A PDA-based Sign Translator

Jing Zhang[1], Xilin Chen[2], Jie Yang[2], Alex Waibel[1, 2]

[1] *Mobile Technologies, LLC*

[2] *Interactive Systems Labs, School of Computer Science, Carnegie Mellon University*

jingzhang@computer.org, {xlchen, yang+, ahw}@cs.cmu.edu

Abstract

In this paper, we propose an effective approach for a PDA-based sign system, and it presents user the sign translator. Its main functions include 3 parts: detection, recognition and translation. Automatic detection and recognition of text in natural scenes is a prerequisite for automatic sign translator. In order to make the system robust for text detection in various natural scenes, the detection approach efficiently embeds multi-resolution, adaptive search in a hierarchical framework with different emphases at each layer. We also introduce an intensity-based OCR method to recognize character in various fonts and lighting condition, where we employ Gabor transform to obtain local features, and LDA for selection and classification of features. The recognition rate is 92.4% for the testing set got from the natural sign. Sign is different from the normal used sentence. It is brief, with a lot of abbreviations and place nouns. We here only briefly introduce a rule-based place name translation. We have integrated all these functions in a PDA, which can capture sign image, auto segment and recognize the Chinese sign, and translate it into English.

1. Introduction

A sign is an object that suggests the presence of a fact. 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 used for a designation, direction, safety advisory, or command. Signs are everywhere in our lives, and they make our lives easier when we are familiar with them, but they pose problems or even danger when we are not. With the progress in computer technology, the small devices such as PDA become more and more popular. It is possible and convenient for a foreign tourist to use such a small device to overcome the sign barrier.

The first issue for sign detection and recognition is the sign image input. At present, the PDA market is shared by Palm OS based PDA and WIN CE based PDA. Both of them can attach a small camera. Palm M100/M105 is the Palm OS based system, and it uses the KODAK PALMPIX camera as image input, while some new model such as SONY CLIE NR 70 even has an embedded camera. For the WIN CE based system, some models such as HP jornada 5XX serials support a CF card camera, and CASIO also provides JK-710DC digital camera card, which can be used for almost all CASIO's PDA. The image input device enhances the PDA and makes it possible for multimodel applications.

The work is related to existing researches in text detection from general backgrounds [7, 17, 23], video OCR [19], and recognition of text on special objects such as licenses plates and containers. Mullot et. al reported early attempts on container and car license plate recognition in 1991[16]. Some other researchers published their efforts on container's text detection and recognition later [1, 2, 8, 10].

Automatic detection of text in natural scenes is a very difficult task. The primary challenge lies in variations of text: it can vary in font, size, orientation, and position of text, and it may be blurred from motion, and may be occluded by other objects. Originating in 3-D space, text as signs in scene images can be distorted by slant, tilt, and shape of objects on which they are found [17]. For the Chinese text particularity, in addition to the basically horizontal left-to-right orientation, the text orientations include vertical, circularly wrapped around another object, slanted, sometimes tapering (such as in a distinct angle away from the camera), and even mixed orientations within the same text area, such as text on a T-shirt or wrinkled sign.

"Video OCR", which is to recognize text from a video stream, was motivated by digital library application and visual information retrieval tasks. The text, especially the subtitle in video, provides meaningful information, so video OCR became one of the important parts in video labeling and retrieving. In such a text detection and recognition task, the image sequence provides some useful information used to detect text and enhance the image's resolution [11, 12, 19, 22].

Compared with video OCR tasks, the recognition of text in natural scenes faces more challenges. The user's movement can cause unstable input images. Non-

professional capturing equipment in PDA can make the input image poorer than that in other video OCR tasks such as detecting captions in broadcast news programs. In addition, it has to be implemented in real time with limited resources, such as a palm-size PDA (Personal Digital Assistant).

We will have to face more troubles when we try to implement an image-based application on a PDA platform:

1. Limited computational resource. Almost all CPUs for embedded system are integer CPU, which has no floating computation component. This brings some advantage, such as low power consumption, small chip volume, easy for thermal design, etc. However, the compiler will use a float emulation library to implement floating computation, which deduces greatly the efficiency. Up to now, the most powerful CPU for the Palm OS based system is the 66MHz DragonBall CPU, while the most powerful CPU for WIN CE based system is a 206MHz StrongARM CPU. Both of them are integer CPU;

2. Limited memory. For the reason of the power consumption, volume and price, all PDA systems are configured with limited memory. A typical Palm OS based system has normally 8-16MB memory, and a WIN CE based system has 16-64MB memory. The memory is used for both storage and program, and all the running programs are in the program space. This is very critical for image-based application compared with the desktop system, which can have GB memory (including the virtual memory);

3. Small display. All PDA use a small display. Most of the Palm OS based system use 160 x 160 display, and all of the Palm PC and Pocket PC use 320x240 display, which is the quarter of standard VGA. It is too small and too hard to design an ideal user interface compared with 800 x 600 SVGA or 1024 x 768 XGA display.

All of these limit the speed, scope and capability of the PDA based application.

The paper is arranged as following: The sign detection and segmentation algorithms are introduced in section II. The character recognition is given in section III. A rule-based place name translation method is briefly introduced in section IV. Some issues on implementation in PDA are given in section IV. At last, we reach the conclusion.

2. Character detection and layout analysis

Two different methods area analysis and edge analysis have been successfully used for detecting text in an image. Area analysis is based on analyzing certain features in an area, e.g., texture and color analysis [7,

23]. DCT(Discrete Cosine Transformation) and wavelet transform are widely used for area analysis [11, 13]. A major advantage of the DCT area analysis method is that DCT coefficients can be obtained directly from the JPEG or MPEG image, while the wavelet transform can provide more stable features compared with DCT method. A disadvantage of area-based methods is that they are sensitive to lighting and character-scale changes. They often fail to detect text if the text size is too large or too small. The edges analysis can provide more stable features and is more suitable for detecting text in natural scenes. However, we have to pay special attention to filtering noises, because noises can add extra edges. A statistic based classifier or neural network is usually applied for the decision in area-based method, and a syntactic based classifier is usually for the edge-based method.

A natural sign image may be with affine transformation, high light, shadow and even mirror imaging. In order to address challenges of detecting text in natural scenes, we use a hierarchical framework with different emphases at each layer. We try to use the edge to detect the possible text regions. Once having the initial candidates, we use the local information, such as color and shape, provided by the first layer to adaptively search the neighborhood of the candidates to refine the candidates. The advantage of this strategy is to avoid to search over the entire color space in the whole image. At the last stage of detection, we combine the layout analysis and character refine together to achieve a high detection rate. The layout of a sign usually depends on language and surrounding shape of the arranged character. The schema of text detection is as figure 1.
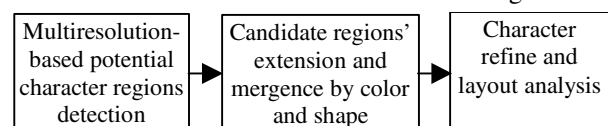

Figure 1. Multi-resolution text detection schema

The lighting condition is diverse under natural scenes. Normally, the gradient of intensity (edge) is more stable than intensity itself in lighting changes, so the edge based method is applied for text detection in our system, and we have the following hypotheses:

1. The color and intensity of the text are designed with highly contrast to its background.
2. Each character is formed by one or several connected regions.
3. The characters in the same context have almost the same size, foreground and background patterns.

The detection algorithm for potential characters is in following:

1. To get initial edge set using LOG (Laplacian of Gaussian).

2. To get the surrounding rectangle of each edge patch, and to filter all the patches by the basic geometric constraint, and those whose surrounding rectangle is too large, too sparse or too simple will be removed.
3. To merge the patches whose surrounding rectangles or extended surrounding rectangles are overlapped each other.

Although the text in the sign is always with easy distinguished colors from its background, the lighting may change it blurred when imaging. Here we model the character and its background as a GMM (Gaussian Mixture Model) in equation (1).

$$f(c) = (1-\alpha)G_{Back}(\mu_b, \Sigma_b) + \alpha G_{Front}(\mu_f, \Sigma_f) \ ,(1)$$

where $G_{Back}(\cdot)$ is the color distribution of the background, $G_{Front}(\cdot)$ is the color distribution of the foreground, α provides the cue on the complexity of the character, $\|\mu_f - \mu_b\|$ identifies the contrast for a color space invariant to the lighting condition, and Σ_b, Σ_f hint the font style. We can also describe the character's GMM with $(\alpha, \mu_b, \Sigma_b, \mu_f, \Sigma_f)$. In practice, we selected the RGB and HSI color space, and took R, G, B, H and I as the components in the color model.

The last step of text detection is layout analysis. The objective of layout analysis is to align characters in an optimal way, so that the characters that belong to the same context will be aligned together, shortly, to put all the characters in the same sentence or phrase together. Usually, the text layout has some intrinsic and extrinsic cluster features. The intrinsic features are those whose features will not change with the camera position, and the extrinsic features will change with the camera position. The intrinsic feature includes font style, color and contrast, and the extrinsic features includes character size, sign shape, etc. Both the intrinsic and extrinsic features can provide some cue to analyze the layout. The procedure of layout analysis is in following:

1. To get some the intrinsic and extrinsic character features, such as size and color distribution.
2. To fit the layout by Hough transform. If the candidate characters with nearly same features can form a certain shape, such as horizon line, vertical line or semicircle, we can believe that these characters are in the same context.

All the detected candidate characters may be in the same one context or may be in different contexts. Fig.2 is an example of sign character detection, and it gives a rectangle around the characters in the same context. It will be further segmented into individual character images before being sent to recognition model finally. The segmentation is done by the horizontal and vertical histograms, and the detail method is introduced in [4]. Fig.3 is the segmented individual character images, and their size is not the completely same even in the same context.

Figure 2. Sign character detection

Figure 3. Segmented character images

3. Recognition of character in natural scenes

OCR is one of the most successful fields in pattern recognition. For clearly segmented printed materials, state-of-the-art techniques offer virtually error-free OCR for several important alphabetic systems including Latin, Greek, Cyrillic, and Hebrew alphabets and their variants. However, when the number of character set in the language becomes large, such as the Chinese or Korean writing systems, or when the characters can not be separated from one another, such as Arabic or Deva-nagari print, the error rates of OCR systems are still far from that of human readers, and the gap between the two is exacerbated when the quality of the character image is compromised, e.g., the image is captured using a video or camera. Especially, the text image can be easily deformed because of an inappropriate camera view angle in a sign recognition and translation task.

It is a key issue how to get robust features of the character in natural scenes in order to adapt to small-shaped varieties of the character. Basically, there are intensity-based and binarilization-based method to get image features. The advantage of the intensity-based method is that it can avoid the information lost compared with the binary image based method. Pa-vlidis may be the first one who introduced the idea of getting features directly from intensity image for OCR [18]. The disadvantage is that the description of the gray image is complex while that of the binary image can easily be done by both geometric and algebraic descriptor.

Gabor wavelet is used to get the character features in our system. Gabor wavelet is a sinusoidal plane wave with particular frequency and orientation, modulated by a Gaussian envelope. It is suitable for extracting orientation-dependent frequency contents of patterns. For Gabor wavelet's good mathematic properties, it has been widely applied for data compression [20], face recognition [21], texture analysis [15] and other aspects

of image processing in recent years. In OCR field, Gabor was first used for handwriting Chinese character recognition by Deng et al [3], but most of these works are based on the binary character image [5, 6]. Yoshimura and his colleagues ever reported their effort on intensity character recognition for the video stream [24], and they used Gabor for feature description and LVQ for feature selection.

The characters embedded in natural scenes are usually under un-uniformed lighting condition, some with obvious contrast, some not, and some with high light while others may in dark, which will produce additional varieties to the features for the same character. So we utilize the local stroke intensity normalization before Gabor transform to deducing these negative factors. The local intensity normalization intends to keep the character with almost the same intensity distribution. The size normalization is also been done for each character before intensity normalization. Fig.4 is the normalized character images.

Figure 4. Normalized character images

For a given pixel (x_1, y_1) with gray level $I(x_1, y_1)$ in an image, its Gabor feature can be regarded as the convolution

$$J_j(x_1, y_1, k, \theta) = \int I(x_1 - x, y_1 - y)G(x, y, k, \theta)\mathrm{d}x\mathrm{d}y$$
. $\quad$ (2)

Where,

$$G(x, y, k, \theta) = G_1(x, y)\exp(iR) - G_1(x, y)\exp(-\frac{\sigma^2}{2}) \quad \text{is}$$

the complex-valued 2D Gabor function modulated by a Gaussian envelope, and where

$$G_1(x, y) = \frac{k^2}{\sigma^2}\exp\left[-\frac{k^2(x^2 + y^2)}{2\sigma^2}\right], \quad R = kx\cos\theta + ky\sin\theta,$$

and $k = \frac{2\pi}{\tau}$. The parameter σ is the deviation of the Gaussian envelope, τ and θ are the wavelength and orientation of Gabor function respectively. Frequency responses of Gabor filters in four orientations (0, 45, 90 and 135) are illustrated in figure 5.
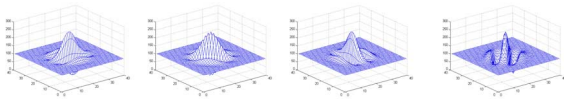

Figure 5. Four directions' Gabor filter

Suppose that m frequencies and n orientations are used for extracting Gabor feature, we can have a vector of $m \cdot n$ complex coefficients for each position, which is used to represent the position's local features. In our application, we divide a character into 7x7 grids as shown in figure 6, which results in $49m \cdot n$ dimensions' feature vector for a character.
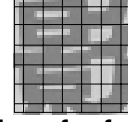

Figure 6. Regions for feature extraction

We use LDA to deduce the feature redundancy in the feature vector, which can also be used for feature optimization. It is a transform that can maximize the between-class scatter matrix S_b and minimize the within-class scatter matrix S_w simultaneously, as equation (3).

$$\arg_W\left\{\max \frac{\mathbf{W}^T \mathbf{S}_b \mathbf{W}}{\mathbf{W}^T \mathbf{S}_w \mathbf{W}}\right\} \quad (3)$$

Feature vector $\mathbf{x}$ in the original feature space is projected to this new space and have the new feature vector $\mathbf{y}$ as equation (4), which is used for classification,

$$\mathbf{y} = \mathbf{W}^T \mathbf{x} \quad (4)$$
.

We use KNN classifier for character recognition finally.

The system can recognize all level 1 characters in Chinese national standard character set GB2312-80 (total 3755 different characters). We have collected more 2000 Chinese sign images, which contains more than 8000 characters. These signs are captured by a still digital camera from natural scenes of indoors and outdoors in China. We randomly select 1630 character images from this sign library, and the different characters in the testing set roughly cover 1/5 of the level 1 Chinese characters, with variations in fonts, lighting conditions, rotation, and even affine transform. Figure 7 is some examples from the testing set. The accuracy is 92.46% without any postprocessing. The most important is that the testing set is completely independent from the training set. So it is quite a satisfied result.

Figure 7. Some examples from the testing set

To test the robust of recognition, we add Guassian noise on the character images. For a given pixel whose intensity is $I(x, y)$ in the character image, we have

$$I'(x, y) = I(x, y) + n(x, y), \quad (5)$$

where $n(x, y) \sim N\left(0, (I(x, y) \cdot \gamma)^2\right)$.

Parameter γ represents the intensity of the noise. Figure 8 illustrates the impact on a Chinese character from testing set when we added noise respectively $\gamma = 0.05$, $\gamma = 0.10$, $\gamma = 0.15$ and $\gamma = 0.20$.

The top curve of Figure 9 shows the recognition rate of 1630 characters in testing set when we added different intensity of Gaussian noise. It is only about 1% decrease of recognition rate when 10% noise is added, and about 6.5% decrease when 20% noise is added, compared with that of the original testing set. This Figure also illustrates the effective of local intensity normalization, which shows that the recognition accuracy is about 24% to 28% higher than that without normalization.
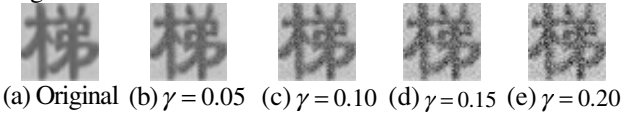

(a) Original (b) $\gamma = 0.05$ (c) $\gamma = 0.10$ (d) $\gamma = 0.15$ (e) $\gamma = 0.20$

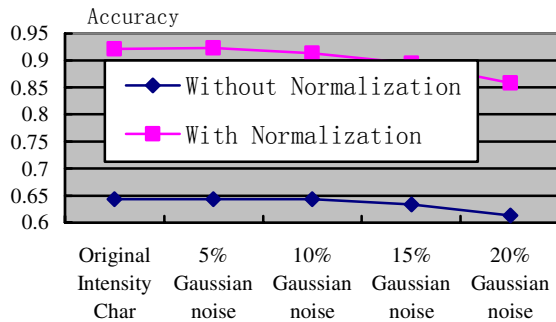**Figure 8. An example with different noises**

Figure 9. The noise's impact on character recognition

4. Sign Translation

The system finally presents users the sign translation. Sign translation is also a kind of machine translation, but it has something special. The rule-based machine translation is used in our system, and the relative methods can be found in [9, 14]. Here, we just briefly discuss about the place name identification and translation. The place name as a special noun is the frequently used part in sign. In Chinese, the place name usually ends with some special characters, such as "街" (street), "区" (district), "巷" (lane), etc. Following the word segmentation module, we add a rule-based parse module for this. We define these place nouns as:

Place noun prefix + Place noun suffix,

where the suffix is from the fixed suffix set, which includes the characters "街", "区", "路", "村", "巷", etc. Some rules about the prefix is in following:

Prefix = noun | noun + position noun | position noun + noun | numeral | …

When we meet a place noun suffix, the place noun parse module will use the backtracking method to get one or more prefix candidates. The succedent parse module will select one from them if more than one candidates is gotten.

**Figure 10.	Translation examples**

Figure 10 is 2 examples of the sign translation. The second is the place name translation, and the prefix part "光辉"is translated into Chinese PinYin "Guang hui".

5. System integration

We took the HP Jornada 568 as the platform for our sign recognition and translation system. It is configured with a 206MHz StrongARM CPU and 64MB RAM which is shared by storage and program space. It provides a CF type I slot which is attached by a camera as the imaging device. It has a 320x240 LCD with 65536 colors.

Figure 11. Sign detection, recognition and translation in PDA

We have done some special optimizations on both speed and memory for the resource limit of the small device. There are quite much float computation in the system, especially during the Gabor and LDA procedure, and it will deduce greatly the running speed in the small device. We ever had the test that it will take more than 20 seconds to recognize a Chinese word from Chinese national standard character set GB2312-80 (total 3755 different characters) when the float computation is used. We substituted the floating computation with normalized integer computation to speed up it, which will surely bring the loss of the precision. The basic ideal is to normalize the operands to integer under the given operator and to ensure no overflow will occur for both operand and result. It took only 1 second to recognize a

Chinese word after speeding. We know that the file system in PDA is also stored in RAM, it does not have tracking and seeking time, which is about ms for hard drive in the desktop system. So database can be access directly from file system without the additional memory allocation, which saves greatly the memory occupation.

Figure 11 gives the system demo running in the HP Pocket PC Jornada 568. It works in two modes at present, automatic mode and manual mode. The automatic mode can automatic detect and segment sign text from the image, automatic recognize Chinese characters and translate them into English. The manual mode is to recognize and translate sign by the user giving a sign's rectangle region.

6. Conclusion

In this paper we presents a PDA-base sign detection, recognition and translation system, which propose a kind of tourist tool to try to let tourist to understand a sign in a foreign country that may specify warnings or hazards. We address the special aspects of sign text's detection, OCR and machine translation. All these are implement on a convenient small device Pocket PC. This system can recognize and translate Chinese sign now, whose character set is formed by level 1 Chinese national standard character set. The further work includes: 1. Using a language model to connect recognition and translation and making the result more understandable; 2. Working on the multiple languages and make it become a reconfigurable system with different source and target language.

Reference

[1]   E. Barnes, Image recognition for shipping container tracking and I.D, Advanced Imaging, Vol. 10, No.1 pp. 61-62, 1995.

[2]   Y. Cui, and Q. Huang, Character Extraction of License Plates from Video, Proc. of CVPR, pp. 502-507, 1997.

[3]   D. Deng, K. P. Chan, and Y. Yu, Handwritten Chinese character recognition using spatial Gabor filters and self-organizing feature maps, Proc. of ICIP, Vol. 3, pp. 940-944, 1994.

[4]   J. Gao, J. Yang, Y. Zhang, and A. Waibel, Text Detection and Translation from Natural Scenes, Technical Report CMU-CS-01-139, Computer Science Department, Carnegie Mellon University, June, 2001.

[5]   Y. Hamamoto, S. Uchimura, K. Masamizu, and S. Tomita, Recognition of handprinted Chinese characters using Gabor features, Proc. of ICDAR, Vol. 2, pp. 819-823, 1995.

[6]   Q. Hou, Y. Ge, and Z. Feng, High performance Chinese OCR based on Gabor features, discriminative feature extraction and model training, Proc. of ICASSP, Vol. 3, pp. 1517-1520, 2001.

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

[8]   S. Kumano, K. Miyamoto, M. Tamagawa, H. Ikeda, and K. Kan, Development of container identification mark recognition system, Transactions of the Institute of Electronics, Information and Communication Engineers D-II, Vol. J84D-II, No.6 pp. 1073-1083, 2001

[9]   H. Lee, J. Dai, and Y, Chang, "Parsing Chinese Nominalization based on HPSG", Computer Processing of Chinese & Oriental Languages, Vol. 6, No. 2, Dec. 1992.

[10]   C. M. Lee, and A. Kankanhalli, Automatic extraction of characters in complex scene images, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 9, No. 1, pp. 67-82, 1995

[11]   H. Li, D. Doermann, and K. Omid, Automatic Text Detection and Tracking in Digital Video, IEEE Trans. on Image Processing, Vol. 9, No. 1, pp. 147-156, 2000

[12]   R. Lienhart, Automatic Text Recognition for Video Indexing, Proc. of ACM Multimedia, pp. 11-20, 1996.

[13]   Y. Lim, S. Choi, S., S. Lee, Text extraction in MPEG compressed video for content-based indexing, Proc. of ICPR, Vol. 4, pp. 409-412, 2000.

[14]   Q. Liu, and S. Yu, "TransEasy: a Chinese-English machine translation system based on hybrid approach", Procedings of the Third Conference of the Association for Machine Translation in the Americas, 1998.

[15]   R. Mehrotra, K. R. Namuduri, N. Ranganathan, "Gabor filter-based edge detection", Pattern Recognition, Vol. 25, No. 12, pp. 1479-1494, 1992

[16]   R. Mullot, C. Olivier, J. L. Bourdon, P. Courtellemont, J. Labiche, and Y. Lecourtier, Automatic extraction methods of container identity number and registration plates of cars, Proc. of Int. Conf. on Industrial Electrlnics, Control and Instrumentation, Vol. 2591, pp. 1739-44, 1991.

[17]   J. Ohya, A. Shio, and A. Akamatsu, Recognition of characters in scene images. IEEE Trans. on PAMI, Vol. 16, No. 2, pp. 214-220, 1994.

[18]   T. Pavlidis, Recognition of Printed Text Under Realistic Conditions, Pattern Recognition Letters, Vol. 14, No. 4, pp. 317-326, 1993.

[19]   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.

[20]   H. Szu, B. Telfer, and J. Garcia, "Wavelet transforms and neural networks for compression and recognition", Neural Networks, Vol.9, No. 4, pp. 695-708, 1996

[21]   L. Wiskott, J. M. Fellous, N. Kruger, and C. Malsburg, "Face Recognition by Elastic Bunch Graph Match", IEEE Trans. on PAMI, Vol. 19, No. 7, pp. 764-768, 1997

[22]   E. K. Wong, and M. Chen, A Robust Algorithm for Text Extraction in Color Video, Proc. of ICME, 2000.

[23]   V. Wu, R. Manmatha, and E. M. Riseman, TextFinder: An Automatic System to Detect, IEEE Trans. on PAMI, Vol. 21, No. 11, pp. 1224-1229, 1999.

[24]   H. Yoshimura, M. Etoh, K. Kondo, and N. Yokoya, Grayscale Character Recognition by Gabor Jets Projection, Proc. of ICPR, Vol. 2, pp. 335-338, 2000.