Text Detection and Translation from Natural Scenes

Jiang Gao, Jie Yang, Ying Zhang, and Alex Waibel

June 2001
CMU-CS-01-139

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

This research is partially supported by DARPA under TIDES project.

Keywords: Text detection, machine translation, OCR, intelligent interface, multimedia system

Abstract

We present a system for automatic extraction and interpretation of signs from a natural scene. The system is capable of capturing images, detecting and recognizing signs, and translating them into a target language. The translation can be displayed on a hand-held wearable display, or a head mounted display. It can also be synthesized as a voice output message over the earphones. We address challenges in automatic sign extraction and translation. We describe methods for automatic sign extraction. We extend example-based machine translation technology for sign translation. We use a user-centered approach in the system development. The approach takes advantage of human intelligence if needed and leverages human capabilities. We are currently working on Chinese sign translation. We have developed a prototype system that can recognize Chinese signs input from a video camera that is a common gadget for a tourist, and translate the signs either into English text or a voice stream. We have built up a database containing about 800 Chinese signs for development and evaluation. We present evaluation results and analyze errors. The sign translation, in conjunction with spoken language translation, can help international tourists to overcome language barriers. The technology can also help a visually handicapped person to increase environmental awareness.

1 Introduction

We work, live, and play in a so-called information society where 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 or even danger. For example, a tourist might not be able to understand a sign in a foreign country. A visually handicapped person can be in danger if he/she misses signs that specify warnings or hazards. In this research, we are interested in signs that have direct influence upon a tourist from a different country or culture. These signs include, at least, 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.

Although much research has been directed to automatic speech recognition, handwriting recognition, optical character recognition (OCR), speech and text translation, little attention has been paid to automatic sign recognition and translation in the past. At the Interactive Systems Laboratory of Carnegie Mellon University, we are developing technologies for automatic sign extraction and interpretation. Sign translation, in conjunction with spoken language translation, can help international tourists to overcome language barriers. It is a part of our efforts in developing a tourist assistant system (Yang, 1999a). The proposed 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.

A successful sign translation system relies on three key technologies: sign extraction, OCR, and language translation. At current stage of the research, we focus our efforts on sign detection and translation while taking advantage of state-of-the-art OCR technologies. Fully automatic extraction of signs from the environment is a challenging problem because signs are usually embedded in the environment. Compared to video OCR tasks, which is to recognize texts from video images, 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. Sign translation has some special problems compared to a traditional language translation task. In general, the text used in a sign is short and concise. Lexical mismatches and structural mismatches become more severe 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 will present technologies to address these challenges.

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 an image. The selected part of the image is then processed, recognized, and translated. 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 are currently working on Chinese sign translation. We choose to work on Chinese sign translation for several reasons. First, Chinese is a major language and very different from European (phonologic) languages. Second, a foreign tourist might have serious language barrier in China because English is not commonly used there. Third, statistics shows that more people will visit China in the future. Finally, technologies developed for Chinese sign translation can be extended to other languages. We have built up a database containing about 800 Chinese signs taken from China and Singapore. The database can be used for development and evaluation. We have developed a prototype system that can recognize Chinese signs input from a video camera that is a common gadget for a tourist, and translate the signs into either English text or a voice stream.

The organization of this report is as follows: Section 2 describes challenges in sign recognition and translation. Section 3 discusses methods for sign detection. Section 4 extends EBMT technology for sign translation. Section 5 addresses issues in system integration, and gives experimental results. Section 6 concludes the report.

2 Challenges

The procedure of automatic sign translation is as follows: capturing the image with signs, detecting signs in the image, recognizing signs, and translating results of sign recognition into a target language. We currently focus on developing technologies for automatic sign extraction and translation, while taking advantage of state-of-the-art OCR technologies.

2.1 Sign Extraction from Natural Scenes

Sign extraction is related to character or text extraction/ recognition. The previous research falls into three different areas: (1) automatic detection of text areas from general backgrounds (Cui and Huang 1997, Jain and Yu 1998, Li and Doermann 1998, Lienhart 1996, Ohya 1994, Sato 1998, Wong 2000, Wu 1999, Zhong and Jain 1995); (2) document input from a video camera under a controlled environment (Taylor 1999); and (3) enhancement of text areas for character recognition (Lienhart 1996; Ohya 1994, Sato 1998, Watanabe 1998, Wu 1999, Li and Doermann 2000).

The first area is relevant to this research. In this area, most research focus on extracting text from pictures or video, though a few are focused on character extraction from vehicle license plates (Cui and Huang 1997). Many video images contain text contents. Such texts can be part of the scene, or come from computer-generated text that is overlaid on the imagery (e.g., captions in broadcast news programs).

The existing approaches for text detection lie in the following categories:

- Edge filtering (Zhong and Jain 1995);
- Texture segmentation (Wu 1999);
- Color quantization (Jain and Yu 1998);
- Neural networks and bootstrapping (Lienhart 1996, Li and Doermann 1998).

Each of these approaches has its advantage/drawbacks concerning reliability, accuracy, computational complexity, difficulty in improvement, and implementation. 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 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. Compared with other object detection tasks, many information sources are unavailable for a sign detection task, such as:

- No motion information: signs move together with background;
- No approximate shape information: text areas assume different shapes;
- No color reflectance information: signs assume different colors.

Signs can vary in font, size, orientation, and position of sign texts, be blurred from motion, and occluded by other objects. Originating in a 3-D space, text on signs in scene images can be distorted by slant, tilt, and shape of objects on which they are found (Ohya 1994). Furthermore, unconstrained background clutters can look like signs in their appearance, causing false detection. Moreover, languages impose another level of variation in text. For example, Chinese characters are composed of many segments and the layout of Chinese characters in signs, which is based on pictographic characters, differs from the layout used in European languages. Handling Chinese characters requires a more elaborate method of modeling. Figure 1 shows an example of a Chinese sign with both vertical and horizontal layouts as well as distortion because of warping.

Figure 1. An example of a sign with different layout and deformation.

2.2 Sign Interpretation

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 function 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 difficulty 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) (Brown 1996) 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 database search method to deal with names, phrases, and symbols related to tourists.

3 Sign Extraction

We propose a new sign extraction algorithm based on an adaptive search strategy. Compared with the existing text detection algorithms, this new adaptive strategy can better handle the dynamics of sign extraction in natural scenes. We embed this adaptive strategy in a hierarchical algorithm structure, with different emphases at each layer. The sign extraction algorithm consists of:

- A multi-resolution edge detection algorithm;
- Adaptive searching in the neighborhood of initial candidate regions;
- Layout analysis of the detected sign regions.

In this three layers structure, the first layer detects possible sign regions. Once having the initial candidates, the system uses the local information, such as color and shape, provided by the first layer to adaptively search the neighborhood of candidates. This adaptive algorithm is in the second layer. Compared with the existing algorithms, the adaptive search strategy is novel. The advantage of this strategy is that we can use the locality and color information provided by the first layer to refine the detection results, rather than searching over entire color space in the whole image. Since a Chinese sign is composed of characters, and a character is typically multi-segments, Chinese signs have more complex layout than that of signs in European languages. In order to achieve high detection rate, an elaborate layout analysis algorithm is essential. This algorithm is in the third layer. Layout constraints are also used in the adaptive algorithm in the second layer to find possible sign regions.

The hierarchical structure of the proposed algorithm framework is shown in Figure 2. Contribution of the 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.

Figure 3 shows an example of a detection result. White rectangles indicate detected sign regions. In our Chinese sign 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. In Figure 3, the system captures two signs with different backgrounds from a non-favorite angle. The system still could successfully detect all the signs. This demonstrates that the detection framework provides considerable flexibility to allow the detection of slanted signs and signs with non-uniform character sizes. In the following sub-sections, we will describe each layer of the framework in detail.

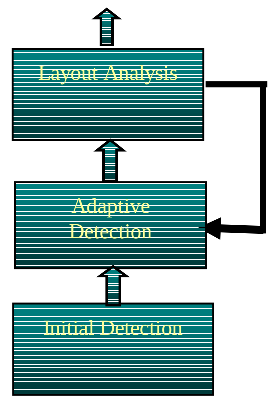

Figure 2. The block diagram of sign extraction algorithm.

Figure 3. An example of sign extraction result.

3.1 Multi-Resolution Edge Detector

In the first layer of the framework, an edge detection algorithm is used to obtain the initial candidates of sign regions. Since lighting and contrast vary dramatically in a natural environment, the detection algorithm needs to perform reliably under different circumstances. We utilize a multi-resolution approach to compensate these variations and the noise of edge detection algorithm. We apply edge detection algorithm using different scale parameters, and then fuse the results from different resolutions.

Figure 4 illustrates the algorithm. The two signs (in the lower left and upper right position of the image) have different contrast and lighting conditions, but can be optimally detected using edge detection algorithm under different scales. The sign in the left can be segmented in scale 1, but cannot be detected in scale 2; while part of the sign on the right can be detected in scale 2, but is difficult to extract in scale1. The integration result is obtained by combining the detection results from different scales.

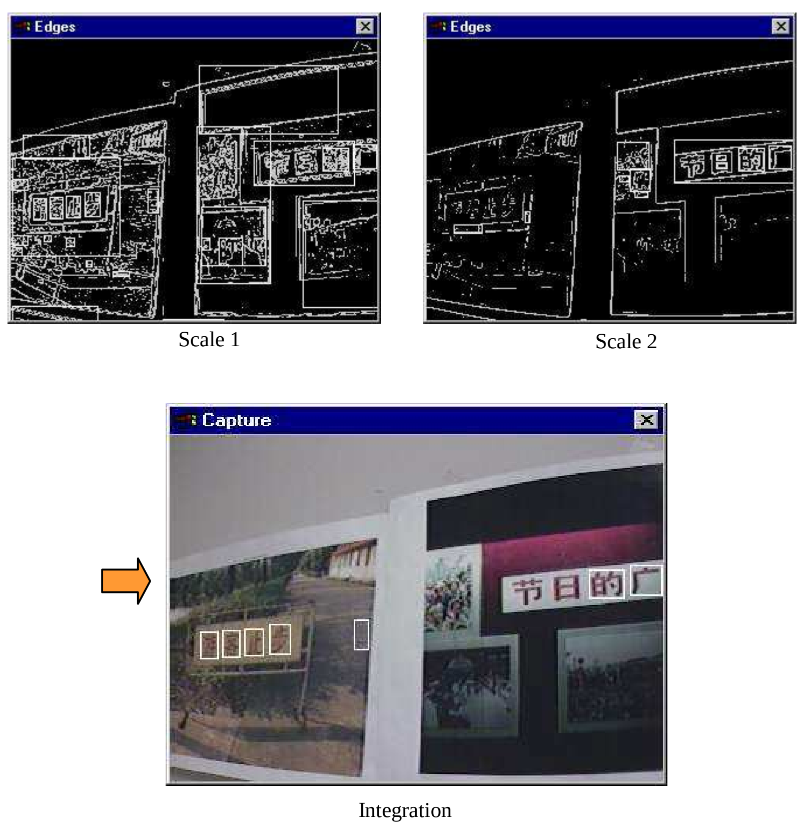

Figure 4. Initial sign extraction from different scales.

3.2 Adaptive Searching and Modeling

3.2.1 Adaptive searching based on layout syntax

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. We discuss these issues in some more details in the following.

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. As shown in Figure 3, the characters in the same sign tend to be in similar appearance, such as using the same font, color, and size. 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.

3.2.2 Color modeling: Gaussian mixtures modeling

Now we have the searching area and direction control, and also have criteria to discriminate different sign areas. The next step is to extract the characters from the searching areas. Character extraction can be a simple case or a complex case.

In a simple case, the color or intensity of a sign do not change significantly. In this case, we can use color information of the initially detected text to extract characters with similar attributes from the searched areas, based on the layout syntax. In some situations, however, colors within a sign region change dramatically due to lighting sources or the quality of the sign. In such a case, the color information may fail to provide a clue on the sign. In this complex case, we can still use the approximate location, direction, and size of the sign characters to extract the characters from the background, but we cannot use the color information of the initially detected text directly.

Most OCR algorithms separate text from background using a “binarization” process. The hidden assumption of such approach is that both color of background and text are in uniform distributions. This assumption is often invalid in sign extraction. We use a Gaussian mixtures model to segment signs.

By using a Gaussian mixtures color model, we can model both the text region and background using arbitrary number of basis functions (clusters). The Gaussian mixtures model utilized is as follows: The probability distribution of D-dimensional color vectors x is represented as a weighted mixture of K basis functions (components) as:

$$p(x) = \sum_{k=1}^{K} w_k \cdot \alpha_k(x) \,, \tag{1}$$

(a)

(b) (c)

Figure 5. Adaptive character extraction using color modeling. (a) Original sign, (b) Color space modeled by two Gaussian mixtures, (c) Color space modeled by three Gaussian mixtures.

w_k is the mixture weight. The basis functions $\alpha_k(x)$ are chosen to be Gaussians of the form:

$$\alpha(x) = \frac{1}{(2\pi)^{D/2}|C|^{1/2}} \exp\{-\frac{1}{2}(x - m_x)^T C^{-1}(x - m_x)\} . \tag{2}$$

m_x and C are the expectation and covariance matrix of D dimensional vectors x, respectively. $|C|$ is the determinant of C.

In Gaussian mixtures color modeling, we use the basis functions to represent regions of different color property. Given the model parameters, the probability that the region k being responsible for generating pixel with color vector x can be computed as:

$$\gamma_k(x) = \frac{w_k \cdot \alpha_k(x)}{p(x)} . \tag{3}$$

An EM algorithm is utilized to determine the optimal parameters C_k and w_k for this Guanssian mixtures model, given the number K of Gaussian mixtures in the region. The EM algorithm maximizes the likelihood:

$$L = \prod_{n=1}^{N} p(x) . \tag{4}$$

N is the number of pixels in search region. The model parameters are initialized using fast clustering algorithm.

A crucial problem of this method is how to determine the number of Gaussian mixtures that are appropriate to model the color space. As shown in Figure 5, due to the lighting and noise conditions in natural scenes, the color compositions of signs differ in different locations. Inappropriate selection of Gaussian mixtures numbers will result in errors in text detection. We solve this problem by taking into account of the layout syntax. We determine the number of Gaussian mixtures by considering if the characters can be extracted from the background, and the characters have to satisfy the layout syntax as described in section 3.2.1. The algorithm is given in Figure 6.

Let $K = 2$;

{

Extract character/text regions based on K Gaussian mixtures color modeling, for each search region $SR(\cdot)$, until there is no $SR_K(c, p, s, t_0)$, such that: $SR_K(c, p, s, t_0) \supset SR0_K(c, p, s, t_0)$;

If $SR_K(c, p, s, t) \supset SR_{K-1}(c, p, s, t)$

FLAG =1;

If ($K < K_{\max}$) AND (FLAG =1)

$K = K + 1$;

Else

$K_{final} = K$, break;

}

Figure 6. Algorithm for determining the number of Gaussian mixtures.

The number of Gaussian mixtures can change in each adaptation area. Figure 5 shows an example of sign extraction using the Gaussian mixture models. We are interested in extracting the sign from the area within the white rectangle. Figure 5 (a) is the original sign, and Figure 5 (b) and Figure 5 (c) are segmentation results using two and three Gaussian mixtures, respectively. The rightmost character is confused with the background if using only two Gaussin mixtures, but can be extracted using three Gaussian mixtures.

3.3 Layout Analysis

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. Figure 8 is an example of Chinese sign. Each character in the sign is composed of several connected sub-components, sometimes the sub-components align in the same way as characters in a sign. This case is very common in Chinese signs. However, such a sign poses a big problem to automatic sign layout analysis: how can a system know if a text region is a character or only segment of a character, without recognition of the whole text area? We use layout analysis techniques, which utilize various heuristics, to deal with this problem. Examples of these criteria include:

1. If the components are near to each other;
2. If the components satisfy certain aspect ratios;
3. After assigning character labels to individual components, which assignment utilize most of the components, and "most like" a layout of sign text.

These criteria can be formulated as (definition of symbols are given in Figure 7) :

1. The components should be near to each other:

$$Dist(i, j) / \max(Width(i), Width(j)) < C1. \tag{5}$$

2. The aspect ratios (AP) of the components fall in a certain range:

$$C2 < AP_i < C3. \tag{6}$$

3. After assigning character labels to individual components, The components should align with each other:

$$Align(i, j) > C5, \tag{7}$$

$$Align(i, j) = C4 \cdot \frac{Ol(i, j)}{\max(Height_i, Height_j)}, \tag{8}$$

and utilize most of the components. In (8) $Ol(i, j)$ is the length of overlaid part of the component i and component j.

In our system, the criteria are designed to allow tolerance to considerable slanting of sign images, non-uniform character sizes, and images captured from unfavorable angles.

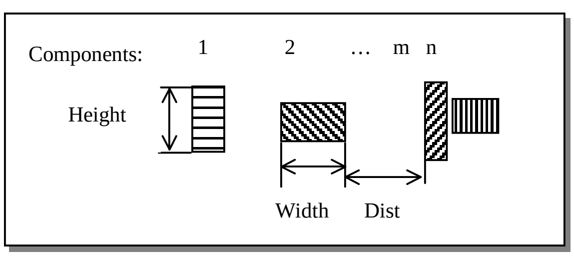

Figure 7. Layout relation: definitions.

In most cases, applying these heuristic rules can find the true sign areas correctly, though there are some situations where several sub-components of the characters in a sign will also be considered as a sign area. Since there do exist some cases where a big sign contains a smaller one in it, we decide that this type of errors is to be solved using human-computer interaction described in section 5.

Figure 8. Chinese characters are multi-segment.

Figure 9. A confusing case for layout analysis.

Layout analysis is still a challenging problem in some special cases. Figure 9 gives a confusing case of Chinese sign layout analysis. Without recognition and non-local decision, it's difficult to determine the layout accurately.

4 Sign Interpretation

We start with the EBMT software (Brown 1996, Brown 1999). 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 (Brown 1997). In the 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 a "best"' translation is guided by a trigram model of the target language and a chart table (Hogan and Frederking 1998). In this work, we extended the EBMT system to handle Chinese input by adding support for the two-byte encoding used for the Chinese text (GB-2312).

The segmentation of Chinese words from character sequences is important for translation of Chinese signs, because the meaning of a Chinese sentence is based on words, but there is no explicit tags around Chinese words. A module for Chinese word segmentation is included in the system. This segmentor uses a word-frequency list to make segmentation decisions.

We tested the EBMT based method using randomly selected 50 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 (Kubler 1993), 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.

We have not yet taken full advantage of the features of the EBMT software. In particular, it supports equivalence classes that permit generalization of the training text into templates for improved coverage. We intend to test automatic creation of equivalence classes from the training corpus (Brown 2000) in conjunction with other improvements reported herein. In addition, we will take advantage of domain information to further improve the translation quality.

5 System Integration

5.1 System Structure

The system structure is shown in Figure 10. A video camera is used to capture the scene images. The video stream or picture is then input to the sign extraction module to compress the information and only focus on the text areas. The sign extraction results are displayed to allow user interaction with the system. The user interface of the system is explained in detail in section 5.2.

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. The system presently translates Chinese signs to English. In fact, under the Interlingua (Hutchins 1986) framework, both the source and target languages can be expanded extensively. Such extension makes a system work for multiple languages.

In addition to visual output, the system has audio output. We use Festival (Black 1998) for speech synthesis. Festival is a general multi-lingual speech synthesis system developed at Center for Speech Technology Research, University of Edinburgh. Festival is a full text-to-speech (TTS) system with various APIs, and an extensive environment for development and research of speech synthesis techniques. For more detailed information, see (Black 1998).

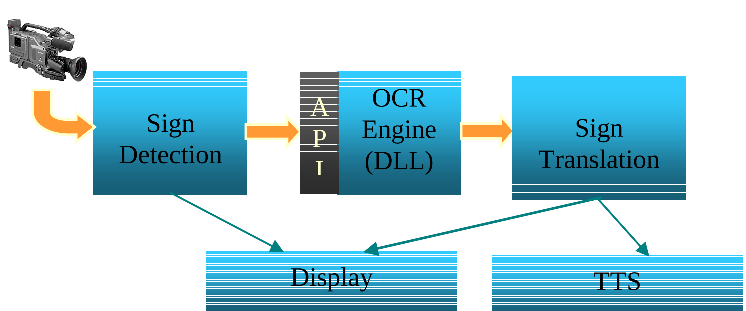

Figure 10. System Architecture.

5.2 User Interface of 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 sign regions extracted automatically. 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.

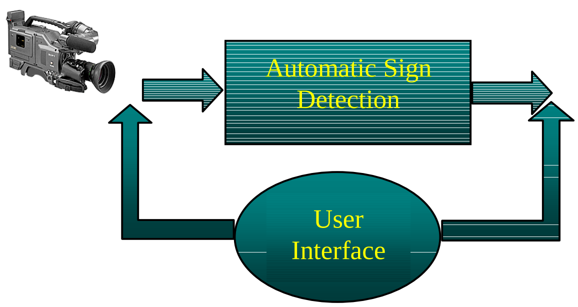

Figure 11. The function of the user interface.

The function of the user interface is summarized in Figure 11. Figure 12 is a screen shot of the current user interface. On the interface, the "unfreeze" button tells the system to input the next scene from a video camera. A user can directly draw on the captured image around the interested areas for recognition/translation, or simply touch on the "Auto Detection" button to let the system do the detection, then click the detected areas to confirm. The black rectangle indicates the detected sign region. The translation result is overlaid on the image near the location of the sign. Figure 12 shows an example of automatic detection and translation.

Figure 13 is an intermediate result of preprocessing/binarization for OCR module, and translation.

Figure 12. Screen shot of the user interface.

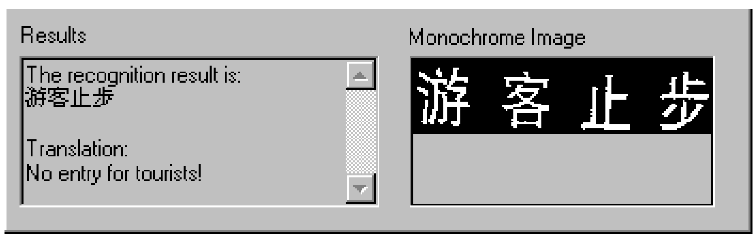

Figure 13. Preprocessing/binarization for OCR.

5.3 System Evaluation

We have evaluated the system for sign extraction and translation. We tested our sign extraction algorithm using 50 sign images randomly selected 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. Test results of automatic detection on 50 Chinese signs

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

We have tested robustness of the detection algorithm by changing conditions, such as image resolution, camera view angle, and lighting conditions. The algorithm 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. In addition, the algorithms can also effectively extract signs in other languages as shown Figure 20 and Figure 21. This demonstrates the applicability of the sign extraction algorithms to other languages.

Some results of sign detection are given in Figure 16 – Figure 21. The rectangles indicate the detected sign regions. Figure 16 and Figure 17 give examples of layout analysis. Figure 18 and Figure 19 are two examples of sign detection under slant and varying lighting conditions. Figure 20 is an example of sign detection in low-resolution image, and Figure 21 is detection of handwritings. Both examples in Figure 20 and Figure 21 are in English.

There are several ways to further improve the sign detection accuracy. For example, it is possible to eliminate false detection by combining sign detection with OCR. The confidence of the sign extraction system can be improved by incorporating the OCR engine in an early stage. Figure 14 shows the idea of using OCR to verify the detected sign regions so that the system can eliminate or reduce false detections.

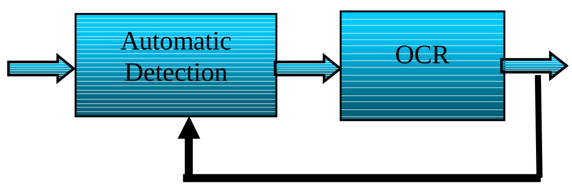

Figure 14. Integration of OCR and sign detection.

We also tested the EBMT based sign translation method. We assumed perfect sign recognition in our tests. Some examples of sign translation from the randomly selected 50 signs are given in Table 2. These examples are selected from reasonable translation results.

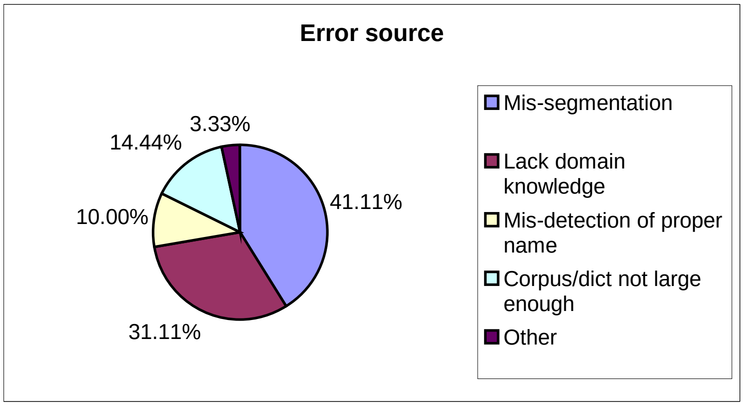

Figure 15. Error analysis of the translation experiment.

Figure 15 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 accuracy of the system largely.

Table 2. Examples of translation from Chinese to English

Original Chinese	*English Translation*
行人出口	Pedestrian exit
儿童医院	Child hospital
查询电话	Inquiry telephony
各种车辆	Various kinds vehicle
古生物化石陈列室	Old living creature fossil exhibit room

6 Conclusions

We present an automatic sign extraction and interpretation system in this report. Sign translation can assist international tourists to overcome language barriers. The technology can also help a visually handicapped person to increase environmental awareness. Our research is currently focused on automatic sign extraction and translation while taking advantage of state-of-the art OCR technology. We have proposed a framework for automatic extraction of signs from natural scenes. The framework considers critical challenges in sign extraction and can extract signs robustly under different conditions, such as different image resolutions, camera viewing angles, and lighting. We have extended EBMT method to sign translation and demonstrated its effectiveness and efficiency. We have successfully applied a user-centered approach in developing a sign translation system. The system takes advantage of human intelligence if needed and leverages human capabilities by providing sign recognition and translation services. We have developed a prototype system and evaluated the effectiveness of the proposed algorithms. We are further improving the robustness and accuracy of our system. We are particularly interested in enhancing translation quality, including translation for an imperfect OCR system.

Acknowledgements

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] Black, A.W., Taylor, P., and Caley, R., Festival, *www.cstr.ed.ac.uk/projects/festival.html*, The Centre for Speech Technology Research (CSTR) at the University of Edinburgh, 1998.

[2] Brown, R.D., Adding Linguistic Knowledge to a Lexical Example-Based Translation System. *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pp. 22-32, Chester, England, August, 1999.

[3] Brown, R.D., Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation. *Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-97)*, pp. 111-118, Santa Fe, New Mexico, July, 1997.

[4] Brown, R.D., Automated Generalization of Translation Examples. *Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000)*, pp. 125-131, 2000.

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

[6] 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.

[7] Hogan, C. and Frederking, R.E., 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.

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

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

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

[11] 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.

[12] Li, H. and Doermann, D., Superresolution-Based Enhancement of Text in Digital Video. *ICPR*, pp. 847-850, 2000.

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

[14] 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.

[15] 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.

[16] 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.

[17] Waibel, A., Interactive Translation of Conversational Speech, *Computer*, vol. 29, no. 7, 1996.

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

[19] Wong, E. K. and Chen, M., A Robust Algorithm for Text Extraction in Color Video, *Proceedings of IEEE Int. Conference on Multimedia and Expo (ICME2000)*, 2000.

[20] 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.

[21] 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. 1999a.
[22] Yang, J., Zhu, X., Gross, R., Kominek, J., Pan, Y., and Waibel, A., Multimodal People ID for a Multimedia Meeting Browser, *Proceedings of ACM Multimedia 99*, 1999b.
[23] Zhong Y., Karu, K., and Jain, A.K., Locating Text in Complex Color Images, Pattern Recognition, vol. 28, no. 10, pp. 1523-1536, 1995.

Figure 16. An example of sign detection: layout analysis I.

Figure 17. An example of sign detection: layout analysis II.

Figure 18. An example of sign detection: lighting and slant I.

Figure 19. An example of sign detection : slant, with false alarms.

Figure 20. An example of sign detection: handwritting.

Figure 21. An example of sign detection: low resolution.