Embedded Reading Device for Blind People: a User-Centred Design

Peters, Jean-Pierre; MANCAS-THILLOU, C.; FERREIRA, S.

Published in:
Proc. of Emerging Technologies and Applications for Imagery Pattern Recognition

Publication date:
2004

Document Version
Peer reviewed version

Embedded Reading Device for Blind People: a User-Centred Design

Jean-Pierre PETERS,
University of Namur, Dept of Psychology, Belgium,
jpp@psy.fundp.ac.be

Céline THILLOU,
Faculté Polytechnique de Mons, Belgium,
celine.thillou@tcts.fpms.ac.be

Silvio FERREIRA,
Faculté Polytechnique de Mons, Belgium,
silvio.ferreira@tcts.fpms.ac.be

Abstract

A handheld PDA-based system is being developed to help blind people in their daily tasks. The design combines in a continuous process Users' involvement and Engineers' effort. This interaction is made efficient thanks to a specialist able to communicate with both parties and to extract useful knowledge for them. The system can be viewed as a main loop including the User taking the snapshot, Text/Picture detection, Optical character recognition, Text-to-speech synthesis, Feedback to the User, until a useful output is reached. Each task is carried out by algorithms integrating both technical performance and user requirements.

1. Introduction

Giving disabled people the greatest accessibility to their environment is the objective of CRETH, the Centre for Resources and Evaluation of Technologies adapted to the Handicapped at the Department of Psychology of the University of Namur (FUNDP). The action of this Centre is threefold: accompanying the handicapped people and offering them the most appropriate available technology for a better integration in the social and occupational life, promoting and helping the development of emerging technologies able to fill the gap between individuals and their environment, and developing a theoretical research based on the accumulated knowledge about the "people - technologies" binomial relationship.

Based on recent scientific breakthroughs about character recognition, speech synthesis and on the recent evolutions of technologies integrating digital cameras and powerful data processing in Personal Digital Assistant, SYPOLE [1] allows to link all these techniques together to improve the daily life of partially-sighted or blind people and to give them a better social integration.

The challenge is to combine in a small device constraints coming from the needs of the disabled, their environment and the absence of immediate feedback. This requires a deep collaboration between the partners: the User, the psychologist (mentioned below as the "Mediator") and the Engineer. How to manage human aspects and the engineering sphere will be discussed.

2. A user-centred approach

For many years, a lot of centres for ergonomics and human factors developed theories and practical advice for the designers. Yet very often they are far away from the technical centres and their conclusions do not seem relevant enough to be taken into account in the technical design. In the case of disabled people, an insidious factor is added: people believe they know what is suitable or not for the handicapped people. This situation is similar to what can be observed in a hospital with the patients. So, those people are not really involved in the design procedures. But things are gradually getting better: from a situation where the market was the only witness of the success (or failure), the users of the adapted technologies were then requested to validate the development done, later to make some suggestions; and nowadays they are generally solicited during the specification phase. So, the involvement of the handicapped users has moved towards the beginning of the design process. But what we experiment in the Sypole project is the benefit of a continuous interaction between the users and the engineers during the whole development process. This step requires a specific strategy.

First of all, a third actor is included in the staff: a disabled people practitioner (the Mediator). This member of the development team is, in this detailed project, a psychologist, but for other research we also work with occupational therapists or other related profiles. Much is expected from him (her), such as thorough and practical experience of the disabled's world, a continuous follow-up of the everyday life of the handicapped and at the same time a good technical

background. Considering both aspects is of primary importance to act as an active interface between the two spheres: a good understanding of the daily needs of the partially-sighted person and a good perception of the potentialities of the developed technologies. Direct contacts between the engineers and the handicapped remain important and they have to continue; but if this approach remains the only one, we will only be able to solve particular cases. The intervening mediator provides a more relevant analysis: he gives a better understanding of the needs and the wishes, he has the skill and the tools to classify the requirements depending on the handicap and on the objectives of the social re-insertion; finally, he offers the generalisation of the Users' advice.

All of this is studied and tried at the CRETH centre [2] where psychologists, functional therapists and technicians work together to suggest to the handicapped people accurate solutions based on new technologies (if those are suitable). Existing equipments are searched, classified, tested in laboratory, assessed in each individual case and sometimes given for a period of trial. Indeed, such systems are often expensive and the first choice has to be the right one. The other task, – more long term-oriented – is to promote new developments and possibly actively participate in them. An interesting example is described below.

3. A user-centred design: SYPOLE

Though we seldom stop to think about it, sighted individuals are continually bombarded every day by the printed world. Some of the sources of this abundance of print media in our environment include transportation, advertising, news and commercial signs. However, this is a phenomenon that people who are blind currently do not experiment. Most of their vision troubles prevent them from having access to textual information. Even the process of eating out is complicated by the fact that few restaurants have menus in Braille.

All of these facts underscore the necessity for a portable, autonomous, small-size and easy to use automatic text reader. With the emergence of multimedia technology and powerful mobile devices like a PDA, it is now possible to imagine an inexpensive system able to capture images in real time and transform image into speech information. Algorithms are currently being developed to characterise the visual content of images and to recognise text by Optical Character Recognition (OCR). Our objective is to make these algorithms working in real time into a device devoted to helping people who are blind or visually impaired.

While several devices have been developed in the past to assist the reading of printed text, they have all fallen short of the user expectations. Most have been too cumbersome or not readily available to be practical and truly portable. Sometimes they even create more problems than they solve. This is why a user-centred approach has been chosen for this development.

First of all, precious resources should be focused on the problems that will have the greatest impact on the day-to-day life of the user. In that way, short text and standard pictures like labels, logos, stamps, banknotes, CD boxes, etc. form a good starting point. In order to enhance the integration of the User in the process, we adopt a strategy of successive refinements. The User can practice partial results and therefore give the developer useful information.

Figure 1: Ergonomics of the system based on the PDA fitted with a camera.

From a current scanned image, classic OCR can segment text and pictures, and run the recognition. In our case, we have to face three major constraints: the PDA technology, the environment, and the person who takes the snapshot. The PDA (see figure 1) has to offer the human interface (keys and voice recognition), process the image and pronounce some feedbacks and the answer, all of this with limited resources of computational power and memory. The environment is all around the User, from the lighting conditions to the system itself. Illumination very often causes many problems: lack of contrast, variable brightness along the image, reflection on the surface, etc. Moreover, the User cannot provide help to the system. So, relevant information can be out of the picture, geometric

distortion appears due to bad camera or object positioning, one finger can overlap a part of the field, movement can be induced by pressing a key on the PDA; the list is far from being exhaustive.

As a good capture is better than many long pre-processing, special effort is undertaken to enhance the handling of the camera. We can see such a trial on figure 2. A software tool is developed to train the User to take a photo. The goal is to obtain a photo appropriate to the algorithms and not just an artistic photo, which is very different! Thanks to this tool, the Mediator can receive from the User interesting data on the manipulation, enter the problem carefully and give relevant feedback in two main topics: the Man-Machine interface and the accurate choice of the algorithms. A derived product will become a training tool for the User.

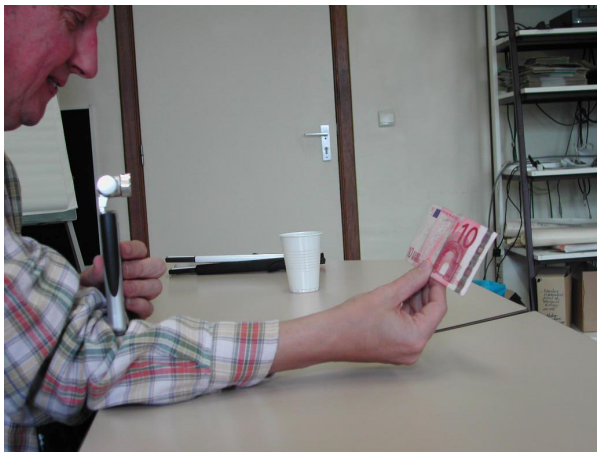

Figure 2: User trying to take a snapshot of banknote

As far as the Man-Machine interface is concerned, it has to be developed from the beginning, contrary to what is often done when the major scientific difficulty in a project is image analysis with high constraints. What is essential is immediately involving the User in the product being developed.

Three key technologies are required for our automatic reading system: text detection, optical character recognition and speech synthesis. All these technologies are described in the next section. For further information, the automatic reading system is detailed on [3].

4. Text Recognition System

4.1 Text detection

Traditionally, document images are scanned with a flatbed, sheet-fed or mounted imaging device. However, digital cameras have shown their potential as an alternative imaging device. But camera-based images require specific processing. The first is detection and localization of the text regions. The idea is to locate the text elements without necessarily recognizing them, cut them out of the image, determine the reading order and finally correct their perspective. Actually, camera-captured images present a bunch of degradations, missing in scanner-based ones, such as blur, perspective distortion, uneven lighting, moving objects or sensor noise. In our case, the user's movement can in addition cause unstable input images and computational resources are limited.

The mainstream of roman languages text regions is characterized by the following features [4]:

- Characters contrast with their background since they are designed to be read easily.

Characters appear in clusters at a limited distance aligned to a virtual line. Most of the time the orientation of these virtual lines is horizontal since that is the natural writing direction.

Text detection techniques can broadly be classified as edge [5] [6], color [7] [8], or texture-based [9] [10]. Edge-based techniques use edges information in order to characterize text areas. Edges of text symbols are typically stronger than those of noise or background areas. These methods operate essentially in greyscale format and do not require much processing time. Nevertheless, they do not cope with complex text images like pictures of magazines or scene images where edge information alone is not sufficient to separate text from a noisy background. The use of color information allows the image to be segmented into connected components of uniform color. A reduction of the color palette is often required. The main drawbacks of this approach consist in the high color processing time and the high sensibility to uneven lighting and sensor noise.

Texture-based techniques try to capture certain texture aspects of text. In our approach, the document image consists of several different types of textured regions, one of which results from the text-content in the image. Thus, we pose the problem of locating text in images as a texture discrimination problem.

Our method for texture characterization is based on Gabor filters which have been used earlier for a variety of texture classification and segmentation tasks [11]. We use a subset of Gabor filters proposed by Jain and

Farokhnia [12] associated with an edge density measure. These features are designed to identify text paragraphs. Each individual filter will still confuse text with non-text regions but an association of filters will complement each other and allow text regions to be identified unambiguously. We use a reduced K-means clustering algorithm to cluster feature vectors. In order to reduce computational time, we apply the standard K-means clustering to a reduced number of pixels and a minimum distance classification is used to categorize all surrounding non-clustered pixels. Empirically, the number of clusters (value of K) was set to three, a value that works well with all test images. The cluster whose centre is closest to the origin of feature vector space is labelled as background while the furthest one is labelled as text. Several text detection results are shown on .Figure 3.

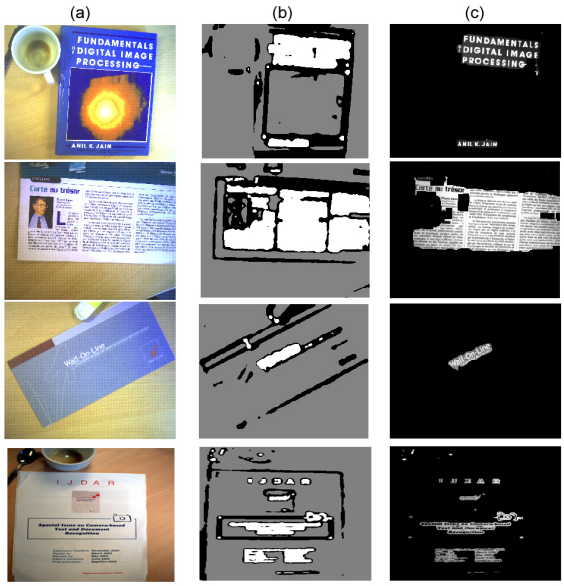

Figure 3:. Several text detection results. (a) Original images (b) Text region clustering (c) Final results

4.2 Perspective correction

As previously suggested, the User disabilities introduce specific constraints for text recognition task. A common problem relates to the position between the "text object" and the camera. The User cannot be certain that the document and the camera are placed facing each other. Documents that are not frontal-parallel to the camera's image plane will undergo a perspective distortion. In general, supposing that the document itself is on a plane, the projective transformation from the document plane to the image plane can be modelled by a 3-by-3 matrix in which eight coefficients are unknown and one is a normalization factor. The removal of perspective can be achieved once the eight unknowns have been found.

The early optical character recognition (OCR) systems were often tested against documents under ideal conditions. Usually, images were electronically converted from paper with a scanner, so the surface texture was fairly even, lighting was well distributed, and the image was captured at an overhead angle. If any of these variables were to be slightly altered, however, many of the assumptions that made these systems successful would not apply. We will describe in the next sections all the steps we implemented to take into account these problems for a better character recognition.

4.3 Denoising and Binarisation

Once text is extracted, it is embedded into a more or less close bounding box and the understanding of this text area can begin.

In many document processing systems, grey-level documents images are first binarised to form two-level (black/white) images. The binarisation process involves the assignment of pixels to either foreground or background objects. The process is often achieved by global or local thresholding [13]. Both global and local thresholding schemes make use of the assumption that foreground and background pixels can be classified by comparing their intensity values with some prescribed or automatically selected thresholds.

In our case, the embedded constraint and the User induce several degradations.

For document images corrupted by various kinds of noise, the assumption is violated wherever there is an outlier, and the binarized images may be severely blurred and degraded. It is, thus, a common task to preprocess the input text region using certain noise deletion algorithms such as our wavelet-based denoising [14]. This particular denoising method has the main advantage to be very general and to either improve image or do nothing but especially not to further degrade images. This is a very important point for our application, because images to handle are various and the User cannot help us on the kind of noise which affects the image.

A following binarization method is applied using color information [17] to extract text as properly as possible.In Figure 4, the usefulness of a proper binarisation algorithm is shown by explaining the result for a blurred color image with different methods. After denoising and color clustering, the result is quite better.

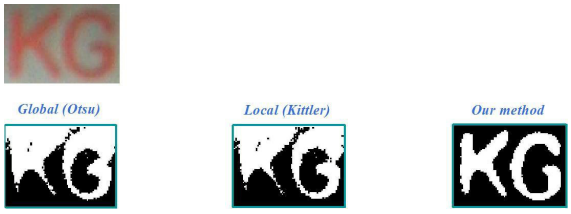

Figure 4: Importance of a proper binarisation: from left to right, global Otsu method, local Kittler method and our method using wavelet-based denoising and color information.

4.4 Character Segmentation

The following step is character segmentation in order to separate each character from the others. Other possibilities based on segmentation-free recognition [15] can be applied. In our case, images are taken under various conditions and as the User can not help on character fonts or sizes, most of these techniques based on correlation matching are not possible. Moreover, they are often more accurate with many assumptions, heavy and not convenient for our embedded platform. Therefore, character segmentation still has to be applied.

The traditional connected component algorithm devised by Rosenfeld and Platz [16], takes advantage of divisions between regions by labeling them as distinct objects.

Another major point in character segmentation is broken and touching characters. In natural scene images, the first category seldom occurs because recognizable characters are thick enough. Nevertheless, touching characters are often present in these kinds of images due to some perspective, large thickness, blur, etc.

Thanks to our algorithm using color information [17] with respect to non-connectivity, we have already minimized the number of touching characters. For those ones left, a very common tool is the caliper distance between the uppermost and bottom most pixels in each column to find the place where to separate components.

4.5 Character Recognition and Error Correction

For character recognition, we use a back-propagation neural network with one hidden layer and 180 nodes in it. Based on edges and geometry, the input vector contains 47 features after a principal component analysis and the output vector is 36 classes for the 26 latin letters (capital or not) and the 10 digits.

Our training database was built with 29,000 degraded characters and our test database, with different samples, is based on around 1,000 degraded characters from a personal database of camera-based images and from ICDAR 2003 robust reading competition.

Our results are 83% for characters extracted from a quite uniform background and we have not had any relevant results for complex backgrounds yet. Actually, our system is currently under process to handle highly degraded characters from complex backgrounds to be robust to give relevant and comparable results.

It is really important not to disappoint blind or visually impaired people and not to give false results because in some applications, the impact can be harmful. For example, in banknote recognition, the User has to be sure of the banknote value which is easy to understand. Therefore, we have included in our system a strict rejection class for this purpose.

To correct some recognition errors, we add one more step. We take into account confidence levels of OCR output in order to choose the right character in the N-best list instead of always considering the best recognized one. This step has to be carried out quickly and not to require too many resources because of our constraint of embedded platforms. Instead of a dictionary we use character n-grams of a given language by using a large database of sentences from a French newspaper during 10 years. In order to find the best path to get the right word, we use the well-known Viterbi algorithm which computes maximal probabilities by iteration.

In order to give an audio answer for blind or visually-impaired people, the key step is the Text-to-Speech synthesizer eLite [18] we use. It is a multilingual research platform which easily deals with important linguistic issues: complex units detection (phone numbers, URLs), trustworthy syntactic disambiguation, contextual phonetization and acronym spelling. Moreover, important research, still in progress, will be integrated within eLite, like page layout detection or non-uniform units-based speech synthesis. Actually, it is very important to have a high-quality speech synthesis in order to get a user friendly device.

5. Conclusion and future work

We described the whole process to build an automatic text recognition system for the blind or visually impaired. The approach that we have chosen aims at fostering the interaction between the Engineer and the User in order to get a product adapted to the latter's needs. The intervention of a mediator, a

psychologist in this case, makes mutual understanding easier and allows the opinions or suggestions expressed by the User to be generalized.

Nevertheless, this algorithm needs further improvements to handle more types of images (more complex backgrounds...) in order to be more robust. Computationally speaking, each step requires an optimized code to run efficiently on a PDA in a realistic waiting period for the User.

6. Acknowledgements

We want to thank the TTS department of Multitel, who provided the E-lite speech synthesizer for our research and Prof. Dr. Bernard Gosselin, Head of Image Group of FPMs, for the direction of the project.

We would also like to express our thanks to V. Collin, Psychologist who is in charge of the Users involvement in Sypole Project, G. Bazier, CRETH administrator and M. Mercier, Head of Department of Psychology, FUNDP.

Finally we also thank Ch. Minetti and F. Dubois, FSA / ULB,

This project is called Sypole and is funded by the Ministère de la Région wallonne in Belgium.

7. References

[1] SYPOLE project website, http://tcts.fpms.ac.be/projects/sypole/sypole.html

[2] M. Mercier "A center of resources and evaluation of technologies for the social and professional insertion of handicapped people (CRETH)", Proceedings of HUSITA5 Conference on Social Services in the Information Society.

[3] S. Ferreira, C. Thillou, B. Gosselin, "From Picture to speech: an innovative OCR application for embedded environment", Proc. of the 14th ProRISC workshop on Circuits, Systems and Signal Processing (ProRISC 2003), Veldhoven (Netherlands).

[4] R.Lienhart, A.Wernicke, "Localizing and segmenting text in images, videos and web pages", IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, n.4, pp. 256-268, 2002

[5] A.K. Jain and B. Yu, "Automatic Text location in images and video frames", Pattern Recognition, vol. 31, n.12, pp. 2055-2076, 1995

[6] M. Pietikaïnen and O. Okun, "Text extraction from grey scale page images by simple edge detectors", Proceedings of the 12th Scandinavian Conference on Image Analysis, pp. 628-635, 2001

[7] W.-Y. Chen and S.-Y. Chen, "Adaptative page segmentation for color technical journals' cover images", Image and Vision Computing, vol. 16, pp. 855-877, 1998

[8] Y. Zhong, K. Karuand A.K. Jain, "Locating text in complex color images", Pattern Recognition, vol. 28, n.10, pp. 1523-1535, 1995

[9] V. Wu, R. Manmatha, "Textfinder: an automatic system to detect and recognize text in images", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 21, n.11, pp. 1224-1229, 1999

[10] A.K. Jain, S. Bhattacharjee, "Text segmentation using Gabor filters for automatic document processing", Machine Vision and Applications, vol. 5, pp. 169-184, 1992

[11] A.C. Bovik, M. Clark, W. Geisler, "Multichannel texture analysis using localized spatial filters", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, pp. 55-73, 1990

[12] A.K. Jain, F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters", Pattern Recognition, vol. 24, pp. 1167-1186, 1991

[13] P. Sahoo, S. Soltani, and A. Wong, "A survey of thresholding techniques", *Computer Vision, Graphics, and Image Processing*, vol. 41, pp. 233–260, 1988.

[14] C. Thillou and B. Gosselin, "Robust thresholding based on wavelets and thinning algorithms for degraded camera images", Proceedings of ACIVS 2004, (2004)

[15] H. Yong Kim, "Segmentation-free printed character recognition by relaxed nearest neighbor learning of windowed operator", XII Simpósio Brasileiro de Computação Gráfica e Processamento de Imagens - Campinas, SP, Brasil , 1999

[16] Gonzalez, Woods, "Digital Image Processing", 1994

[17] C. Thillou, B. Gosselin, "Segmentation-based binarisation for color degraded images", Proceedings of ICCVG 2004, 2004

[18] Multitel website, http://www.multitel.be/TTS/