Wearable Reading Assistant Device with Scene Text Locator for the Blind

Takahiro Sasaki  
Graduate School of Information Sciences  
Tohoku University, Japan  
sasaki@sc.isc.tohoku.ac.jp

Akira Saito  
School of Engineering  
Tohoku University, Japan  
akirab@sc.isc.tohoku.ac.jp

Hideaki Goto  
Cyberscience Center  
Tohoku University, Japan  
hgot@isc.tohoku.ac.jp

Abstract—Finding signboards and capturing the text images in a form suitable for character recognition is difficult for the blind. This report presents a reading assistant device with a scene text locator that shows the text location using sound signals. A fast text detection algorithm using an edge feature and the string’s structural feature is proposed.

Keywords—reading assistant, scene text detection, text locator

I. INTRODUCTION

Some text information acquisition devices have been developed for helping the visually-impaired to access text information in the environment. The device in [1], based on a PDA (Personal Digital Assistant), requires some text acquisition training so the user can hold the target at a proper distance from the device. However, some objects such as signboards cannot be held by the user’s hand. Thus, the users often have difficulties in capturing the entire text in the camera view. In this report, we present a reading assistant device with a scene text locator that produces sound signals for helping the user to direct the camera to the text center. Since detecting text from scene images is an important step for those applications, many approaches have been proposed. Liu et al. proposed an edge-based method, focusing on the text region’s edge density[2]. Huang et al. used text line’s features and Harris’s corner to extract text region[3]. We use an edge feature and the string’s structural feature to improve the accuracy of text detection.

II. READING ASSISTANT DEVICE WITH SCENE TEXT LOCATOR

A. Overview of the device

Our system detects text regions in scene image, converts and plays some guidance sound according to the location of the text region. Our prototype device is equipped with Logitec C905m Web Camera as the image capture device, laptop PC with a 2.53GHz CPU, and bone conduction headphones. We chose TEAC Filltune HP-200 bone conduction headphones system since it does not cover the user’s ears and imposes little impact on the sound perception which is very important for the blind to navigate themselves. The input image is in color at 640×480 pixels.

B. Text detection method

The first step is to detect text regions from each video frame obtained by the webcam. The Sobel filter is applied to the image in order to obtain an edge intensity map. In the edge image, connected components and the bounding boxes are found. When the bounding box’s width or height is smaller than 8 pixel, or the width or height is larger than 200 pixel, the box is ignored as a noise. Inspired by [4], we assume character’s shape is more complicated than the other objects, and we use the DCT-based text detection feature to reduce the noise components. Each bounding box is normalized to 32×32 in size, then DCT is applied to the normalized image. A threshold value is calculated from the entire image. If the DCT value is lower than the threshold, the connected component is filtered out. We use the bounding boxes’ centroids for the noise filtering based on the text string’s linear structure. The first and the second nearest centroid and its distance is calculated for each box. The first and the second largest text regions are picked up for the text location presentation. For each text region, the center of gravity of the bounding box is found and it is used for the target location. When the target is close to center of the input image (|x| < 80 and |y| < 60), a simple base sound is played back. When the position is off the center, sound pan and frequency transformation are applied. The sound volume

C. Text location presentation using sound signals

We use some short guidance signals and sound transform to presents the text location in the current view. Considering that multiple text regions exist, the first and the second largest text regions are picked up for the text location presentation. For each text region, the center of gravity of the bounding box is found and it is used for the target location. When the target is close to center of the input image (|x| < 80 and |y| < 60), a simple base sound is played back. When the position is off the center, sound pan and frequency transformation are applied. The sound volume
Figure 1. (a) Input image, (b) Obtained edgemap, (c) Bounding box of CC, (d) Boxes filtered out by DCT, (e) Connected boxes as text string, (f) Detected text region

Figure 2. Text presentation using sound signals

Figure 3. Evaluation of the sound guidance for text regions

Table I

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>mistake(horizontal)</th>
<th>mistake(vertical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>98.3%</td>
<td>1.7%</td>
<td>0.0%</td>
</tr>
<tr>
<td>3D sound</td>
<td>70.0%</td>
<td>28.3%</td>
<td>1.7%</td>
</tr>
</tbody>
</table>

B. Evaluation on presentation with sound

We tested the sound system independently to measure the sound system’s performance alone. We compare HRTF-based 3D sound system to ours. In this experiment, target’s location is given randomly, and the coordinate area is divided into 9 regions (including center, right to left and up and down). We set the number of target as one. The user listens to the sound and answers the perceived direction out of the 9 directions. The subjects are 4 men, and the number of trials is 20 times for each method. The result is shown in Table I, and the sound system used in our system can present target location much better.

IV. CONCLUSION

We have presented a wearable system that presents text position using sound signals. The system uses edge and structure based text region detection method, and uses sound pan and frequency transform. This work was supported by Grants-in-Aid for Scientific Research No.22300194 from JSPS.

REFERENCES


