp. 169-174, 1996.

[2] R.D. Brown. Automated generalization of translation examples". In Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000), p. 125-131. Saarbrücken, Germany, August 2000.

[3] C. Hogan and R.E. Frederking. An evaluation of the multi-engine MT architecture. Machine Translation and the Information Soup: *Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA '98)*, vol. 1529 of Lecture Notes in Artificial Intelligence, pp. 113-123. Springer-Verlag, Berlin, October.

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

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

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

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

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

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

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

[11] V. Wu, R. Manmatha, and E.M. Riseman, 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.

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