READING LABELS OF CYLINDER OBJECTS FOR BLIND PERSONS

Ze Ye[1], Chucai Yi[2] and Yingli Tian[1,2]
[1]Dept. of Electrical Engineering, The City College of New York
[2]Dept. of Computer Science, The Graduate Center
The City University of New York, USA
e-mail: {zye01@ccny.cuny.edu, cyi@gc.cuny.edu, ytian@ccny.cuny.edu}

ABSTRACT

We propose a camera-based assistive framework to help blind persons to read text labels from cylinder objects in their daily life. First, the object is detected from the background or other surrounding objects in the camera view by shaking the object. Then we propose a mosaic model to unwarp the text label on the cylinder object surface and reconstruct the whole label for recognizing text information. This model can handle cylinder objects in any orientations and scales. The text information is then extracted from the unwarped and flatted labels. The recognized text codes are then output to blind users in speech. Experimental results demonstrate the efficiency and effectiveness of the proposed framework from different cylinder objects with complex backgrounds.

Index Terms—blind person, projection, assistive text reading, text region, stroke orientation, distribution of edge pixels,

1. INTRODUCTION

Depending on the statistics in 2002 [16], more than 161 million persons suffered visual impairment and 37 million of them are blind persons. It is a challenging task for blind persons to distinguish different objects. They need to independently find out specific container from many containers with same shape in their daily life such as shopping, cooking, taking medications, etc. They are used to recognizing an object from its shape and material by touch and smell, but not able to acquire the text information from print labels on the object.. Some reading-assistive systems, such as voice pen, could be utilized in this situation. They integrate optical character recognition (OCR) software to offer the function of scanning and recognition of text for helping blind persons read print documents and books. Normally, it employed the gradient and curvature to capture the text feature [5]. However, these systems are generally designed for scanned document images with simple background and well-organized characters rather than packing box with multiple decorative patterns. In this case, no words and features are distorted. The OCR software cannot directly handle the scene images with complex backgrounds and curved surfaces. Also the system can only extract limited information of the object from partial of the label because it cannot reconstruct the whole surface of container. Thus these assistive text reading systems usually require manual localization of text regions in a fixed and planar object surface, such as a screen and book.

To assist blind persons more conveniently in reading labels from cylinder objects in their hands, we propose a camera-based assistive framework by new methods of an image stitching and text reading to extract significant text information from cylinder objects with complex backgrounds and multiple text patterns. The tasks of our system are image stitching to mosaic all images taken by camera and text extraction to read the involved text information from complex backgrounds. For our application, objects in camera captured images are very probably surrounded by various background outliers, and text characters usually appear in multiple scales, distortions, fonts, colors, and orientations.

The prototype system consists of a wearable camera attached to a cap or a pair of sunglasses, an audio output device such as a Bluetooth or earphones, and a mini-microphone as voice input device. This simple hardware structure ensures portability of the system. A wearable computer/PDA associated with Blind Navigation System [15] can process the input information synchronously and then transforms the input image to output voice for blind users.

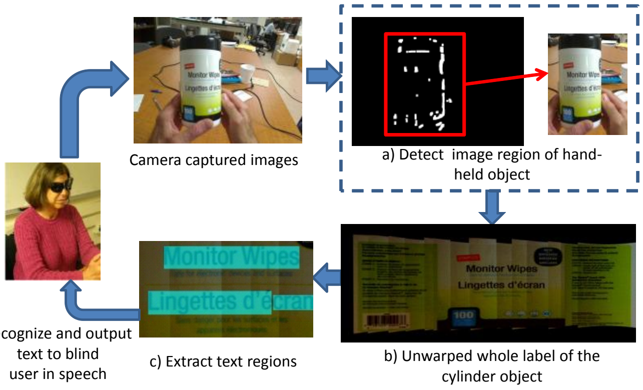
Camera captured images
a) Detect image region of hand-held object
b) Unwarped whole label of the cylinder object
c) Extract text regions
Recognize and output text to blind user in speech
Monitor Wipes
Lingettes d'écran

Fig. 1. Flowchart of the proposed framework to read labels from hand-held objects for blind users.

Fig. 1 describes the flowchart of the proposed framework. This paper focuses on the following main steps: a) detecting the object held by a blind user. In this step, a blind user wearing a camera captures the hand-held object from the cluttered background or other neutral objects in the camera view by slightly shaking the object for 1 or 2 seconds; b) reconstructing the whole label of the cylinder object by our proposed mosaic method which can handle different orientations and scales; c) extracting text to transform image-based information into text codes. In the text localization method, the basic processing cells are rectangle image patches with fixed ratio, where features of text can be obtained from both stroke orientations and edge distributions. There are many ways to detect the text region in natural scenes [1, 8, 9, 10, 11]. An Adaboost machine learning algorithm [6] is employed here as a text region classifier [23] to classify the text and non-text patches. The extracted text regions are then recognized by OCR software and communicate with the blind user in speech.

2. DETECTING IMAGE REGION OF HAND-HELD OBJECT

A camera with a reasonably wide angle is employed in our prototype system, to ensure the hand-held object appears in the camera view since the blind user cannot aim accurately. However, some non-text texture or background outliers will also appear in the camera view, e.g., when a user is shopping at a supermarket. To extract the hand-held object from other objects in the camera view, we employ a motion-based background subtraction (BGS) method to localize the object of interest in camera view.

BGS is an effective approach to detect moving objects for video surveillance systems with stationary cameras. To detect moving objects in a dynamic scene, many adaptive background subtraction techniques have been developed [18,21]. Stauffer and Grimson [17] modeled each pixel as a mixture of Gaussian approximate the model. Their system can deal with slow lighting changes and introducing or removing objects from the scene. Tian *et al*. [20] further improved the multiple Gaussian mixtures based BGS method to handle complex foregrounds and quick lighting changes.

As shown in Fig. 1(a), while capturing images of the hand-held object, the blind user first holds the object still, and then lightly shakes the object for one or two seconds. Here, we apply the efficient multiple Gaussian Mixtures based BGS method to detect the object region while blind user shakes it. More details of the algorithm can be found in [20]. Once the object of interest is extracted from the camera view, we will reconstruct the whole label of the hand-held cylinder object which requires the user rotating the object along the cylinder axis so that different parts of the label can be captured by the camera.

3. UNWARPING CYLINDER OBJECT SUFACE AND RECONSTRCING THE WHOLE LABEL

It is a challenging task to develop a camera-based text reader for blind users to read a label of a cylinder object due to the following issues: 1) Blind persons cannot see how the object appears in the camera view. The object usually appears in arbitrary direction, scales, and rotations. 2) The existing OCR software was designed for recognizing text information from scanned documents and cannot handle non-flat surfaces. 3) Only small portion of the label which faces to the camera is visible. It is very possible to miss the representative information of the object. Therefore, we first develop a robust method to unwarp cylinder surface. Our method is able to handle objects in different scales and orientations as shown in Fig. 2. Then we develop a mosaic method to obtain the print label of the object by requesting the user simply rotate the object along the long axis of the object.

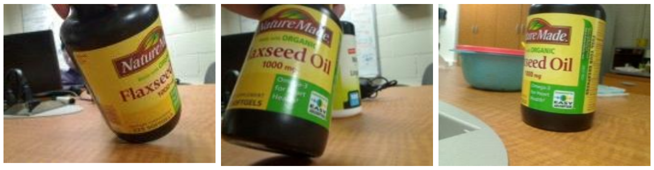

Fig. 2. Camera-captured images of a hand-held cylinder object with different scales and orientations.

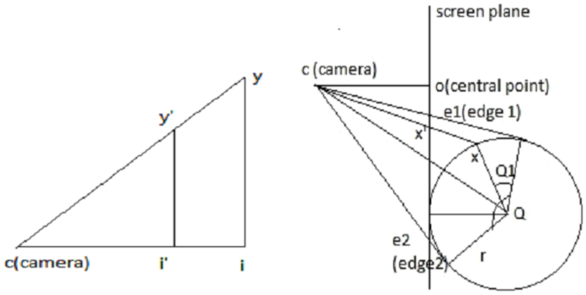

Fig. 3. Geometrical relation between the cylinder surface and its tangent plane.

3.1 Unwarping Cylinder Object Surface

The hand-held objects generally appear in different orientations in camera captured images (see Fig. 2). To unwarp camera-based cylinder object surface, we first normalize the orientation of cylinder objects into vertical based on camera perspective projection. Then each pixel in the cylinder surface will be projected to its tangent plane. The geometric relationship of the cylinder surface and its tangent plan is illustrated in Fig. 3, where point c is the original coordinate of the camera, line $i'y'$ and e_1e_2 represent the tangent plane of the cylinder object after normalized to vertical direction, line iy corresponds to the vertical line on the cylinder object at position x, r is the radius of the cylinder object, f is the focus length, and Q is the frontal part of the cylinder object which is visible in camera view. The geometry relationship between the image of the cylinder object and its flattened image on the tangent plane can be calculated by following equations:

$$x = Q1 * r; \tag{1}$$

$$cx = -2 * r * \left[\frac{r+f}{cos(\angle ocQ)}\right] * cos\left(\frac{Q}{2} - Q1\right) + \left[\frac{r+f}{cos(\angle ocQ)}\right]^2 + r^2; \tag{2}$$

$$\angle ocx = \angle oceQ - arccos\left(\frac{cx^2 + \left[\frac{r+f}{cos\left(\angle ocQ\right)}\right]^2 - r^2}{cx^2 + \left[\frac{r+f}{cos\left(\angle ocQ\right)}\right]^2 - r^2}\right) \tag{3}$$

$$Ox^{'} = f * tan\,(\angle ocx); \tag{4}$$

$$\frac{yi}{y'i'} = \frac{ci}{ci'} = \frac{cx}{cx'} 2 * cx * \frac{r+f}{cos\left(\angle ocQ\right)}; \tag{5}$$

Some unwarped images are displayed in Fig. 4. After obtaining the unwarped image of the cylinder surface, we will then stitch the different parts of the images to obtain the flattened label.

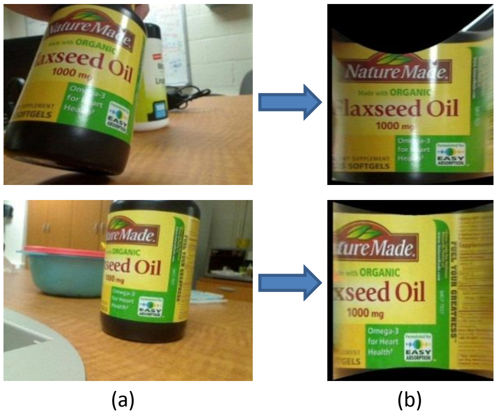

Fig. 4. Example results of unwarp cylinder objects in different orientations. a) Original images of the object; b) normalized images in vertical orientation with flattened surface.

3.2 Reconstructing the Whole Label

In this section, we will describe our mosaic method to reconstruct the whole label of the cylinder object. When camera is capturing object images, multiple images of different parts of the label can be captured when a blind user rotates the object along its vertical axis. In most cases, an object will be captured from multiple perspective, obtaining camera-based object images of different parts. We first apply a preprocessing including histogram equalization, threshold selection, [13] and downsample the image to half size to reduce effects of lighting changes and noises. As shown in Fig. 5, we employ Scale-invariant feature transform (SIFT) features [7] to calculate the matching points between two overlapped images of different portion of the object label. We then use RANSAC algorithm to estimate the transformation matrix in order to stitching all the flattened images together. Fig. 6 displays some stitched labels of the cylinder objects. In these images of whole labels, more meaningful text information can be extracted comparing to images only contain partial of the labels.

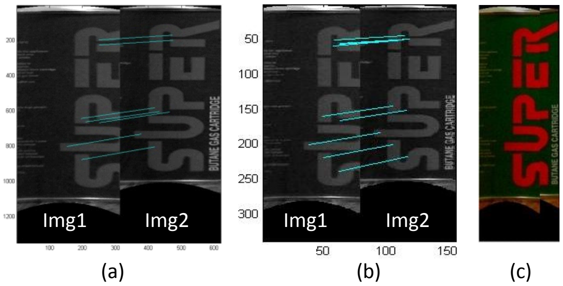

Fig. 5. Matching two consecutive images of object label based on SIFT features at different scales. (a) Matched SIFT features at the original images. (b) Matched SIFT features at the downsampled images. (c) Stitched image of "Img1" and "Img2" by the proposed method.

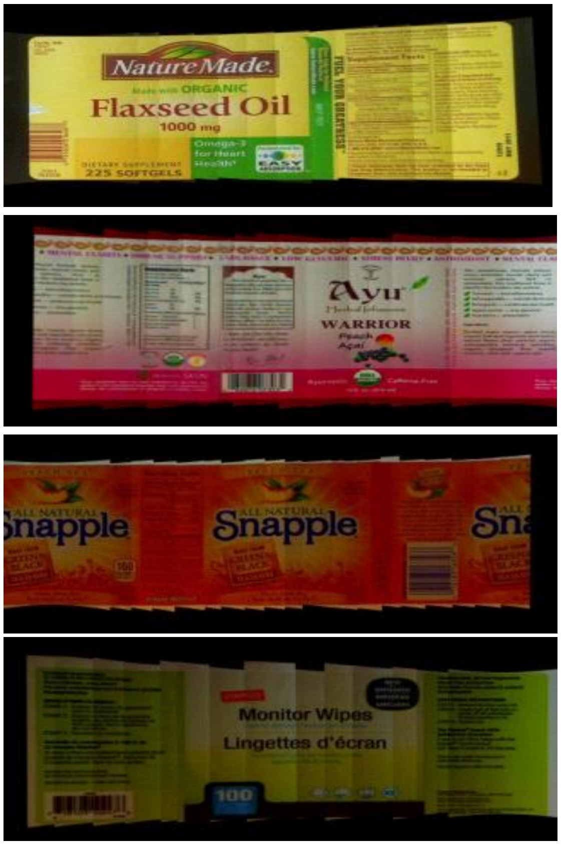

Fig. 6. Reconstructed whole labels of cylinder objects.

4. AUTOMATIC TEXT LOCALIZATION

To extract text information from the reconstructed labels, an effective algorithm of automatic text localization is designed on the basis of previous work [2, 3]. It combines pixel-level text layout analysis and structure-level text feature learning.

4.1 Layout Analysis

According to our observations, the print text on the object labels mostly appears in approximately uniform color and horizontal alignment. Image patches compatible with the two layout characteristics are generated from scene image as candidate regions, because they are very likely to contain text information.

To extract text characters from background outliers in different color, we adopt color reduction method [12] to group pixels in similar colors together, obtaining a color layer. Each scene image is decomposed into several color levels, as shown in Fig. 7. Cluttered background with all kinds of non-text outliers is transformed into relatively simple background noise. Canny edge detection is applied to generate a group of boundaries at each color level. Each boundary is C_1 circumscribed by a rectangle bounding box to measure its size and location. Several empirical geometrical constraints, are defined to filter out the boundaries of background noises, as Eq. (6). In our experiments, the height of a character boundary does not exceed 10, and the width-to-height ratio is between 0.3 and 1.5.

$$\begin{gathered} h(C) > 10 \\ 0.3 \leq w(C)/\, h\,(C) \leq 1.5 \end{gathered} \tag{6}$$

where $h(\cdot)$ and $w(\cdot)$ denote the height and width of a boundary.

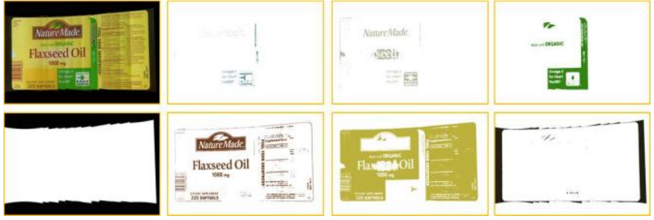

Fig. 7. A constructed label image and its 7 color levels obtained by color reduction. In the second image at the bottom row, the title is separated from background.

Next, we search for text regions at each color level by horizontally aligned boundaries. Since label text always appears in the form of text strings rather than a single character, a character boundary should be accompanied by several siblings. For each boundary in the edge map of a color level, we check its sibling boundaries with similar height, vertical location, and reasonable distance. In our experiments, we define two boundaries C_1 and C_2 as siblings if they satisfy the following conditions.

$$\begin{gathered} 0.8 \leq h(C_1)/h(C_2) \leq 1.2 \\ |y(C_1) - y(C_2)| \leq 0.5 * \max\big(h(C_1), h(C_2)\big) \\ |x(C_1) - x(C_2)| \leq 2 * \max\big(w(C_1), w(C_2)\big) \end{gathered} \tag{7}$$

where $x(\cdot)$ and $y(\cdot)$ denote the centroid coordinate of a boundary, and $h(\cdot)$ and $w(\cdot)$ denote the height and width of a boundary. For each boundary with at least two siblings, on its left and right respectively, we initialize an align group. On each color level, all the align groups are found out based on the sibling conditions, and then we merge those with intersections into larger align group.

For the boundaries within an align group on a color level, both layout characteristics are satisfied. Then we crop out image regions covering the align groups, which are used as candidate regions for further step of text prediction.

4.2 Text Feature Learning

The layout analysis distinguishes text from background outliers by using size, location, color, and alignment of character boundaries. However, it ignores the inner structure of text character. Some background object and texture, like window, brick and grid, are also extracted as candidate regions. Thus we will further filter out the false positive candidate regions. At first, we build a dataset of candidate regions, which are all obtained from layout analysis. This dataset contains about 2000 text patches as positive samples and about 4000 non-text background patches are used as negative samples, as shown in Fig. 8. In the following paragraphs of the paper, we denote a candidate region by sample. All the samples in this dataset are normalized into width 96 and height 48. We will train a text classifier to predict whether a given sample truly contains text information (positive sample) or not (negative sample).

To obtain a robust text classifier, we map each pixel in a sample into a feature-specific value associated with character appearance and structure. It will transform the intensity-based sample into a feature-based map (feature map), and generate representative and discriminative text features. In our system, three feature maps are proposed including gradient, stroke width consistency, and stroke orientation.

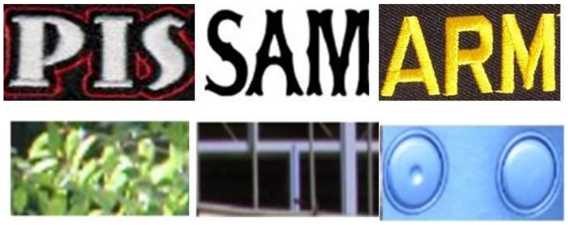

Fig. 8. Some examples of training samples for text extraction. Top row shows the positive samples that are text patches, and bottom row shows the negative samples that do not contain text information but have similar structure to text patch.

Character structure includes regular-shaped boundaries and plain torso. Thus gradient distribution plays an important role in text structure modeling. In our system, Sobel operator is adopted to calculate the gradients at each pixel p_0 in horizontal $G_x(p_0)$ and vertical directions $G_y(p_0)$, which are combined into gradient magnitude by $S(p_0) = \sqrt{G_x^2(p_0) + G_y^2(p_0)}$. We obtain a feature map of gradient magnitude.

Stroke serves as basic unit of character structure. In our work, stroke is defined as a connected image region with half-closed boundary in consistent width and orientation. We generate another two feature maps based on stroke width and stroke orientation respectively. First, stroke width transform (SWT) method [4] is applied to localize text stroke by stroke width consistency. SWT calculates stroke width by probe rays from edge pixel along gradient direction, as shown in Fig. 9. In edge pixel p_1, a probe ray is generated along gradient direction, and we extend the ray until it touches another edge pixel p_2. If the gradient at p_2 has similar magnitude and opposite direction to the gradient at p_1, a segment p_1p_2 is generated across the stroke. Segment length $|p_1p_2|$ is calculated as stroke width and any pixel p_0 at the segment is labeled by $S(p_0) = |p_1p_2|$.

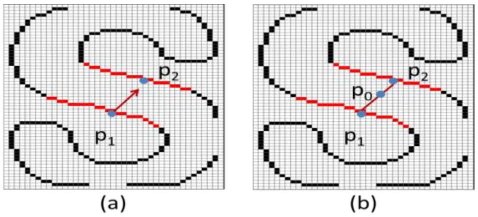

Fig. 9. Stroke width transform. (a) Probe ray starting from edge pixel $\boldsymbol{p_1}$ along gradient direction at $\boldsymbol{p_1}$, and it encounters $\boldsymbol{p_2}$ with similar gradient magnitude and opposite gradient direction. (b) Segment length is calculated as stroke width to label each pixel at the segment $\boldsymbol{p_1}\,\boldsymbol{p_2}$.

Furthermore, character structure is closely related to the distribution of stroke orientations. Thus we design a stroke orientation transform (SOT) method to calculate feature map of stroke orientation. For a pixel p_0, we set a circular range as $R(p_0) = \{p|d(p,p_0) \le 36\}$, where $\mathrm{d}(.)$ is set as Euclidean distance. Then the edge pixel $p_e \in R$ with the minimum Euclidean distance from p_0 is calculated by (8), where p is edge pixel within the range of p_0. As shown in Fig. 10, the distribution of stroke orientations is quantized into 9 bins within the degree range $(-90^o, 90^o]$. Orientation distribution of a text patch has two usual characteristics. First, due to the highest frequency of vertical and the second highest of horizontal strokes in English letter, the orientation bin at 0 has the highest value, and that at 90 has the second highest value. Second, the orientation distribution of text patch is approximately symmetrical between -45^o and 45^o, because slant strokes usually appear in pair, like ‘A’ and ‘X’.

$$p_e = \underset{p\in\mathrm{R}(\mathrm{p}_0)}{\operatorname{argmin}}\, d(p,p_0)$$
$$S(p_0) = \left(\arctan\left(G_y(p_e),\ G_x(p_e)\right)\right) \qquad (8)$$

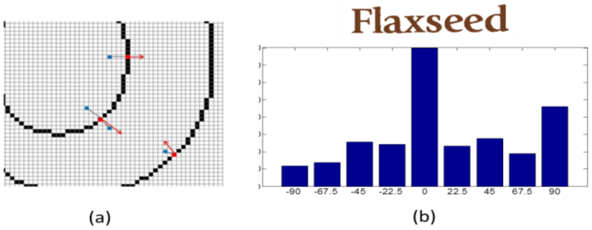

Fig. 10. Stroke orientation transform. (a) Any pixel (blue dot) is mapped into its nearest edge pixel (red dot), and labeled by gradient orientation (red arrow) at the edge pixel. (b) The distribution of stroke orientations for the image patch “Flaxseed”, where the orientations are quantized into 9 bins within the degree range $(-90^o, 90^o]$.

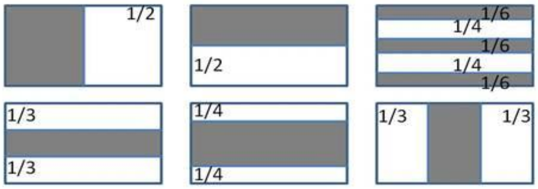

Fig. 11. Block patterns to extract feature vectors from the feature maps.

Three feature maps of a sample are transformed into a feature vector by using 6 block patterns as shown in Fig. 11. Each block pattern consists of several partitions, either white or gray. They are transformed into the same size as feature maps, and then act as a mask of feature extraction. Then measurement value is calculated, for example, the sum of feature values in gray partitions is subtracted from that in white partitions. We calculate all measurement values from the combinations of feature maps and block patterns, and cascade them into a feature vector for each sample. The feature vectors are regarded as observation points in sample space, and input into cascade-Adaboost learning model [22] to train a text classifier. This classifier can predict whether a sample (candidate region) truly contains text or not.

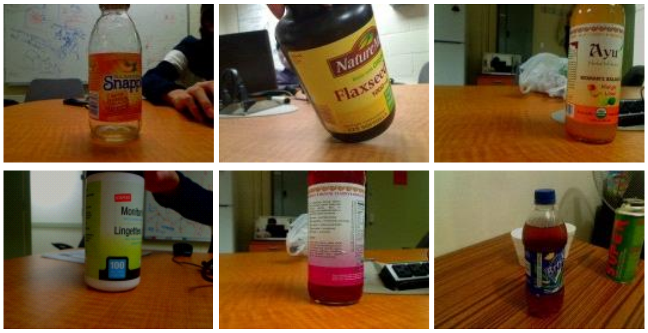

Fig. 12. Some image samples in our dataset.

5. EXPERIMENTS

We build a dataset of camera-based cylinder objects, which contains 300 images from 16 cylinder objects in total. These images are captured by a Logitech-webcam, and the image resolution is 1280*1024 pixels. Not all the objects appear in upright position. Some of the cylinders have more than 70 degrees between the camera optical axis and the cylinder object vertical axis. Since text layout analysis based on adjacent character grouping can only handle text strings with three or more character members, the system discards the images containing ground truth text regions with text has less than 3 characters. Thus all objects are selected captured as testing images to evaluate our system. Fig. 12 shows some examples of the objects in our dataset. The localization algorithm is performed on the scene images of Robust Reading Dataset to calculate image regions containing text information. Fig. 13 depicts some results of localized text regions, marked by blue rectangle boxes. Our system can efficiently obtain information from the object labels.

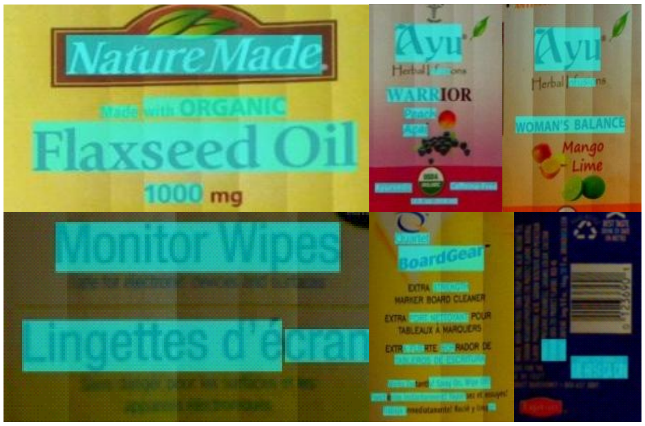

Fig. 13. Blue rectangles were drawn out after text localization. This algorithm has high accuracy for large words that have 3 or more than 3 characters. OCR will be applied to the extracted text regions to recognize the text codes.

6. CONCLUSION

In this paper, we have proposed an assistive system to read text labels from cylinder objects for blind persons by combining a perspective projection algorithm with a text extraction method. Our system can fully reconstruct the whole label of cylinder objects. To localize text in camera captured images, adjacent character grouping is performed to calculate candidates of text patches prepared for text classification. The Adaboost-based text classifier is applied to obtain the text regions. Off-the-shelf OCR is then employed to perform word recognition in the localized text regions and transform into audio output for blind users.

Our future work will focus on building a general geometry model to reconstruct the objects' surface with different shapes and extending localization algorithm to process text strings with less than 3 characters. We will also address the significant human interface issues associated with reading region selection by blind users.

7. ACKNOWLEDGEMENT

This work was supported in part by NSF grant IIS-0957016, EFRI-1137172, NIH 1R21EY020990, FHWA grant DTFH61-12-H-00002, and Microsoft Research.

8. REFERENCE

[1] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," In *CVPR*, Vol. 2, pp. II-366 – II-373, 2004.

[2] X. Chen, J. Yang, J. Zhang and A. Waibel, "Automatic detection and recognition of signs from natural scenes," In *IEEE Transactions on image processing*, Vol. 13, No. 1, pp. 87-99, 2004.

[3] Shiu, Y.C, Huang C "Locating cylindrical objects from perspective projections" Aerospace and Electronics Conference, 1991. NAECON 1991., Proceedings of the IEEE 1991 National

[4] B. Epshtein, E. Ofek and Y. Wexler, "Detecting text in natural scenes with stroke width transform," In *CVPR*, pp. 2963-2970, 2010.

[5] Open source code of optical character recognition, Tesseract OCR: https://code.google.com/p/tesseract-ocr/.

[6] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," In *Int. Conf. on Machine Learning*, pp.148–156, 1996.

[7] Lowe, David G. (1999). "Object recognition from local scale-invariant features". *Proceedings of the International Conference on Computer Vision*. **2**. pp. 1150–1157.

[8] K. I. Kim, K. Jung, and J. H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," *In IEEE Trans. on PAMI*, 2003.

[9] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi, "Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model," *IEEE Trans on Image Processing*, Vol. 16, No. 8, pp. 2117-2128, 2007.

[10] S. M. Lucas, "ICDAR 2005 text locating competition results," *Proceedings of the ICDAR*, Vol. 1, pp 80–84, 2005.

[11] L. Ma, C. Wang, B. Xiao, "Text detection in natural images based on multi-scale edge detection and classification," In *the Int. Congress on Image and Signal Processing (CISP)*, 2010.

[12] N. Nikolaou and N. Papamarkos, "Color Reduction for Complex Document Images," *International Journal of Imaging Systems and Technology*, Vol.19, pp.14-26, 2009.

[13] N. Otsu, "A threshold selection method from gray-level histograms," In *IEEE Tran.s on system, man and cybernetics*, pp. 62-66, 1979.

[14] T. Phan, P. Shivakumara and C. L. Tan, "A Laplacian Method for Video Text Detection," In *Proceedings of ICDAR*, pp.66-70, 2009.

[15] L. Ran, S. Helal, and S. Moore, "Drishti: an integrated indoor/outdoor blind navigation system and service," In *Pervasive computing and communications*, pp. 23-40, 2004.

[16] S. Resnikoff, D. Pascolini, D. Etya'ale, I. Kocur, R. Pararajasegaram, G. P. Pokharel, et al, "Global data on visual impairment in the year 2002." In *Bulletin of the World Health Organization*, 844- 851, 2004.

[17] C. Stauffer and W.E.L. Grimson, "Adaptive Background mixture Models for Real-time Tracking", CVPR99, June, 1999.

[18] H. Schneiderman and T. Kanade, "A statistical method for 3D object dection applied to faces and cars," In *CVPR* 2000.

[19] P. Shivakumara, T. Phan, and C. L. Tan, "A gradient difference based technique for video text detection," *The 10th ICDAR*, pp.66-70, 2009.

[20] Y. Tian, Max Lu, and A. Hampapur, "Robust and Efficient Foreground Analysis for Real-time Video Surveillance," IEEE CVPR, San Diego. June, 2005.

[21] P. Viola and M. J. Jones, "Robust real-time face detection," In *IJCV* 57(2), 137–154, 2004.

[22] C. Yi and Y. Tian, "Text string detection from natural scenes by structure based partition and grouping," In *IEEE Transactions on Image Processing*, 2011.

[23] C. Yi and Y. Tian. Assistive Text Reading from Complex Background for Blind Persons. In *ICDAR Workshop on Camera-based Document Analysis and Recognition (CBDAR)*, Springer LNCS-7139, pp.15-28, 2011