Literature Review

The phone reader

Department of Computer Science, Rhodes University

Marilyn Bihina

Supervisor: James Connan

December 3, 2012

1 Introduction

In this literature review, we study and analyze different mobile applications that were designed to read text, recognize text or recognize objects in a picture and inform the user about the result of the request. This project is mainly related to image processing in order to extract text from images. Section 2.2 of this chapter investigates image processing, and Section 2.3 presents applications that can read text from an image. Section 2.4 looks at systems that recognize objects in an image. Section 2.5 presents systems recognizing objects with extended functionalities, followed by section 2.6 which describes the tools needed and used to create the applications presented previously; and the last section, before we conclude, defines a plan of action.

2 Image processing

Image processing is the field of study related to the Phone Reader project. In this section, we give a brief overview of image processing.

2.1 Definitions

Image processing consists of converting an image into a digital form [5], and then performing operations on it such as extracting its content, or the information in it. It is also used for

object recognition.

A digital image is an array of square picture elements or pixels arranged in columns and rows [16] . There are colour images, grayscale images and binary images. Colour images can be converted to grayscale in order to facilitate the extraction of some information within the image.

A grey scale image is an 8-bit image, in which each pixel has an assigned intensity between 0 (black) and 255 (white).

A binary image is an image in which pixels can only have two values: black (0) or white (1).

Most common image formats are: GIF, JPEG, TIFF, PNG, PS, and PSD.

2.2 Image processing methods

There are two types of methods used for image processing [5]:

- Analog image processing or visual techniques of image processing: used for printouts and photographs.
- Digital image processing: processing digital images by using a computer. This technique includes three phases for processing images: pre-processing, enhancement and display, information extraction. Let us briefly define each of those phases [4]:
    * Image pre-processing or image restoration consists of correcting the image from different errors, noise and geometric distortions.
    * Image enhancement improves the visual aspect of the image, after the correction of errors, to facilitate the perception or interpretability of information in the image.
    * Information extraction utilizes the computer's decision-making capability to identify and extract specific pieces of information or pixels.

The different image processing techniques used in the Phone Reader Project help in extracting the text contained in the image taken by the user.

3 Text reading systems

This section presents several mobile applications able to read text from an image. Some of them require OCR (Optical Character Recognition), and others do not.

3.1 Mobile text reading applications requiring OCR

The Phone Reader Application, system designed for the blind and visually impaired, was implemented in 2009 by Computer Science Honours students from the university of Western Cape [8]. This application [9] runs on Android mobile phones with embedded wireless Internet connections, and requires a wireless router to manage the session between the phone and the server, as the Phone Reader server application resides on a server computer. The client-server architecture is used for this system, and is considered to allow faster processing, and to produce better results for the translation, with the use of online dictionaries. When the user takes a photo of a text, that photo is then processed with MODI (Microsoft Office Document Imaging) on the remote server. Afterwards, the TTS engine reads the extracted text from the server through the phone's speaker. But the quality of the image has to be very good in order for the system to produce accurate results. This means that before the image is sent to the Optical Character Recognition (OCR) program, it must already be of a good quality, otherwise the OCR will not be able to efficiently extract the content. This presents a weakness of the system because, as it was created for the blind and visually impaired, it should take into consideration the fact that those users cannot always manage to take good images. Taking good images means being aware of lighting conditions, or of the best position for the camera.

Agnes Kukulska, is an author that highlights the possibility of using a type of "phone reader application" to learn a language in her paper [11]. She discusses how mobile applications can contribute to learning a language with systems like Captura Talk [1], an Android application designed in UK in 2008. She explains that this application enables the user to take a photo of a text and hear the text being read back to him, therefore it is the type of application that can be used for people learning English. This application uses a commercial OCR (ABBYY Mobile OCR Engine 3.00) and a Text-To-Speech (TTS) engine. It can also recognize and translate text in more than twenty languages, and includes a talking word processor.

Similarly, the National Federation of the Blind (USA) and Kurzweil Technologies have created a mobile application for the blind and dyslexic: the knfbREADER [20] in 2008. However, this application and the Captura Talk are not identical; the knfbREADER does not include the talking word processor, but enables the user to enlarge the text for a better view.

[1] www.capturatalk.com

It is integrated in the Nokia N82 cell phone and runs on the Symbian Operating System. The images are processed with Kurzweil image processing software. According to Dr. Marc Maurer, President of the National Federation of the Blind, this was the first cell phone able to read text from a picture taken by the user; and it will promote equal opportunity for the blind. Those two systems have many features and provide good quality images, but are both very expensive.

Contrary to the software described previously, R-MAP Android [19] is a low cost system especially designed for the blind and visually impaired that facilitates their use of mobile phones to read text. This application does not require an Internet connection because it uses an OCR program and a text-to-speech engine integrated in the Android mobile phone designed for blind people, and it does not offer translation of texts. This paper [19] contains a very rich list of related works, which goes from Braille to the knfbREADER. It also lists a large number of OCRs: open source OCR such as GOCR, OCRAD, or Tesseract; and commercial OCRs like ABBYY FineReader or the Microsoft office Document Imaging, also used for the Phone Reader Application [9] mentioned earlier in this section. The cost of very powerful mobile applications for the visually impaired, like the knfbREADER is a concern raised in this paper. All visually impaired people cannot afford such an expensive application. This system provides a user friendly interface, easy to handle for blind users: the buttons to access the different functionalities of the software are located at the corners of the screen and are therefore, easy to access. Also if the text in the picture taken is not readable, the system will inform the user who could make another attempt. The system uses Tesseract OCR and TTS software available on the Android mobile phone. But this application does not seem to be widely known, nor used despite its low cost.

In this subsection, we have reviewed types of mobile systems that can produce accurate audible results from an image with text by using an OCR program. There are different OCR programs, some are free and open source, others are commercial OCRs and can recognize text in a wide variety of languages. Using a commercial OCR program results in creating a costly application.

3.2 Mobile text reading applications not requiring OCR

Trinetra Grocery shopping assistant [13] assists blind users when they are doing their grocery shopping, so that their phone can tells them which item is on the shelf in the store, based on

the product-level identification. The product is identified with the Barcoda Pencil, then its barcode is sent via Bluetooth to the Trinetra system which will read out loud the information found on the product (brand and name) to the user. Trinetra system is meant to be cost-effective for the blind user. The user uses a Barcoda Pencil, a Bluetooth headset, and the Nokia 6620 cell phone. This system does not use an OCR program because there are no pictures taken, but it uses TALKS TTS program to produce the speech to the user. One limit of this system is the fact that the user has to carry two accessories besides his cell phone to use the system, and also has to manage to locate the barcode on each product.

Moreover, Trinetra also offers a currency identifier system [13]. It consists of taking a picture of an US currency note and hear its value. This system uses the phone's camera to take the photo of the currency note and the Microsoft's online Lincoln object-recognition technology to identify the currency note. Many different images of the US currency notes need to be stored in the system's database for it to compare the image sent by the user to the images contained in the database. When a currency note is identified, the TALKS TTS reads out loud the result to the blind user. Once again, there is no OCR required.

In this section, we can notice that extracting readable information in images does not necessarily require the use of an Optical Character Recognition (OCR) program. Instead, data about an object can be stored and retrieved from a database.

4 Object recognition systems

This section presents systems that provide object recognition with different techniques.

4.1 Systems using crowd-sourcing for object recognition

Some systems like Vizwiz [3] or CrowdSearch [22] use crowd-sourcing to improve object recognition from images:

- Vizwiz is a mobile application that runs on iPhone for blind users. It sends a picture taken with a phone to a remote server and ask a question about the content of the image; the answers are given by persons especially employed for that task. The user can then hear the answer to his question, sometimes receiving that answer can take a few minutes, which can be annoying for the user. However, this method provides more accurate results as the answers are not programmed with a computer, but are verified

by humans.

- In this paper about CrowdSearch [22] the author declares that Google Goggles does not work with all types of images because of the high rate of errors that can be provided from a poor quality image taken from a phone, but was mainly designed to recognize building landmarks. CrowdSearch is an image search system of which the results are validated by humans working on the query image. According to the author, local image search on mobile phones can only be efficient energy-wise, but it limits the accuracy of the results whereas remote processing on powerful servers is more accurate and fast. This application was designed for the Apple iPhones and each query sent by the user has a cost, a defined deadline and is validated by human validators after the automated image search. As this author explains, it might really be hard to have very accurate results if the research is done locally because of the limited size of the database located on the phone. The author does emphasize the fact that CrowdSearch is a costly system and that human validation may increase the delay of the results, but it gives more accurate results.

Croud searching is a method that requires human intervention to process a request from a user, but it is a method that provides better results as they are not programmed results. The user is under the impression that he is in a conversation with another person.

4.2 Mobile applications using visual search on specific types of objects

There are systems that offer visual search on some specific types of objects. This means that the user takes a photo of an object and the system will find information on it. But the results are not always audible outputs.

- Kooaba [2] is an image recognition system that runs on iPhone. With it, a user can have information on a picture he has taken with his camera phone. It is mostly designed to recognize prints (newspaper, brochure), paintings, products (books, Cd's), and places. A huge database is used and contains over 28 millions items for object recognition.

- Google Goggles is an image search system [21] that can scan for text contained in a picture using OCR. It can translate text in a few languages [7], but it does not

[2] www.kooaba.com

automatically read out the extracted text to the user. Google Goggles' limited range of image searched is mentioned in papers like the one about CrowdSearching [22] or the one about Mobile Image Recognition [21], which both explain that this application was not designed for the recognition any type of objects, but mostly for landmarks or products.

- SnapTell [3] is a system mentioned in the article about mobile image recognition [10], that runs on Android and iPhone. It was designed for books and CD covers recognition, and is based on the same principle as Kooaba.

Visual search techniques can be used by mobile applications, and also provide accurate results about a user request on a specific object. But those applications are not always well-suited for people with reading disabilities. And without the help of a translation program, they can be useless for non-native speakers, but useful for illiterate users.

5 Text and object recognition systems offering extended functionalities

Some systems allow the recognition of text and object from images, and provide even more functionalities, other than the translation.

As mentioned earlier, this paper [10] highlights the fact that Google Goggles can mostly recognize landmarks and logos; and that camera phones generally produce low quality images contrarily to the paper about Mobile Product Recognition [21]. The mobile image recognition system presented in this paper captures and recognizes video frames of a document (mostly newspaper), not snapshots of images, enabling the system to process different frames of the same video until one is recognized. The video taken concerns the same document captured from different angles. It is not a client-server architecture because this author believes that that type of architecture causes delays to respond to the user, but instead it is a client-only architecture in which the database remains on the phone and is periodically updated from the database generator. The size of the database is computed by estimating the number of newspaper pages a user can want to read daily, and in this case it reached about 10.5MB. Once an image is recognized, a thumbnail of the document is displayed with "hot spots" which

[3] www.snaptell.com

are each linked to some electronic content. The user can then access that electronic content by clicking on a "hot spot". The image recognition algorithm used in this system processes the image to improve its low quality. The recognition process takes about 6 seconds. This idea sounds original, but does not seem very useful as the output is not read to the user, and there are no translations available. That makes it unusable for visually impaired people, or for the non-native speakers. It does not seem to be useful for a wide range of users. In this case, only persons who can see and read can use this system, and if a user has access to the Internet, as this system requires, and needs information on a newspaper, the easiest way would be to open that newspaper website, and click on each interesting link. The necessity of this system is not clearly explained in this paper.

Like the Trinetra Grocery shopping assistant [13], this research article [12] presents a system that only provides an audible output, and it describes the knfbREADER as a system having a high level of accuracy for text recognition only for documents without many colours or images; and the AdvantEdge Reader as an efficient portable scanning and reading device only for flat documents. The device presented in this paper is called Multifunctional assistant for the blind, which is a PDA phone (a personal organizer assistant and cellphone combined in one device [15]) that uses a multi-functionalities system including an in-house optical character recognition(OCR), audio messages recording, object recognition, banknotes recognition, colour recognition and the ability to listen to audio speeches or records. Depending on the user's level of vision, the system offers different designs of user interfaces. The only output is a synthetic voice used to guide the user when he is using the system. To recognize objects, they propose a tag-based object recognition which will enable the user to stick a dedicated tag on an object, take a photo of the tag, and record a description of that object. Hence each time the photo of this object will be taken, the system will then play its previously recorded message to the blind user. The banknote recognition in this system is made for the Euro currency but the author mentions the possibility of easily extending the system to recognize other currencies. Once a picture of a banknote is taken, this system starts by locating the "region of interest" which is the zone on the banknote where its value will be read; this region is the tag in the case of object recognition. The image is then processed and only the numbers representing the value of the banknote are extracted and read to the user. The colour recognition is also done by processing a picture taken and identifying the main colour of the object. But the problem of the quality of the picture taken from a camera phone remains because of

the fact that OCR engines are meant to work with scanned documents, this system can only process 1.3 megapixels images. The approach aims to use specific tags for similar-by-touch objects, but this can cause a problem when the system needs to differentiate them and verify if a tag already has an accompanying description. The number of tags required for the classification of similar-by-touch objects can be very large, which means that the user might have to regularly buy them.

Contrarily to the previous authors, the author of the Mobile Product Recognition paper (MPR) [21] considers the camera phone as a very efficient tool for image processing and defines a few image retrieval systems like Google Goggles, SnapTell, or Kooaba as too slow to process a query compared to his system. It is a system similar to Google Goggles, a mobile visual search system as defined in the paper, meaning that when an image is taken by the user, the system seeks information about it on a remote server using a wireless Internet connection. The MPR system uses a set of functionalities such as feature compression image to accelerate the query processing time, and achieves the queries in less than 2 seconds. The system does not send the whole query image to the server, but just its smaller compressed features. The image processing is implemented on the phone and the query is processed by a search in the database. But this author's point of view about the efficiency of a camera phone for image processing is not shared by many other authors, all the previous authors have an opposite point of view, and even in this Science Daily article [18], the camera phone is described as a not efficient enough tool to take good quality images.

In conclusion, reading a text from an image can be extended to seeking information about that image. It can be the colour of an object, the value of a banknote or even a special description assigned by the user to the object, or image. Some applications do not work with snapshots from a camera, but with video frames, to find the best frame to process.

6 Common tools used by object recognition and text reading mobile applications

To develop a mobile application that will process an image and then read the content to the user, we have noticed that essential tools are required.

6.1 Mobile operating systems

- **Android**

  Android [17] is an operating system developed by Google's team for mobile devices. It is an open development Linux based platform. It was originally designed to support applications written in Java, but with Android NDK (Native Developpement Kit) [1], applications can also be written in C or C++. A Text-to-Speech engine is also provided with the Android mobile phones [19]. Android applications can be developed using Windows, Mac OS and Linux systems [6], thus no special hardware is required. The most recommended IDE for developing Android apps is Eclipse. Dalvik is the virtual machine environment for mobile devices created by Google, and when projects are compiled, each application runs on its own VM, not on the Java VM.

- **iOS**

  iOS is the mobile operating system created by Apple that runs on the iPhone since 2007. Its development requires Macintosh computers running on Mac OS X [6]. iOS apps are written in Objective-C with the modern IDE Xcode, quite similar to Netbeans, Eclipse, or Visual Studio.

6.2 OCR

An Optical Character Recognition (OCR) program is a program that converts an image that contains text into an editable format [14]. It can take as input different document formats such as PDF, PNG, TIF and produces an editable document as an output of various formats like TXT, DOC, or PDF. In order to extract the text from a scanned document, or a digital image, the OCR performs the following steps: after loading the input image, the OCR detects features like the resolution and inversion, and font size. Each OCR expects the image to have a predefined background or foreground colour (most often back and white). A deskewing algorithm can also be applied when necessary. Finally there is a page layout analysis that is performed to detect position of important areas in an image.

6.3 Text-To-Speech Engine

A text-to-speech software or speech synthesis program is used to read a text out loud [2] to a user. The one included in Android 1.6 can also translate texts in English, French, German,

Italian and Spanish. A TTS engine uses natural human voices to avoid sounding like a robot during the reading of a text to a user. There are also many different free TTS engines available online where the user can choose the voice he prefers the text to be read with.

7 Plan of Action

To implement this Phone Reader mobile application, we will perform the following tasks:

1. Develop a web server on which a user can upload images that need to be processed.
2. Pre-process the images that have been uploaded in order for the OCR to recognize words more easily.
3. Evaluate the pre-processing of images.
4. Send the image to the OCR for text extraction.
5. Implement a user interface for the client.
6. Perform the translation of the extracted text, if required by the user.
7. Connect the application to a Text-To-Speech software for the reading of the text obtained to the user.
8. Test and improve the application.
9. Write the documentation of the project.

8 Conclusion

We have reviewed different mobile applications designed to help users with reading disabilities when they come across a text they cannot read. We can then say that the Phone Reader project is not a brand new topic, related works have been implemented since 2008 by major companies whose software are expensive. A few other systems have been implemented by researchers that are not big companies, but are not widely used. Nevertheless, those applications designed for visually impaired can also be used for non-native speakers when they include a translation functionality, and for illiterates if they generate audible results. Furthermore it has been noticed that camera phones often take low quality images, hard to process efficiently

by an OCR. This is the reason why to implement this Phone Reader application, we will have to mainly work on the image processing aspect, so that the OCR will easily recognize characters from the text in the processed image. Thus the TTS engine will be able to give more accurate results.

.

References

[1] Android. What is the NDK? Online. Available from: `http://developer.android.com/tools/sdk/ndk/overview.html`.

[2] Android. TextToSpeech. Online, November 2012. Available from: `http://developer.android.com/reference/android/speech/tts/TextToSpeech.html`.

[3] Bighamy, J. P., Jayant, H. C., Jiy, H., Littlex, G., Miller, A., Millerx, R. C., Millery, R., Tatarowiczx, A., Whitez, B., Whitey, S., and Yehz, T. Vizwiz: Nearly realtime answers to visual questions. *UIST* (October 2010).

[4] Dr. Kumar, N. Digital image processing techniques for image enhancement and information extraction. *Proceedings of Workshop on Remote sensing and GIS applications in water resources engineering, Bangalore IV* (1997), 33–43.

[5] Dr Rao, K. Overview of image processing. *Readings in Image Processing* (25-26 September 2004), 1–7.

[6] Goadrich, M. H., and Rogers, M. P. Smart smartphone development: ios versus android. *SIGCSE* (March 2011).

[7] Google. Introducing Google Play. Online, 2012. Available from: `https://play.google.com/store/apps/details?id=com.google.android.apps.unveil&hl=en`.

[8] Hossein, and Shayesteh, H. The Phone Reader. Online, 2009. Available from: `http://www.cs.uwc.ac.za/~hashayesteh/`.

[9] Hossein, and Shayesteh, H. Phone reader application. Tech. rep., University of Western Cape, Faculty of Computer Science, November 2009.

[10] Hull, J. J., Liu, X., Erol, B., Graham, J., and Moraleda, J. Mobile image recognition: Architectures and tradeoffs. *HotMobile* (February 2010).

[11] Kukulska-Hulme, A. Learning cultures on the move: where are we heading? *Journal of Educational Technology and Society 13(4)* (2010), p4–14.

[12] Mancas-Thillou, C., F. C., Demeyer, J., Minetti, C., and Gosselin, B. A multifunctional reading assistant for the visually impaired. *EURASIP Journal on Image and Video Processing Volume 2007, Article ID 64295, 11 pages* (September 2007).

[13] Narasimhan, P., Gandhi, R., and Rossi, D. Smartphone-based assistive technologies for the blind. *CASES* (October 2009).

[14] Nicomsoft. Optical Character Recognition (OCR) How it works. Online, 2012. Available from: `http://www.nicomsoft.com/optical-character-recognition-ocr-how-it-works/`.

[15] pcmag.com. Definition of: PDA phone. Online, 2012. Available from: `http://www.pcmag.com/encyclopedia_term/0,1237,t=PDA+phone&i=56917,00.asp`.

[16] Sachs, J. Digital image basics. *Digital Light & Color* (1999), 1–14.

[17] Saha, A. K. A developer‘s first look at android. *Linux for you* (January 2008), 48–50.

[18] ScienceDaily. Better pictures with mobile devices. *ScienceDaily* (12 October 2011).

[19] Shaik, A. S., Hossain, G., and Yeasin, M. Design, development and performance evaluation of reconfigured mobile android phone for people who are blind or visually impaired. *SIGDOC* (September 2010).

[20] Technologies, K., and the National Federation of the Blind. First cell phone that reads to the blind and dyslexic. *Voice of the Nation's Blind* (27 January 2008), 1–3.

[21] Tsai, S. S., Chen, D., Chandrasekhar, V., Takacs, G., Cheung, N., Vedantham, R., Grzeszczuk, R., and Girod, B. Mobile product recognition. *MM* (October 2010), 1–4.

[22] Yan, T., Kumar, V., and Ganesan, D. Crowdsearch: Exploiting crowds for accurate real-time image search on mobile phones. *MobiSys* (October 2010).