CHAPTER 1
INTRODUCTION

There are over billions of visually challenged people worldwide. Recent developments in computer vision, digital cameras, and portable computers make it feasible to assist these individuals by developing camera-based products that combine computer vision technology with other existing commercial products such optical character recognition (OCR) systems.

Reading is obviously essential in today's society. Printed text is everywhere in the form of reports, receipts, bank statements, restaurant menus, classroom handouts, product packages, instructions on medicine bottles, etc. And while optical aids, video magnifiers, and screen readers can help blind users and those with low vision to access documents, there are few devices that can provide good access to common hand-held objects such as product packages, and objects printed with text such as prescription medication bottles. The ability of people who are blind or have significant visual impairments to read printed labels and product packages will enhance independent living and foster economic and social self-sufficiency.

Today, there are already a few systems that have some promise for portable use, but they cannot handle product labeling. For example, portable bar code readers designed to help blind people identify different products in an extensive product database can enable users who are blind to access information about these products through speech and braille. But a big limitation is that it is very hard for blind users to find the position of the bar code and to correctly point the bar code reader at the bar code. Some reading-assistive systems such as pen scanners might be employed in these and similar situations. Such systems integrate OCR software to offer the function of scanning and recognition of text and some have integrated voice output. However, these systems are generally designed for and perform best with document images with simple backgrounds, standard fonts, a small range of font sizes, and well-organized characters rather than commercial product boxes with multiple decorative patterns. Most state-of-the-art OCR software cannot directly handle scene images with complex backgrounds.

Several portable reading assistants have been designed specifically for the visually impaired. KReader Mobile runs on a cell phone and allows the user to read mail, receipts, fliers, and many other documents. However, the document to be read must be nearly flat,

placed on a clear, dark surface (i.e., a non-cluttered background), and contain mostly text. Furthermore, KReader Mobile accurately reads black print on a white background, but has problems recognizing colored text or text on a colored background. It cannot read text with complex backgrounds, text printed on cylinders with warped or incomplete images (such as soup cans or medicine bottles). Furthermore, these systems require a blind user to manually localize areas of interest and text regions on the objects in most cases.

Fig 1.1: Few Portable Systems

Although several reading assistants have been designed specifically for the visually impaired, to our knowledge, no existing reading assistant can read text from the kinds of challenging patterns and backgrounds found on many everyday commercial products. Such text information can appear in multiple scales, fonts, colors, and orientations. To assist blind persons to read text from these kinds of hand-held objects, we have conceived of a camera-based assistive text reading framework to track the object of interest within the camera view and extract print text information from the object. The proposed algorithm can effectively handle complex background and multiple patterns, and extract text information from both hand-held objects and nearby signage.

In assistive reading systems for blind persons, it is very challenging for users to position the object of interest within the center of the camera's view. As of now, there are

still no acceptable solutions. We approach the problem in stages. To make sure the hand-held object appears in the camera view, we use a camera with sufficiently wide angle to accommodate users with only approximate aim. This may often result in other text objects appearing in the camera's view (for example, while shopping at a supermarket). To extract the hand-held object from the camera image, we develop a motion-based method to obtain a region of interest (ROI) of the object. Then, we perform text recognition only in this ROI.

It is a challenging problem to automatically localize objects and text ROIs from captured images with complex backgrounds, because text in captured images is most likely surrounded by various background outlier "noise" and text characters usually appear in multiple scales, fonts, and colors. For the text orientations, this paper assumes that text strings in scene images keep approximately horizontal alignment. Many algorithms have been developed for localization of text regions in scene images. We divide them into two categories: rule-based and learning-based.

Rule-based algorithms apply pixel-level image processing to extract text information from predefined text layouts such as character size, aspect ratio, edge density, character structure, color uniformity of text string, etc. Phan et al. analyzed edge pixel density with the Laplacian operator and employed maximum gradient differences to identify text regions. Shivakumara et al. used gradient difference maps and performed global binarization to obtain text regions. Epshtein et al. designed stroke width transforms to localize text characters. Nikolaou and Papamarkos applied color reduction to extract text in uniform colors. In color-based text segmentation is performed through a Gaussian mixture model for calculating a confidence value for text regions. This type of algorithm tries to define a universal feature descriptor of text.

Learning-based algorithms, on the other hand, model text structure and extract representative text features to build text classifiers. Chen and Yuille presented five types of Haar-based block patterns to train text classifiers in an Ad boost learning model. Kim et al. considered text as a specific texture and analyzed the textural features of characters by a support vector machine (SVM) model. Kumar et al. used globally matched wavelet filter responses of text structure as features. Ma et al. performed classification of text edges by using histograms of oriented gradients and local binary patterns as local features on the SVM model. Shi et al. employed gradient and curvature features to model the grayscale curve for handwritten numeral recognition under a Bayesian discriminant function. In our research

group, we have previously developed rule-based algorithms to extract text from scene images. A survey paper about computer-vision-based assistive technologies to help people with visual impairments can be found.

In solving the task at hand, to extract text information from complex backgrounds with multiple and variable text patterns, we here propose a text localization algorithm that combines rule based layout analysis and learning-based text classifier training, which define novel feature maps based on stroke orientations and edge distributions. These, in turn, generate representative and discriminative text features to distinguish text characters from background outlier.

Fig 1.2: Examples of printed text from hand-held objects with multiple colors, complex backgrounds, or non flat surfaces.

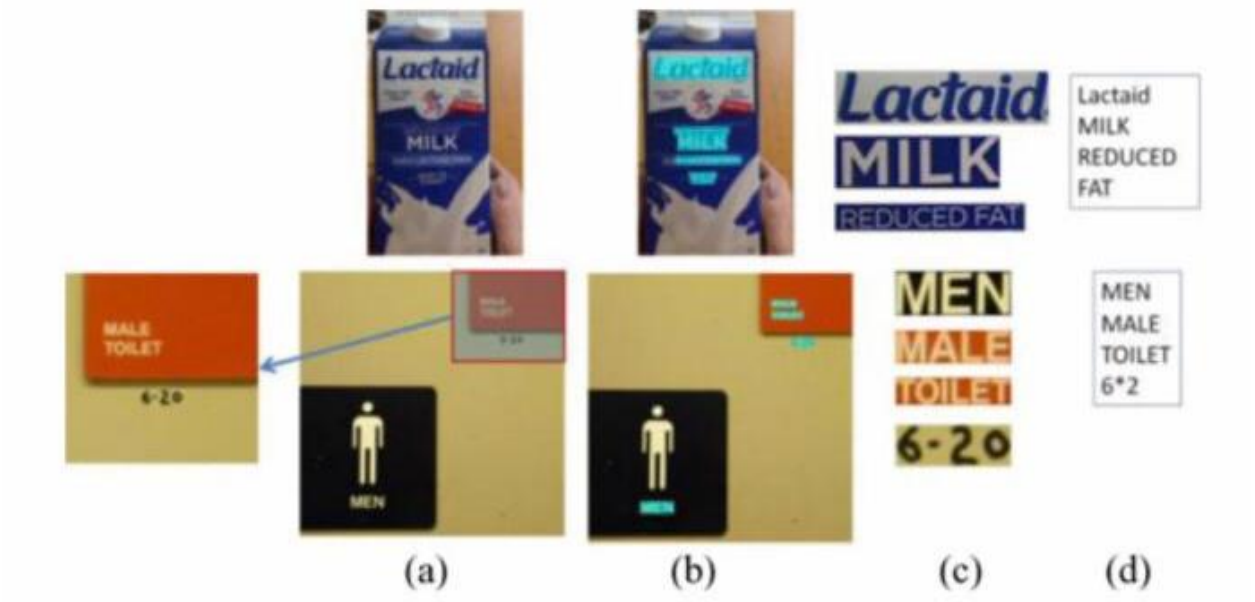

Fig 1.3: Two examples of text localization and recognition from camera captured images. (Top)Milk box. (Bottom)Men bathroom signage. (a)Camera captured images. (b) Localized text regions (marked in blue). (c) Text regions cropped from image. (d) Text codes recognized by OCR. Text at the top-right corner of bottom image is shown in a magnified callout.

CHAPTER 2

LITERATURE SURVEY

1. "Detecting and reading text in natural scenes," in Proc. Comput, X. Chen and A. L. Yuille, Vision Pattern Recognit., 2004, vol. 2, pp. II-366–II-373.

This paper gives an algorithm for detecting and reading text in natural images. The algorithm is intended for use by blind and visually impaired subjects walking through city scenes. We first obtain a dataset of city images taken by blind and normally sighted subjects. From this dataset, we manually label and extract the text regions. Next we perform statistical analysis of the text regions to determine which image features are reliable indicators of text and have low entropy (i.e. feature response is similar for all text images).We obtain weak classifiers by using joint probabilities for feature responses on and off text. These weak classifiers are used as input to an AdaBoost machine learning algorithm to train a strong classifier. In practice, we trained a cascade with 4 strong classifiers containing 79 features. An adaptive binarization and extension algorithm is applied to those regions selected by the cascade classifier. A commercial OCR software is used to read the text or reject it as a non-text region. The overall algorithm has a success rate of over 90%.

2. Automatic detection and recognition of signs from natural scenes:

An approach to automatic detection and recognition of signs from natural scenes, and its application to a sign translation task. The proposed approach embeds multiresolution and multiscale edge detection, adaptive searching, color analysis, and affine rectification in a hierarchical framework for sign detection, with different emphases at each phase to handle the text in different sizes, orientations, color distributions and backgrounds. We use affine rectification to recover deformation of the text regions caused by an inappropriate camera view angle. The procedure can significantly improve text detection rate and optical character recognition (OCR) accuracy. Instead of using binary information for OCR, we extract features from an intensity image directly.

This approach Propose a local intensity normalization method to effectively handle lighting variations, followed by a Gabor transform to obtain local features, and finally a linear discriminant analysis (LDA) method for feature selection. We have applied the approach in developing a Chinese sign translation system, which can automatically detect and recognize Chinese signs as input from a camera and translate the recognized text into English.

3. "Wearable Obstacle Avoidance Electronic Travel Aids for Blind: D.Dakopoulos and N.G.Bourbakis, 2004, vol. 2, pp. II-366–II-373.

The last decades a variety of portable or wearable navigation systems have been developed to assist visually impaired people during navigation in known or unknown, indoor or outdoor environments. There are three main categories of these systems: electronic travel aids (ETAs), electronic orientation aids (EOAs), and position locator devices (PLDs). This paper presents a comparative survey among portable/wearable obstacle detection/avoidance systems (a subcategory of ETAs) to inform the research community and users about the capabilities of these systems and about the progress in assistive technology for visually impaired people. The survey is based on various features and performance parameters of the systems that classify them in categories, giving qualitative-quantitative measures. Finally, it offers a ranking, which will serve only as a reference point and not as a critique on these systems.

CHAPTER 3

SYSTEM ANALYSIS

System analysis as "the process of studying a procedure or business in order to identify its goals and purposes and create systems and procedures that will achieve them in an efficient way". Another view sees system analysis as a problem-solving technique that breaks down a system into its component pieces for the purpose of the studying how well those component parts work and interact to accomplish their purpose. The field of system analysis relates closely to requirements analysis or to operations research. It is also "an explicit formal inquiry carried out to help a decision maker identify a better course of action and make a better decision than she might otherwise have made.

3.1 EXISTING SYSTEM

Walking safely and confidently without any human assistance in urban or unknown environments is a difficult task for blind people. Visually impaired people generally use either the typical white cane or the guide dog to travel independently. But these methods are used only to guide blind people for safe path movement, and these cannot provide any product assistance like shopping etc.

Today, there are already a few systems that have some promise for portable use, but they cannot handle product labelling. For example, portable bar code readers designed to help blind people identify different products in an extensive product database can enable users who are blind to access information about these products. The limitation is that it is very hard for blind users to find the position bar code and to correctly point the bar code reader at the bar code.

3.2 PROPOSED SYSTEM

The system proposes a camera-based label reader to help blind persons to read names of labels on the products. Camera acts as main vision in detecting the label image of the product or board then image is processed internally. And separates label from image, and finally identifies the product and identified product name is pronounced through voice.

Then received label image is converted to text. Once the identified label name is converted to text and converted text is displayed on display unit connected to controller. Now converted text should be converted to voice to hear label name as voice through earphones connected to audio.To read printed text on hand-held objects for assisting blind person in order to solve the common aiming problem for blind users. This method can effectively distinguish the object of interest from background or other objects in the camera view. To extract text regions from complex backgrounds, we have proposed a novel text localization algorithm based on models of stroke orientation and edge distributions. OCR is used to perform word recognition on the localized text regions and transform into audio output for blind users. Furthermore, we will address the significant human interface issues associated with reading text by bind users.

CHAPTER 4

DESIGN

Design is a visual look or a shape given to a certain object, in order to make it more attractive, make it more comfortable or to improve another characteristic. Designers use tools from geometry and art. Design is also a concept used to create an object.

4.1 FRAMEWORK

The framework presents a prototype system of assistive text reading. The system framework consists of three functional components: scene capture, data processing, and audio output. The scene capture component collects scenes containing objects of interest in the form of images or video. In our prototype, it corresponds to a camera attached to a pair of sunglasses. The data processing component is used for deploying our proposed algorithms, including 1) object- of- interest detection to selectively extract the image of the object held by the blind user from the cluttered background or other neutral objects in the camera view; and 2) text localization to obtain image regions containing text, and text recognition to transform image-based text information into readable codes. We use a mini laptop as the processing device in our current prototype system. The audio output component is to inform the blind user of recognized text codes.

Fig 4.1: Snapshot of our demo system, including three functional components for scene capture, data processing, and audio output.

4.2 IMAGE CAPTURING AND PRE-PROCESSING

The video is captured by using webcam and the frames from the video is segregated and undergone to the preprocessing. First, get the objects continuously from the camera and adapted to process. Once the object of interest is extracted from the camera image and it converted into gray image. Use haar cascade classifier for recognizing the character from the object. The work with a cascade classifier includes two major stages: training and detection. For training need a set of samples. There are two types of samples: positive and negative.

4.3 AUTOMATIC TEXT EXTRACTION

In order to handle complex backgrounds, two novel feature maps to extracts text features based on stroke orientations and edge distributions, respectively. Here, stroke is defined as a uniform region with bounded width and significant extent. These feature maps are combined to build an Adaboost based text classifier.

4.4 TEXT REGION LOCALIZATION

Text localization is then performed on the camera-based image. The Cascade-Adaboost classifier confirms the existence of text information in an image patch but it cannot the whole images, so heuristic layout analysis is performed to extract candidate image patches prepared for text classification. Text information in the image usually appears in the form of horizontal text strings containing no less than three-character members.

4.5 TEXT RECOGNITION AND AUDIO OUTPUT

Text recognition is performed by off-the-shelf OCR prior to output of informative words from the localized text regions. A text region labels the minimum rectangular area for the accommodation of characters inside it, so the border of the text region contacts the edge boundary of the text characters. However, this experiment show that OCR generates better performance text regions are first assigned proper margin areas and binarized to segments text characters from background. The recognized text codes are recorded in script files. Then, employ the Microsoft Speech Software Development Kit to load these files and display the audio output of text information. Blind users can adjust speech rate, volume and tone according to their preferences.

CHAPTER 5

METHODOLOGY

Image recognition technology has a great potential of wide adoption in various industries. In fact, it's not a technology of the future, but it's already our present. Such corporations and startups as Tesla, Google, Uber and Adobe Systems etc. heavily use image recognition. To prove that the technology marches around the world let's look at the recent statistics. Researchers predict that the global market of image recognition. That's quite a sound figure. So no wonder that more and more so-called imagetech application that leverage image recognition emerge for various purposes and business verticals.

5.1 FRAMEWORK AND ALGORITHM OVERVIEW

The data processing component is used for deploying our proposed algorithms, including 1) Object-of-interest detection to selectively extract the image of the object held by the blind user from the cluttered background or other neutral objects in the camera view, 2) Text localization to obtain image regions containing text, and text recognition to transform image-based text information into readable codes. We use amin laptop as the processing device in our current prototype system. The audio output component is to inform the blind user of recognized text codes. A Bluetooth earpiece with mini microphone is employed for speech output.

This simple hardware configuration ensures the portability of the assistive text reading system. Figure depicts a work flowchart of the prototype system. A frame sequence *V* is captured by a camera worn by blind users, containing their hand-held objects and cluttered background. To extract text information from the objects, motion based object detection is first applied to determine the user's object of interest *S* by shaking the object while recording video ground from motion-based object detection, and *R* represents the calculated foreground object at each frame. The object of interest is localized by the average of foreground masks.

Next, our novel proposed text localization algorithm is applied to the object of interest to extract text regions. At first, candidate text regions are generated by layout analysis of color uniformity and horizontal alignment. After text region localization, off-the-shelf

OCR is employed to perform text recognition in the localized text regions. The recognized words are transformed into speech for blind users.

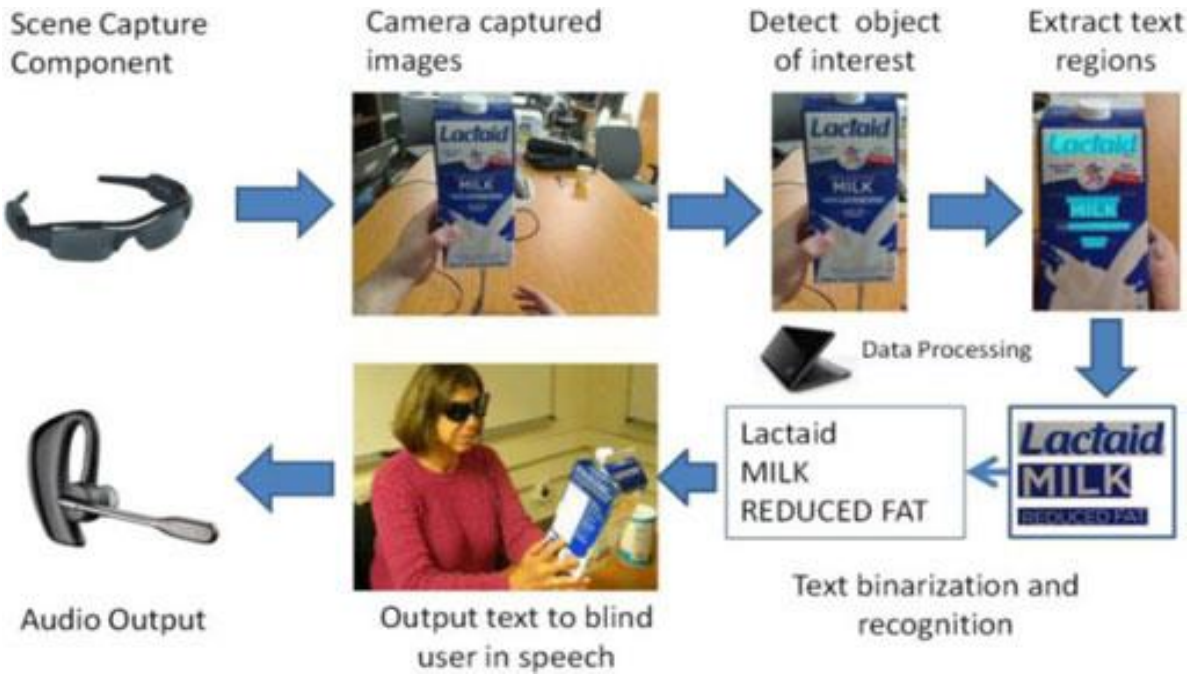

Fig 5.1: Flowchart of the proposed framework to read text from hand-held objects for blind users.

Our main contributions embodied in this prototype system are:

1) Novel motion-based algorithm to solve the aiming problem for blind users by their simply shaking the object of interest for a brief period.

2) A novel algorithm of automatic text localization to extract text regions from complex background and multiple text patterns; and

3) A portable camera-based assistive framework to aid blind persons reading text from hand-held objects. Algorithms of the proposed system are evaluated over images captured by blind users using the described techniques.

5.2 OBJECT REGION DETECTION

To ensure that the hand-held object appears in the camera view, we employ a camera with a reasonably wide angle in our prototype system (since the blind user may not aim accurately). However, this may result in some other extraneous but perhaps text-like objects appearing in the camera view for example, when a user is shopping at a supermarket).

To extract the hand-held object of interest from other objects in the camera view, we ask users to shake the hand-held objects containing the text they wish to identify and then

employ a motion-based method to localize the objects from cluttered background. Background subtraction (BGS) is a conventional and effective approach to detect moving objects for video surveillance systems with stationary cameras. To detect moving objects in a dynamic scene, many adaptive BGS techniques have been developed.

Stauffer and Grimsson modeled each pixel as a mixture of Gaussians and used an approximation to update the model. A mixture of K Gaussians is applied for BGS, where K is from 3 to 5. In this process, the prior weights of K Gaussians are online adjusted based on frame variations. Since background imagery is nearly constant in all frames, a Gaussian always compatible with its subsequent frame pixel distribution is more likely to be the background model. This Gaussian-mixture-model based method is robust to slow lighting changes but cannot handle complex foregrounds and quick lighting changes.

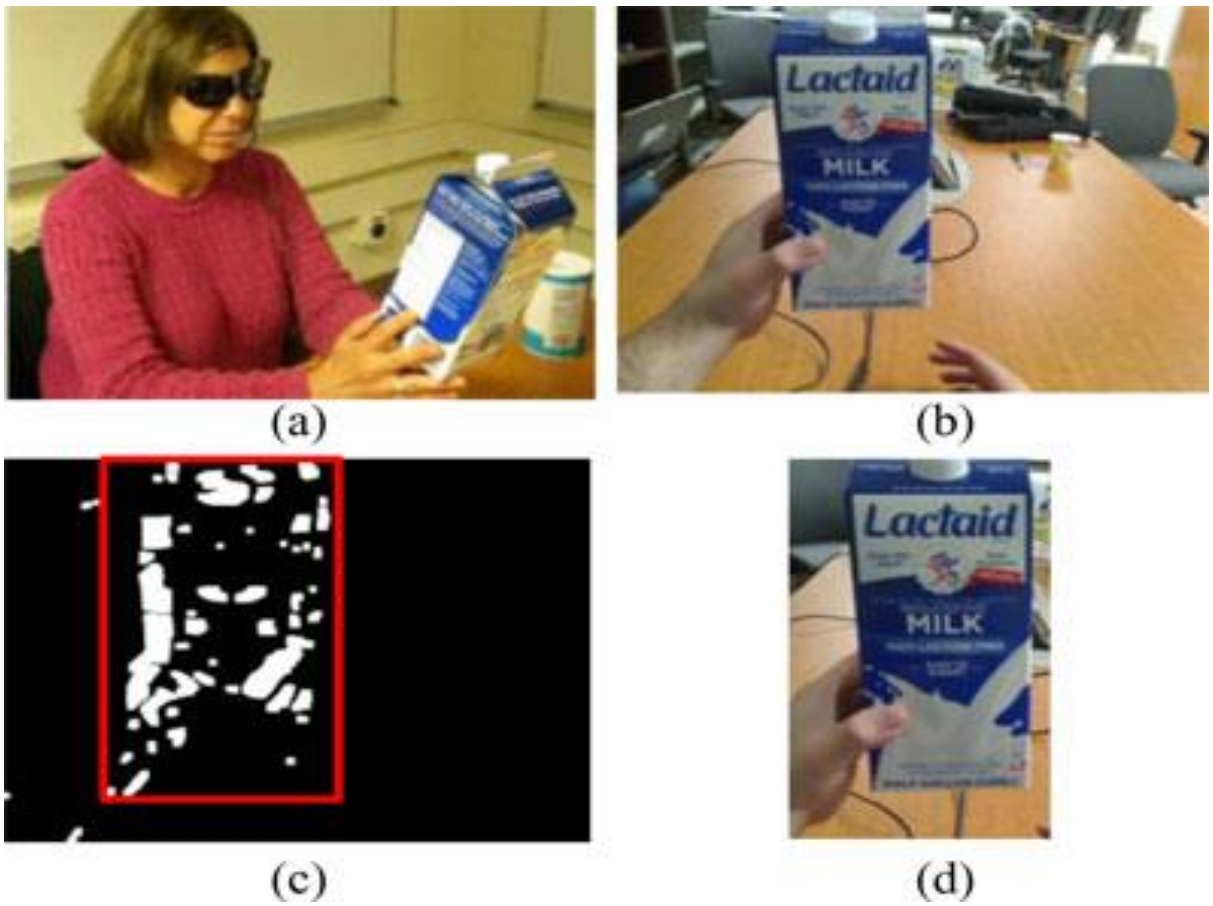

Fig 5.2: Localizing the image region of the hand-held object of interest. (a) Capturing images by a camera mounted on a pair of sunglasses. (b) Example of a captured image. (c) Detected moving areas in the image while the user shaking the object (region inside the bounding box). (d) Detected region of the hand-held object for further processing of text recognition.

Tian et al. further improved the multiple Gaussian-mixture based BGS method to better define foreground while remove background objects. First, texture information is employed to remove false positive foreground areas. These areas should be background but are often determined as foreground because of sudden lighting changes. A texture similarity measure is defined to evaluate whether the detected foreground motion is caused by lighting change or moving object. Second, in addition to quick lighting changes, BGS is also influenced by shadows. Many systems use color information to remove the shadow, but this does not work on grayscale videos. To solve this problem, the normalized cross correlation of the intensities is used for shadow removal. The grayscale distribution of a shadow region is very similar to that of the corresponding background region, except is a little darker. Thus, for a pixel in BGS-modeled foreground areas, we calculate the NCC between the current frame and the background frame to evaluate their similarity and remove the influence of shadow.

As shown in fig 5.2, while capturing images of the hand-held object, the blind user first holds the object still, and then lightly shakes the object for 1 or 2 s. Here, we apply the efficient multiple Gaussian-mixture-based BGS method to detect the object region while blind user shakes it. More details of the algorithm can be found. Once the object of interest is extracted from the camera image, the system is ready to apply our automatic text extraction algorithm.

To detect moving objects in a dynamic scene, many adaptive BGS technique have been developed. Stauffer and Grimsson modeled each pixel as a mixture of Gaussians and used an approximation to update the model. A mixture of *K* Gaussians is applied for BGS, where *K* is from 3 to 5. In this process, the prior weights of *K* Gaussians are online adjusted based on frame variations. Since background imagery is nearly constant in all frames, a Gaussian always compatible with its subsequent frame pixel distribution is more likely to be the background model.

This Gaussian-mixture-model based method is robust to slow lighting changes but cannot handle complex foregrounds and quick lighting changes. Tianfurther improved the multiple Gaussian-mixture based BGS method to better define foreground while remove background objects. First, texture information is employed to remove false positive foreground areas. These areas should be background but are often determined as foreground because of sudden lighting changes.

A texture similarity measure is defined to evaluate whether the detected foreground motion is caused by lighting change or moving object. Second, in addition to quick lighting changes, BGS is also influenced by shadows. Many systems use color information to remove

the shadow, but this does not work on grayscale videos. To solve this problem, the normalized cross correlation of the intensities is used for shadow removal.

The grayscale distribution of a shadow region is very similar to that of the corresponding background region, except is a little darker. Thus, for a pixel in BGS-modeled foreground areas, we calculate the *NCC* between the current frame and the background frame to evaluate their similarity and remove the influence of shadow.

While capturing images of the hand-held object, the blind user first holds the object still, and then lightly shakes the object for 1 or 2 s. Here, we apply the efficient multiple Gaussian-mixture-based BGS method to detect the object region while blind user shakes it. More details of the algorithm can be found. Once the object of interest is extracted from the camera image, the system is ready to apply our automatic text extraction algorithm.

5.3 AUTOMATIC TEXT EXTRACTION

We designed a learning-based algorithm for automatic localization of text regions in image. In order to handle complex backgrounds, we propose two novel feature maps to extracts text features based on stroke orientations and edge distributions, respectively. Here, stroke is defined as a uniform region with bounded width and significant extent. These feature maps are combined to build an adaboost based text classifier.

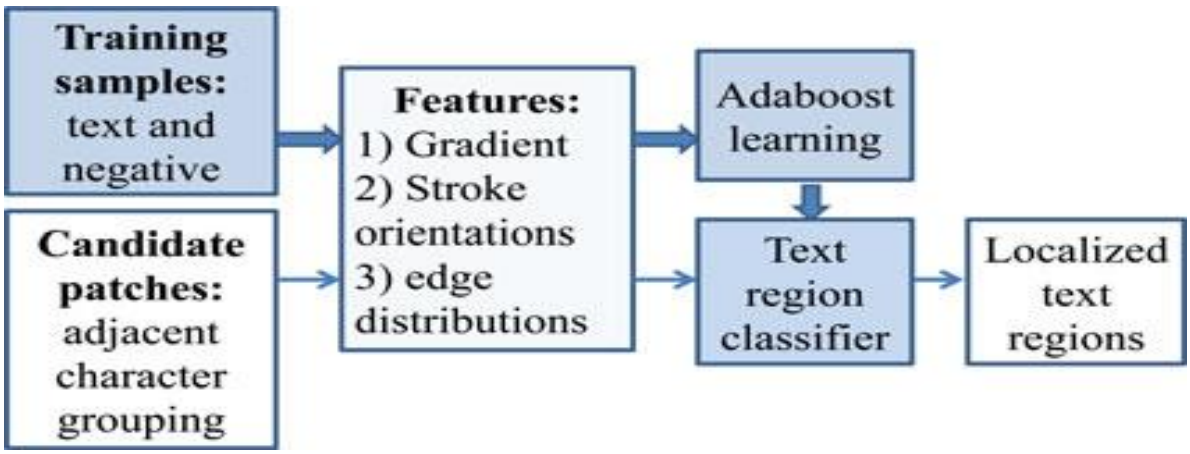

Fig 5.3: Diagram of the proposed Adaboost-learning-based text region localization algorithm by using stroke orientations and edge distributions.

5.3.1 TEXT STROKE ORIENTATION

Text characters consist of strokes with constant or variable orientation as the basic structure. Here, we propose a new type of feature, stroke orientation, to describe the local structure of

text characters. From the pixel-level analysis, stroke orientation is perpendicular to the gradient orientations at pixels of stroke boundaries. To model the text structure by stroke orientations, we propose a new operator to map a gradient feature of strokes to each pixel. It extends the local structure of a stroke boundary into its neighborhood by gradient of orientations. We use it to develop a feature map to analyze global structures of text characters.

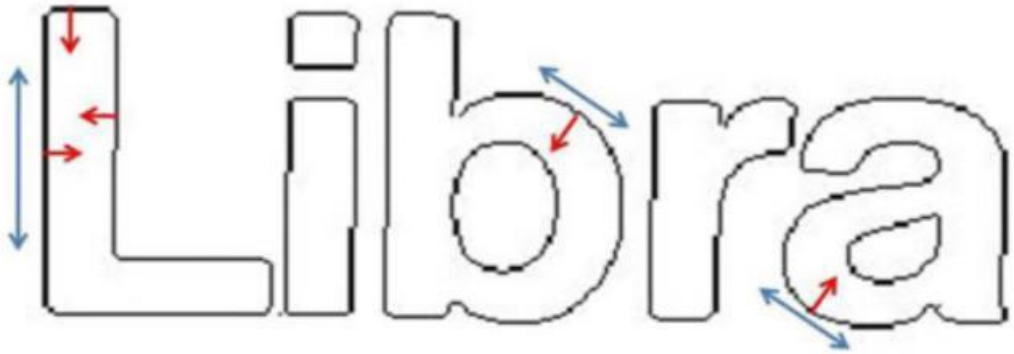

Fig 5.4: Sample of text strokes showing relationships between stroke orientations and gradient orientations at pixels of stroke boundaries. Blue arrows denote the stroke orientations at the sections and red arrows denote the gradient orientations at pixels of stroke boundaries.

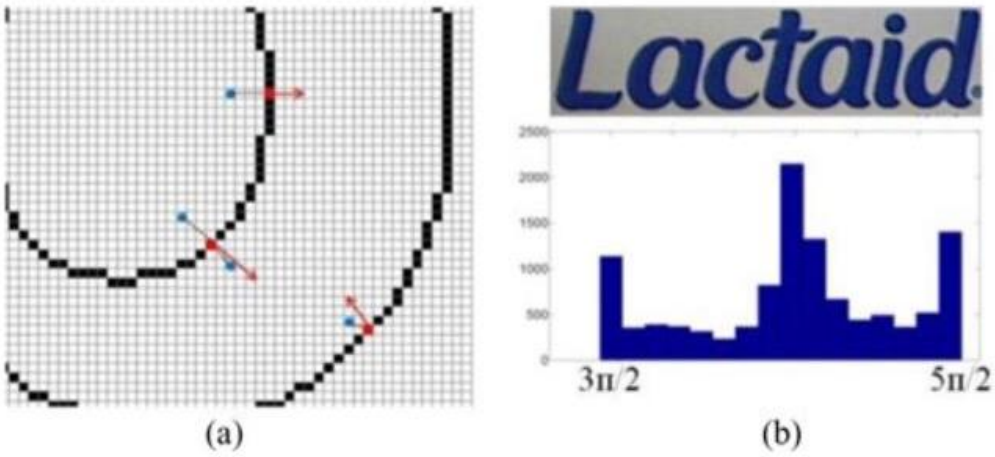

Fig 5.5: (a) Example of stroke orientation label. The pixels denoted by blue points are assigned the gradient orientations (red arrows) at their nearest edge pixels, denoted by the red points. (b)210×54textpatchandits16-binhistogram of quantized stroke orientations.

5.3.2 DISTRIBUTION OF EDGE PIXELS

In an edge map, text characters appear in the form of stroke boundaries. The distribution of edge pixels in stroke boundaries also describes the characteristic structure of text. The most used feature is edge density of text region. But the edge density measure does not give any spatial information of edge pixels. It is generally used for distinguishing text regions from relatively clean background regions.

To model text structure by spatial distribution of edge pixels, we propose an operator to map each pixel of an image patch into the number of edge pixels in its cross neighborhood. At first, edge detection is performed to obtain an edge map, and the number of edge pixels in each row *y* and each column *x* is calculated.

5.3.3 ADABOOST LEARNING OF TEXT FEATURES

Based on the feature maps of gradient, stroke orientation, and edge distribution, a text classifier is trained from an Adaboost learning model. Image patches with fixed size (height 48 pixels, width 96 pixels) are collected and resized from images taken from the ICDAR-2011 robust reading competition to generate a training set for learning features of text. We generate positive training samples by scaling or slicing the ground truth text regions, according to the aspect ratio of width *w* to height *h*.

To train a robust text classifier, we ensure that most positive training samples contain two to four text characters. We build a relationship between the width-to-height aspect ratio and the number of characters of ground truth text regions. It shows that the ground truth regions with two to four text characters have width-to-height ratios between 0.8 and 2.5, while the ones lower than 0.8 mostly have less than two characters and the ones higher than 2.5 mostly have more than four characters. Therefore, if the ratio is $w/h < 0.8$ with too few characters, the region is discarded. If the ratio $w/h \geq 2.5$ corresponding to more than four text characters, we slice this ground truth region into overlapped patches with width-to-height ratio 2:1. If the ratio w/h falls in [0.8,2.5), we keep it unsliced and scale it to width-to height ratio2:1. Then, the samples are normalized into width 96 and height 48 pixels for training.

The negative training samples are generated by extracting the image regions containing edge boundaries of non-text objects. These regions also have width to-height ratio 2:1, and we similarly scale them into width 96 and height 48. In this training set, there are a total of 15 301 positive samples, each containing several text characters, and 35 933 negative samples without containing any text information for learning features of background outliers. Some training examples are shown in the figure.

Fig 5.6: Examples of training samples with width-to-height ratio 2:1. The first two rows are positive samples and the other two rows are negative samples.

5.3.4 TEXT REGION LOCALIZATION

Text localization is then performed on the camera-based image. The Cascade-Adaboost classifier confirms the existence of text information in an image patch but cannot handle the whole image, so heuristic layout analysis is performed to extract candidate image patches prepared for text classification. Text information in the image usually appears in the form of horizontal text strings containing no less than three-character members.

Therefore, adjacent character grouping is used to calculate the image patches that contain fragments of text strings. These fragments consist of three or more neighboring edge boundaries that have approximately equal heights and stay in horizontal alignment. But not all the satisfied neighboring edge boundaries are text string fragments.

Fig 5.7: Adjacent characters are grouped to obtain fragments of text strings, where each fragment is marked by a colored rectangle. The extracted image patches will be processed and input into text classifier.

Thus, the classifier is applied to the image patches to determine whether they contain text or not. Finally, overlapped text patches are merged into the text region, which is the minimum rectangle area circumscribing the text patches. The text string fragments inside those patches are assembled into informative words.

5.4 TEXT RECOGNITION AND AUDIO OUTPUT

Text recognition is performed by off-the-shelf OCR prior to output of informative words from the localized text regions. A text region labels the minimum rectangular area for the accommodation of characters inside it, so the border of the text region contacts the edge boundary of the text character. However, our experiments show that OCR generates better performance if text regions are first assigned proper margin areas and binarized to segment text characters from background.

Thus, each localized text region is enlarged by enhancing the height and width by 10 pixels, respectively, and then, we use Otsu's method to perform binarization of text regions, where margin areas are always considered as background. We test both open- and closed-source solutions that allow the final stage of conversion to letter codes (e.g. Omni Page, Tesseract, ABBY Reader).

The recognized text codes are recorded in script files. Then, we employ the Microsoft Speech Software Development Kit to load these files and display the audio output of text information. Blind users can adjust speech rate, volume, and tone according to their preferences.

5.4.1 DATASETS

Two datasets are used to evaluate our algorithm. First, the ICDAR Robust Reading Dataset is used to evaluate the proposed text localization algorithm. The ICDAR-2003 dataset contains 509 natural scene images in total. Most images contain indoor or outdoor text signage. The image resolutions range from 640 × 480 to 1600 × 1200.

Since layout analysis based on adjacent character grouping can only handle text strings with three or more-character members, we omit the images containing only ground truth text regions of less than three text characters. Thus, 488 images are selected from this dataset as testing images to evaluate our localization algorithm.

To further understand the performance of the prototype system and develop a user-friendly interface, following Human Subjects Institutional Review Board approval, we

recruited ten blind persons to collect a dataset of reading text on hand-held objects. The hardware of the prototype system includes a Logitech web camera with autofocus, which is secured to the nose bridge of a pair of sunglasses. The camera is connected to an HP mini laptop by a USB connection.

Fig 5.8: Examples of blind persons capturing images of the object in their hands.

The laptop performs the processing and provides audio output. In order to avoid serious blocking or aural distraction, we would choose a wireless "open" style Bluetooth earpiece for presenting detection results as speech outputs to the blind travelers in a full prototype implementation.

The blind user wore the camera/sunglasses to capture the image of the objects in his/her hand, as illustrated. The resolution of the captured image is 960 × 720. There were 14 testing objects for each person, including grocery boxes, medicine bottles, books, etc. They were required keep their head (where the camera is fixed) stationary for a few seconds and subsequently shake the object for an additional couple of seconds to detect the region of object of interest. Each object was then rotated by the user several times to ensure that surfaces with text captions are exposed and captured. We manually extracted 116 captured images and labeled 312 text regions of main titles.

5.4.2 EVALUATIONS OF TEXT REGION LOCALIZATION

Text classification based on the Cascade-Ada boost classifier plays an important role in text region localization. To evaluate the effectiveness of the text classifier, we first performed a group of experiments on the dataset of sample patches, in which the patches containing text are positive samples and those without text are negative samples.

These patches are cropped from natural scene images in ICDAR-2003 and ICDAR-2011 Robust Reading Datasets. Each patch was assigned a prediction score by the text classifier; a higher score indicates a higher probability of text information. We define the true positive rate as the ratio of correct positive predictions to the total number of positive samples.

Similarly, the false positive rate is the ratio of correct positive predictions to the total number of positive predictions. Figure plots the variation of true positive against false positive rates. This curve indicates that our text classifier is biased toward negative (i.e., no text) responses because the false positive rate stays near zero until the true positive rate approximately rises to 0.7. This characteristic is compatible with the design of our blind assistive framework, in which it is useful to filter out extraneous background outliers and keep a conservative standard for what constitutes text.

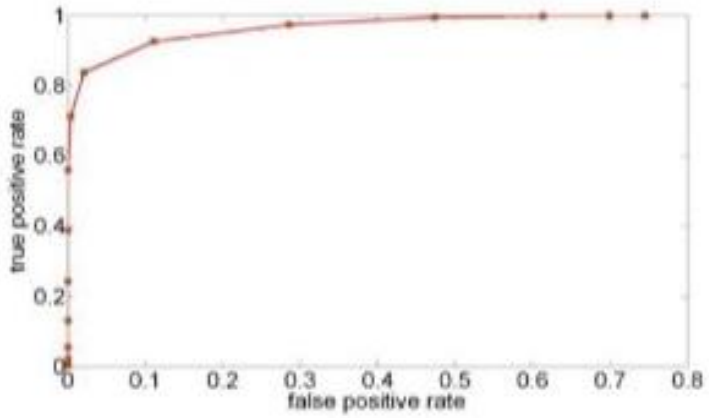

Fig 5.9: Curve of classification performance, where horizontal axis denotes false positive rate and vertical axis denotes true positive rate.

Fig 5.10: Some example results of text localization on the ICDAR-2003 robust reading dataset, and the localized text regions are marked in blue. It shows that our algorithm can localize multiple text labels in indoor and outdoor environments.

Fig 5.11: Some example results of text localization on the ICDAR-2011 robust reading dataset, and the localized text regions are marked in blue. Our algorithm can localize multiple text labels in indoor and outdoor environments.

Fig 5.12: (a) Some results of text localization on the user-captured dataset, where localized text regions are marked in blue. (b)Two groups of enlarged text regions, binarized text regions, and word recognition results from top to down.

Next, the text region localization algorithm was performed on the scene images of ICDAR-2003 Robust Reading Dataset to identify image regions containing text information. Figure depict some results showing the localization of text regions, marked by blue rectangular boxes. To analyze the accuracy of the localized text regions, we compare them with ground

truth text regions and characterize the results with measures we call precision, recall, and f-measure.

For a pair of text regions, match score is estimated by the ratio between the intersection area and the mean area of the union of the two regions Each localized (ground truth) text region generates a maximum match score from its best matched ground truth (localized)text region. Precision is the ratio of total match score to the total number of localized regions.

Some examples of localized text regions are presented in the figure using blue boxes. To improve the performance of blind-assistive technology applications, we adjusted the parameters of text layout analysis to adapt to the hand-held object images.

5.4.3 PROTOTYPE SYSTEM EVALUATION

The automatic ROI detection and text localization algorithms were independently evaluated as unit tests to ensure effectiveness and robustness of the whole system. We subsequently evaluated this prototype system of assistive text reading using images of hand-held objects captured by ten blind users in person.

Two calibrations were applied to prepare for the system test. First, we instructed blind users to place hand-held object within the camera view. Since it is difficult for blind users to aim their held objects, we employed a camera with a reasonably wide angle. In future systems, we will add finger point detection and tracking to adaptively instruct blind users to aim the object. Second, in an applicable blind-assistive system, a text localization algorithm might prefer higher recall by sacrificing some precision. We adjusted the parameters of our text localization algorithm and obtained another group of evaluation result, as precision 0.48, recall 0.72, *f*-measure 0.51. The higher recall ensures a lower miss (false negative) rate. To filter out false positive localizations, we could further employ some post processing algorithm based on scene text recognition or lexical analysis. This work will be carried out in future work.

Next, we evaluated the user-captured dataset of object text. The dataset was manually annotated by labeling the regions of the object of interests and the text regions inside the object of interest regions. In our algorithm evaluation, we defined a region as correctly detected if the ratio of the overlap area of a detected region and its ground truth region is no less than 3/4. Experiments showed that 225 of the 312 ground truth text regions were hit by our localization algorithm. By using the same evaluation measures as above experiments, we obtained precision 0.52, recall 0.62, and *f*-measure 0.52 on this dataset. The precision is lower

than that on the Robust Reading Dataset. The images in the user-captured dataset have lower resolutions and more compact distribution of text information, so they generate low-quality edge maps and text boundaries, which result in improper spatial layouts and text structure features. OCR is applied to the localized text regions for character and word recognition. The Figure shows some examples of text localization and recognition of our proposed framework. We note that the recognition algorithm might not correctly and completely output the words inside localized regions. Additional spelling correction is likely required to output accurate text information. Our text reading system spends 1.87 s on average reading text from a camera-based image. The system efficiency can and will be improved by parallel processing of text extraction and device input/output, i.e., speech output of recognized text and localization of text regions in the next image are performed simultaneously.