Camera Reading For Blind People

Roberto Neto\textsuperscript{a}, Nuno Fonseca\textsuperscript{a}

\textsuperscript{a} Polytechnic Institute of Leiria, Leiria 2411-901 Leiria, PORTUGAL

Abstract

Blindness makes life rather difficult for people who suffer from this health problem, but the use of technology can help in some day-to-day tasks. In this context, the present work focuses the development of a photo-to-speech application for the blind. The project is called Camera Reading for Blind People, and its ultimate purpose is the development of a mobile application that allows a blind user to "read" text (a sheet of paper, a signal, etc.). To achieve that, a set of frameworks of Optical Character Recognition (OCR) and Text to Speech Synthesis (TTS) are integrated, which enables the user, using a smartphone, to take a picture and hear the text that exists in the picture.

1. Introduction

Blind people are unable to perform visual tasks. For instance, text reading requires the use of a braille reading system or a digital speech synthesizer (if the text is available in digital format). The majority of published printed works does not include braille or audio versions, and digital versions are still a minority. On the other hand, blind people are not able to read the simple warnings in walls or signals that surround us. Thus, the development of a mobile application that can perform the image to speech conversion, whether it's a text written on a wall, a sheet of writing paper or in another support, has a great potential and utility.

The technology of optical character recognition (OCR) enables the recognition of texts from image data. This technology has been widely used in scanned or photographed documents, converting them into electronic copies, which one can edit, search, play its content and easily carry [7].

The technology of speech synthesis (TTS) enables a text in digital format to be synthesized into human voice and played through an audio system. The objective of the TTS is the automatic conversion of sentences, without restrictions, into spoken discourse in a natural language, resembling the spoken form of the same text, by a native
This technology has had significant progress over the last decade, with many systems being able to generate a synthetic speech very close to the natural voice. Research in the area of speech synthesis has grown as a result of its increasing importance in many new applications [19].

This paper describes the process of developing a prototype of a mobile application (iOS), which can be used on an iPhone/iPod Touch/iPad, which allows a blind user to use the device camera and to get the reading of the existing written text in the captured image. The system uses already existing OCR and TTS frameworks, combining them in a way that, taken together, they can provide the desired results.

This paper is organized according to the following structure: in section 2, an overview and the state of art regarding OCR and TTS technologies are presented; in section 3, the system is presented; section 4 presents some additional steps to improve the results; and finally, some conclusions and future work are presented.

2. Overview and state of art

This part presents an introduction to OCR and TTS technologies used in this project as well as studies and work within the theme of this project.

2.1. Optical Character Recognition (OCR) overview

Optical character recognition, usually designated by the acronym OCR, is the process of recognition and automatic conversion of existing characters in the written-support image into the text format, which can then be used in various applications. The OCR has been widely studied and has displayed considerable advances regarding the performance and accuracy of the obtained results [9].

The process of optical character recognition can be summarized as a process that follows a set of steps [6]:

- Optical image acquisition
- Location and segmentation
- Preprocessing
- Feature extraction
- Classification
- Postprocessing

The final text is then converted into the desired document format (rtf, txt, pdf, etc.).

2.2. Text-to-Speech (TTS) overview

Voice synthesis, defined as TTS (acronym for Text-To-Speech), is a computer system that should be able to read aloud any text, regardless of its origin [5].

The use of TTS aims to produce human voice artificially. Voice synthesis is a complex process and complex algorithms are needed to produce an intelligible and natural result. TTS synthesis makes use of techniques of Natural Language Processing. Since the text to be synthesized is the first entry of the system, it must be the first to be processed.

There are several techniques to create a synthesized voice [4]:

- Articulatory synthesis
- Formant synthesis
- Concatenation synthesis
- Hidden Markov models synthesis

The main synthesis techniques, presented above, are the methods used in the study and development of speech synthesis systems. However, a way to profit from the inherent advantages of each technique is to use a hybrid of the various techniques in the development of future systems speech synthesis.
The quality of a speech synthesis can be determined by its naturalness and intelligibility. Naturalness is the characteristic that describes how close to the human voice is the sound obtained by TTS; intelligibility refers to the ease with which the sound is understood in complex situations [5,12].

2.3. State of Art

There are some works being done within the image-to-speech research area.

The Phone Reader thesis [3] presents an approach to a system that allows recognizing the text in a photo using the Tesseract OCR, and the reading of the text is made through the TTS, in the selected language. The OCR processing and translation are made on an external server, which requires a data connection for communication between the server and the device.

In the paper Optical Character Recognition [14], an application is presented, that performs the recognition of the existing text in photos via OCR Tesseract and, after that, the speech synthesis of recognized text is made. In this application, all processing is done locally on the mobile device, and there will be no preprocessing or recognition of the limits of the images, so these will be defined as future work.

The paper Open Source OCR Framework Using Mobile Device [20] describes a project in which the image is captured with the device camera, and is converted to a bitmap format. The multi-channel RGB image is converted into single-channel intensity image. Then it made a postprocessing, removing the noise generated by OCR, one based on ASCII code in the recognized text and all alphabetical and numerical characters are removed, no filter being applied. This is indicated as the future work, refining the results by using OCR cross-referencing with the dictionary, and other high level semantic techniques.

In OCRdroid: A Framework to Digitize Text Using Mobile Phones [10], the operation of a scanning system for obtaining text contained in images is discussed. This is a project whose aim is to develop an application for Android, which makes a preprocessing of the images as they are captured, correcting the existing lighting and checking if the images are properly positioned for better recognition of the text, and then sends the obtained image to an external server which processes the image recognition OCR.

Some public mobile applications also exist in this area, that are able to recognize text using OCR and reading it using the system internal TTS (like the iOS VoiceOver feature [2]):

- SayText [16] is an application that enables you to take photographs of texts, and after recognizing the text, you are allowed to send it by email or obtain the reading of it.
- Talking Camera Pro, for visually impaired/blind [18], is an application that recognizes text via OCR, synthesizes it and returns its reading.
- Prizmo - Escaneamento, OCR & Fala [15], is an application which allows recognizing various image and document formats, the result can be exported as PDF / Text, vCard or JPEG / PNG. The application can make the processing of images obtained and also the synthesis of the text.

Despite the existence of studies and applications that address the theme proposed in this project, it seems that existing solutions continue to present several limitations, which relate to the efficiency in recognition, particularly in situations where shooting conditions are not ideal. This research area still needs to achieve greater maturity to overcome the limitations of real-life situations where light conditions, the focus, alignment or the existence of shadows do not allow the effective recognition by OCR.

The project Camera Reading for Blind People is therefore an additional contribution to the area, since it provides additional experiences on the subject and presents some foundations for the development of related applications for iOS devices.

3. Proposed Architecture

In this section, the methodology to pursue the project is presented, from image capture with the camera, through implementation of OCR and TTS tools.

The overall system architecture is represented by the following figure 1:
3.1. **Image capture with the camera**

The camera of the mobile device is critical to use the application, since it will be essential for the user to take the picture of the brackets containing text that will be recognized and synthesized.

The image capture is done inside the application itself, thus avoiding the use of additional applications such as access to the photo gallery.

The device used in the project was an Apple iPhone 5. The display screen size is 4.0’ with a resolution of 640x1136 pixels. The camera of the device has a resolution of 8 megapixels with 3264x2448 pixels, autofocus and LED Flash.

The application will run in full screen mode. The captured image will be the one displayed on the screen. The autofocus and auto flash camera options will be set as switched on.

3.2. **Implementation of OCR tools**

After the image is obtained, the next step is the optical recognition of existing text in it, through use of an OCR framework.

Optical character recognition, usually known as OCR, is the interpretator of images that have text and, after they are scanned, the text contained in them can be recognized, edited and used.

It was found that there are some frameworks available to iOS, having been selected three for testing: the open source framework Tesseract and the commercial frameworks Leadtools and Abbyy.

The Tesseract framework has no costs and the remaining solutions have costs: Leadtools $995 [11] (as seen on March 28th, 2014) and in the case of the framework Abbyy €4900 [1] (as seen on 28th March, 2014). It must be referred that the use of commercial frameworks required the contact with the respective companies and filling out forms to obtain three-month licenses for product reviews.

3.3. **Choice of OCR framework**

To compare the efficiency of recognition of each framework, a set of 30 pictures of texts with different shapes, sizes and layouts was collected. We tried to get a set of images that represent a variety of situations as embracing as possible, with texts in different shapes, sizes and colors, different alignments for the images, different exposures to light at the time of image capture and shadows in an attempt to get the most accurate representation of real situations that will surely happen when using the application.

Some images used in the tests are presented below in figure 2:
Then the text of each photo was transcribed for individual text files in order to be able to make a comparison between the text obtained after optical image recognition and the original text.

3.4. Levenshtein distance

For measurement values that can be used in the comparison between the texts, the optically recognized and the original, a function that reproduces the string comparison algorithm Levenshtein distance was developed.

The Levenshtein distance algorithm [13] was created by the Russian scientist Vladimir Levenshtein in 1965. The basic principle of the algorithm is the measurement of similarity between two strings. This is done by calculating the number of basic operations or changes needed to make two strings equal. Changes may consist of a substitution, insertion or deletion of a character. The Levenshtein distance \( d(p, t) \) between two strings (\( p \) and \( t \) represent the strings) is the minimum number of operations required to make \( p = t \). For example, if \( p = "moose" \) and \( t = "moody" \), the Levenshtein distance is \( d(p, t) = 2 \) since the strings can be made using the same two substitutions (replacement of letter 's' by letter 'd', exchange of letter 'e' by letter 'y').

Taking another example, the Levenshtein distance between the words "kitten" e "sitting" is 3, since with only 3 changes one can make a word turn into another, and there is no way to do it with less than three changes [17].

- kitten
- sitten (substitution of ’k’ for ’s’)
- sittin (substitution of ’e’ for ’i’)
- sitting (insertion of ’g’ at the end)

Tests were made with the 30 images being recognized by OCR, and the difference obtained between the recognized text and the original text was verified. The total of characters of the original texts was registered and the Levenshtein distance obtained between the original text and the recognized text was calculated, as well as the percentage value of that distance, obtained by dividing the distance obtained by the total of characters of the text.

It was then calculated the median value of the distance values obtained. Unlike the average, the median is not influenced by extreme values, since it is essentially linked to the position it occupies in the ordered set. Thus, if some value is too big or small – outliers –, it will not affect the calculation of the median, since it does not alter the order.

The results of Levenshtein distance for the three tested frameworks are presented in the following graph in figure 3:
Figure 3 – Comparison of the string distance from the tested images in frameworks Tesseract, Abbyy and Leadtools (the y-axis values are on a logarithmic scale)

The comparison result of the median of the distances obtained in each of the frameworks is analyzed in the following graph presented in figure 4.

Figure 4 – Comparison of the median value of the string distance in the tested images in frameworks Tesseract, Abbyy and Leadtools

Although the commercial frameworks have better results than the free framework (Tesseract - 23.45%, Abbyy - 18.76% and 18.81% Leadtools), the research work was based on Tesseract framework due to budget limitations.

3.5. Implementation of the tool Text-To-Speech

There are several available frameworks TTS, some free, such as the case of GoogleTTS and OpenEars, or others commercial, such as the case of iSpeech or AcapelaTTS.

Several tests with free frameworks were performed, verifying very similar results in the synthesis of human voice with the provided strings. In spite of existing several frameworks TTS, iOS 7 now supports native TTS support, though the use of a new class has been included in AVFoundation framework, the AVSpeechSynthesizer class, that will be used because it allows to obtain the synthesis of voice desired for the application, without resort to external frameworks.
4. System Optimization

Using the framework Tesseract, we proceeded to a set of tests in order to seek better results in the recognition of images, and greater efficiency in the use of the application. In order to do this, two stages were added: a preprocessing and a postprocessing.

4.1. Preprocessing

Given that the image used in the optical recognition process has a considerable impact on the results, particularly regarding image quality, brightness, contrast, focus, among other features, the project proceeded to attempt to make changes to images shot by the device, to improve the result in his reading OCR.

Thus, we used image filters in order to make a preprocessing of the images so that we could get better results.

A set of tests was run to obtain the median percentage of Levenshtein distance, resulting from the difference between the original text and the recognized text, after the application of each of these filters to the set of the 30 images collected for testing.

Various filters available for iOS (Core Image [8]) were tested, namely CIToneCurve (adjusts tone response of the R, G, and B channels of an image), CIHueAdjust (changes the overall hue, or tint, of the source pixels) CIGammaAdjust (adjusts midtone brightness), CIColorControls (adjusts saturation, brightness, and contrast values), CIColorMonochrome (remaps colors so they fall within shades of a single color), CIExposureAdjust (adjusts the exposure setting for an image similarly to the way you control exposure for a camera when you change the F-stop).

It should be noted that, in the tests for each filter, the values of its various parameters were changed in cycle, seeking the best results. The values for the parameters of each filter were calculated this way, which allowed us to obtain better results. First the filters were tested individually, and after that they were combined trying to find even better results.

After running several tests, it was shown that the combination of filters CIColorControls and CIColorMonochrome produced the best results, so this filter combination will always be applied to images captured by the device immediately before the OCR recognition process.

The CIColorControls filter allows you to adjust the saturation, brightness and contrast in an image. It receives as parameters a CIImage image, and the values for the referred attributes can be changed. The used values were 1.00 to saturation, 0.50 to brightness and 2.14 to contrast.

The CIColorMonochrome filter allows remapping the colors in an image so that they fall within shades of a single color. It receives as parameters a CIImage image, and the values for the input color and the intensity can be changed. The applied values were 0.77 for the intensity and "white" as the color for which it is intended to remap the others.

The results obtained after application of the filter combination are as follows in figure 5:

![OCR Tesseract - Levenshtein distance](image)

**Figure 5** – Values of string distance obtained after applying the filter combination CIColorControls and CIColorMonochrome
After applying this combination of filters, there was a substantial improvement in the results, the average value of the distance between the original text and the recognized decreases from 23.45% to 17.83%.

4.2. Postprocessing

During the testing of the application, particularly when using the image captured with the camera of the device, it was found that in certain readings, the result of OCR recognition was a set of meaningless characters, that in no way reflected the text in the images. This result was caused by the poor quality of the captured image, or OCR recognition errors. As expected, the oral synthesis of these situations turned out to be misleading and unpleasant to the user.

After several tests were performed, and observing the resulting optical text recognition, it was found that the normal pattern of strings in the case of incorrect readings presented several special characters indicated below:

( ) ; * ~ » “ . =

The indicated character set was based on the observation of several samples tested, and can naturally be modified in future versions, or new characters to be included in this set, simply by including them in the function.

To avoid such a situation, regular expressions were used to test if the text returned by the OCR corresponded to a readable text or if it returned many occurrences of the listed special characters. Thus, the created function receives the text recognized by OCR, counts the total number of characters in the text, and counts the total occurrences of the above special characters, calculating a percentage value of these special characters in relation to the total characters of the text, referred to as error rate, obtained via the following formula 1:

\[
\text{Error rate} = \frac{\text{Occurrences of special characters}}{\text{total characters of the text}}
\]

(1)

If this percentage exceeds the defined value, which can be configured by the user, the application presents an error in processing, instead of performing the synthesis of the text, prompting the user to repeat the process of capturing the text.

In order to assess the initial percentage value for the application, and to set the range of values to be changed, we proceeded to a set of tests using the set of images already used in other tests. The percentage of allowed special characters was defined and, when these values were exceeded, the situations were considered as processing error.

The graphical representation of the values obtained is presented below in figure 7:

![Figure 6 – Median values of the distance after post processing](image_url)

By analyzing the results, it can be seen that the smaller the percentage of allowable special characters, the better the results of the obtained percentage of the median error are. This is a consequence of the fact that the images with
worst OCR recognition results are discarded. The exceptional downward trend in the percentage of the median occurs in the value of 1%, since in this value only the reading of two images are regarded as valid, so this value can not be considered as a reference value for drawing some conclusions from their use.

It was found that for percentages below 3% the amount of discarded images is very high, exceeding 50% of the 30 images tested. The use of such a low rate may make the implementation little effective, often showing the error message of recognition.

The default value for the first use of the application shall be 5%, since, after the tests carried out using this value as a reference, it was found that, although the 5 discarded images of the all 30 tested, corresponding to a percentage of approximately 16.7%, the percentage value of the median of distance obtained is 13.43%, which corresponds to a substantial improvement in the figure registered without error prevention, 17.83%.

Of course the user may choose to change the set value for the percentage, which may be increased to get more incorrect results and fewer error messages, or decreased to get more effective results but a larger number of situations of recognition error.

5. Conclusion

This paper presents the development of the project Camera Reading for Blind People, considering OCR and TTS stages, to create an application that was gradually improved and refined over the project.

An analysis was made regarding the OCR and TTS technologies that were used in the development of the application, in order to know the methods behind those, and to understand in greater detail the mechanisms that perform optical character recognition on images and speech synthesis of texts.

The project consisted of the construction of an application composed by several parts, integrating the system of image capture by the mobile device, which is used by an OCR framework for recognition of its text, which is then synthesized through a process of TTS.

Optimizations carried out for improving outcomes resulted in a more efficient application, capable of responding to the challenge set by the theme of the project: a camera that reads texts for the blind.

To improve the quality performance of the system, a pre-processing of the image was carried out before submitting it to optical recognition. After considering various possibilities of image treatment, the choice fell on the use of a combination of filters (CIColorControls and CIColorMonochrome), which showed substantially improved results.

Another optimization applied to the project was an analysis of the optical recognition result in order to assess its usefulness as a presentable result. It was found that, in many situations, due to incorrect OCR reading, the final result represented a set of meaningless characters, but which represented text, consequently the TTS performed their synthesis, which caused an upsetting situation to the user. Thus, a function that uses regular expressions to analyze the result of OCR recognition has been developed; if the result presents a significant percentage of special characters, the system will indicate that there was a reading error and it will ask the user to repeat the process.

Although the final application does not represent a fully usable and totally reliable solution for blind people in daily life yet, the developed program and the obtained results can already be considered interesting, verifying that in some images, the ideal conditions being met, the results are quite interesting, occurring recognition and consequent reading, both very effective and consistent with the original text.

The final result achieved is not perfect, since it has shortcomings regarding the recognition of images in real time because the recognizing process of large and complex texts is sometimes slow. Also, the fact that the process of capturing the images does not provide an automatic system to aid the user to orient the image capture as correctly as possible presents itself as a limitation of this application, since the application is aimed at blind people, who may have more difficulties in accomplishing this task. However, this limitation can and should be targeted for improvement in future work.

The research, implementation and optimization developed allowed the design of a free application that is already in a state of possible use, even with the limitations referenced above, allowing the reading of texts, provided that the light conditions are the ideal for image recording and the equipment is properly directed to the text you want to listen to so that recognition and reading are as satisfactory as possible.
5.1. Future work

The use of commercial frameworks is something that needs to be evaluated, since it displayed better results than the free framework when used without any filter, which suggests that the use of these frameworks in images that have been submitted to filters to improve recognition will lead to even more satisfactory results.

The dataset used for this project could be increased, by seeking a broader and more representative set of real use situations.

Another test that could be carried out to assess the quality of the system involves comparing the perceived audio line with the actual text, asking users to write what they thought they had heard and finally making the comparison between the understood text and the original text.

Another issue that should be targeted for improvement is recognition in real time.

The alignment of the image is a critical aspect that needs future research work. If the blind user register an image with incomplete text (cropped words on the sides of the image), the final result cannot be useful. Therefore, in the future the application should assist the user in aligning the device with the text to be recognized. Recognition can be eventually done through a process of reading the image pixels, seeking to find the limits between lighter patterns, where possibly there is no text, such as the borders of a page, and darker patterns, where potentially the text is located.

Support for other languages will be another issue to be reviewed. At this stage, the project was designed and developed for the English language, but in the future it should allow use in other languages, thus extending the number of people who can benefit from the advantages of the application.

It is also reserved for future work the increment of the menu options, allowing the user to change other application options such as the language or the setting of sounds and orientations.

References

- [1] ABBYY, Abbyy DSKs for developers, online available in http://www.abbyy-developers.eu/en/business/pricing
- [2] APPLE, Voice Over para OS X, online available in http://www.apple.com/pt/accessibility/osx/voiceover/
- [3] BIHINA, M. M. B. The Phone Reader. Grahamstown, South Africa, November, 2012.
- [4] CONDADO, P., Quebra de barreiras de comunicação para portadores de paralisia cerebral. Doutoramento em Engenharia Electrónica e Computação, Ciências de Computação, Universidade do Algarve, 2009
- [5] DUTOIT, T. A Short Introduction to Text-to-Speech, Kluwer Academic Publishers, Dordrecht, Boston, London, 1997.
- [6] EIKVIL, LINE. OCR Optical Character Recognition, 1993
- [7] ELMORE, M. AND MARTONOSI, M. A morphological image preprocessing suite for ocr on natural scene images, 2008
- [8] FONSECA, N., REIS, C., SILVA, C., MARCELINO, L., CARREIRA, V. Desenvolvimento em iOS iPhone, iPad e iPod Touch, Lisboa, FCA – Editora de Informática, Lda, 2012.
- [9] FUJISAWA, H. A view on the past and future of character and document recognition. In Proceedings of the Ninth International Conference on Document Analysis and Recognition – Volume 01, pages 3-7; Washington, DC, USA. IEEE Computer Society, 2007.
- [10] JOSHI, A., ZHANG, M., KADMAWALA, R., DANTU, K., PODURI S. AND SUKHATME, G.S. OCRdroid: A Framework to Digitize Text Using Mobile Phones. Computer Science Department, Electrical Engineering Department University of Southern California, Los Angeles, CA 90089, USA, 2009
- [11] LEADTOOLS, Leadtools web order, online available in http://www.leadtools.com/resources/weborder.htm
- [12] LEMMETTY, S. Review of Speech Synthesis Technology. Masters Thesis, Helsinki University, 1999.
- [13] MANGNES, B. The use of Levenshtein distance in computer forensics. Master's Thesis. Master os Sciente in Information Security. Department of Computer Science and Media Technology, Gjøvik University College, 2005
- [14] MITHE, R., INDALKAR S., DIVEKAR, N. Optical Character Recognition. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-1, March 2013
- [15] PRIZMO, Prizmo - Escaneamento, OCR & Fala. Disponível online em https://itunes.apple.com/PT/app/id366791896
- [16] SAYTEXT, Saytext, online available in https://itunes.apple.com/PT/app/id37637999
- [17] SILVA, A. Uma Avaliação Experimental da Combinação de Métricas de Similaridade para o Alinhamento de Ontologias através de Mineração de Dados. UNIRIO, 2013. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.
- [18] TALKING CAMERA PRO, Talking Camera Pro - for visually impaired/blind, online available in https://itunes.apple.com/PT/app/id544295997
- [19] THOMAS, S. Natural Sounding Text-To-Speech Synthesis Based On Syllable-Like Units, Department Of Computer Science And Engineering Indian Institute Of Technology Madras, 2007.
- [20] Zhou, S.Z., Open Source OCR Framework Using Mobile Devices, Multimedia on Mobile Devices 2008. Edited by Creutzburg, Reiner; Takala, Jarno H. Proceedings of the SPIE, Volume 6821, article id. 682104, 6 pp. (2008)