HAPTIC ACCESS TO CONVENTIONAL 2D MAPS FOR THE VISUALLY IMPAIRED

Konstantinos Kostopoulos, Konstantinos Moustakas, Dimitrios Tzovaras, Giorgos Nikolakis

Informatics and Telematics Institute,
Thermi-Thessaloniki, Greece
{kkostopo;moustak;tzovaras;gniko}@iti.gr

Céline Thillou, Bernard Gosselin

Faculte Polytechnique de Mons, Mons, Belgium
{celine.thillou;bernard.gosselin}@fpms.ac.be

ABSTRACT

This paper describes a framework of map image analysis and presentation of the semantic information to blind users using alternative modalities (i.e. haptics and audio). The resulting haptic-audio representation of the map is used by the blind for navigation and path planning purposes. The proposed framework utilizes novel algorithms for the segmentation of the map images using morphological filters that are able to provide indexed information on both the street network structure and the positions of the street names in the map. Next, off-the-shelf OCR and TTS algorithms are utilized to convert the visual information of the street names into audio messages. Finally, a grooved-line-map representation of the map network is generated and the blind users are able to investigate it using a haptic device. While navigating, audio messages are displayed providing information about the current position of the user (e.g. street name, cross-road notification and so on). Experimental results illustrate that the proposed system is considered very promising for the blind users and has been reported to be a very fast means of generating maps for the blind when compared to other traditional methods like Braille images.

KEYWORDS

Morphological filters – Connected operators – Haptic interaction – Multimodal maps – Blind users

1. INTRODUCTION

The human brain utilizes complex, still unknown procedures in order to perform intuitive task such as to decode the information stored in maps. These procedures are very difficult to imitate using computers. It is obvious that all common maps are perceived using the visual modality thus making maps inaccessible for special population categories like the visually impaired. Moreover, since maps are the major means of navigating into unknown spaces, it is more than clear that the visually impaired are not able to use the major means of navigation.

There have been many research efforts dedicated to the assistance of the navigation of the visually impaired. The University of Michigan provides the visually impaired students with a full tactile map of the University Campus [1]. This specific tactile map, based on the Braille code, provides not only navigational assistance but also useful information concerning campus functionality like the courses program.

For the navigation of visually impaired persons an oral tactile interface [2] has been proposed. The interface consists of a silicon-based mouthpiece with two modules. The first, a 7x7 tactor array, is adjusted to the roof of the mouth and takes over the tactile display. The second, a tongue touch keypad (TTK), is adjusted to the bottom of the mouth and receives instructions from the user. The latter sends feedback using the TTK and receives tactile cues at the roof of the mouth.

Moreover, in [3], a PDA, coupled with an embedded camera, able to recognize road signs or even text from natural figures is proposed. The PDA captures natural figures and detects text areas. Characters are segmented and a respective audio message is composed and displayed to the blind user, providing him/her with significant information about the surrounding space.

Scalable Vector Graphics (SVG) maps [4] have also been proposed. For the navigation of the visually impaired Scalable Vector Graphics maps contain sound effects as also description tags and are based on SVG, which is a modularized language for describing two-dimensional vector and mixed vector/raster graphics in XML.

Furthermore, in [5] an automated indoor navigation system dedicated to the visually impaired is proposed. The above system utilizes a camera mounted to the user to capture images from the surroundings. Then through image processing the obtained data are analysed by a computer which constructs the local map. Simultaneously the computer follows user's movement and notifies the latter about the presence of points of interest (POIs) like door handles or turning points.

Additionally, in [6] a system developed for the training of the visually impaired is proposed. This system enables visually impaired, to study and interact with various objects in specially designed virtual environments, while allowing designers to produce and customize these configurations. Cane simulation [6] is one of the scenarios that the above system has been used to.

The goal of the framework proposed in this paper is to provide the visually impaired with an easy to use means of accessing conventional 2D maps. The user can interact with the produced 3D model of the map and examine its properties. The developed framework analyzes the map image so as to obtain the enclosed information and is structured as follows. An overview of the framework is briefly presented in Section II. In Section III the map image analysis is thoroughly discussed. The implemented OCR algorithm is thoroughly analyzed in Section IV. Experimental results are demonstrated in Section V and in Section VI conclusions are drawn.

2. FRAMEWORK OVERVIEW

Figure 1 illustrates the general architecture of the developed platform. It uses as input a still image of a conventional 2D map. The output of the system is a haptic-audio representation of the 2D map. Four main modules can be identified in the system. In particular:

1) Recognition of street names and points of interest (POIs).

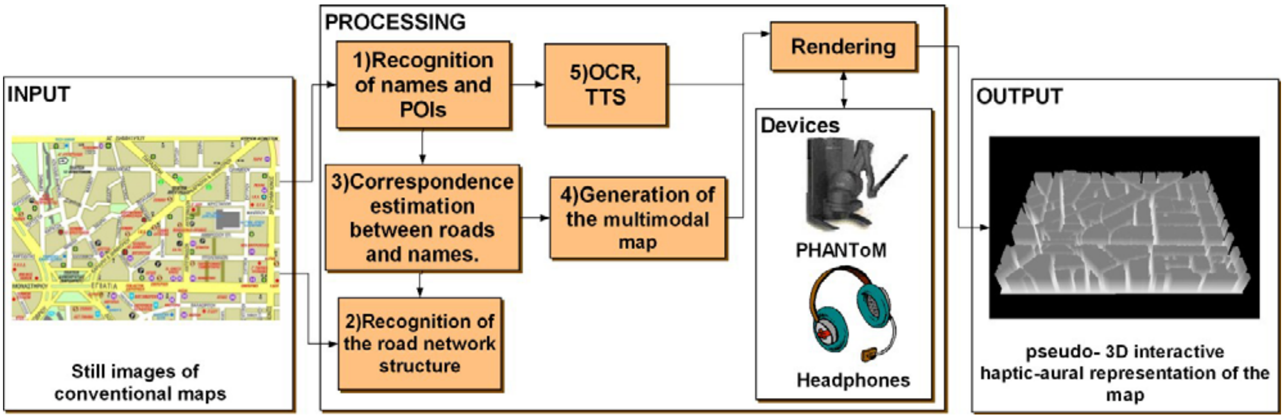

INPUT
Still images of conventional maps
PROCESSING
1)Recognition of names and POIs
5)OCR, TTS
3)Correspondence estimation between roads and names.
4)Generation of the multimodal map
2)Recognition of the road network structure
Rendering
Devices
PHANToM
Headphones
OUTPUT
pseudo- 3D interactive haptic-aural representation of the map

Figure 1: Architecture of the proposed system.

2) Recognition of the road network structure
3) Correspondence estimation between roads and names
4) Optical Character Recognition (OCR) and Text To Speech
5) Generation of the multimodal map.

A number of map prerequisites that the civil map should meet have been carefully chosen in order to design a system that will not depend on the map provider. In particular:

Color constraints: Street names should be represented using a dark color in order to be discernible from the rest of the map and thus advancing the process of their recognition.

Positioning: Street names should be located inside the associated road so as to attain their correspondence.

Resolution: Map resolution should be adequate to utilize an OCR algorithm to retrieve street names.

Special symbols: Symbols that represent points of interest such as hospitals, churches, parking and so on, should be thoroughly defined by the map provider.

3. MAP IMAGE ANALYSIS

The proposed map image analysis focuses on the extraction of the three major map components, street names, road network structure and points of interest (POIs).

3.1. Extracting the semantic information

For the detection of the street names, the primary map (Figure 2-a) is successively subjected to erosion (Figure 2-b) and dithering to two colors (Figure 2-c). Next, the produced image is segmented by applying region growing (Figure 2-d). The retrieved regions represent street names, points of interest and noise.

3.1.1. Extracting points of interest (POIs)

The recognition of points of interest is based on the Angular Partitioning of Abstract Image (APAI)[7],[8]. The latter estimates the resemblance between the templates of the POIs and the retrieved regions. Note that the above matching method is scale and rotation invariant. Figure 3 indicatively illustrates the templates of four symbol.

3.1.2. Extracting street names

After recognizing the POIs, the remaining regions represent street names and noise. The system discards all regions that are too small to represent a street name (Figure 2-d). Before applying the developed OCR system that is thoroughly described in the sequel, the extracted regions of the street names should be aligned in order to simplify the task of the OCR. In the context of the proposed framework, the principal axis of the street name area is identified and the images are rotated accordingly. Then the resulting images are up-sampled and a low pass filter is applied. Finally, the binary image is extracted by dithering to two colors.

Finally, the proposed OCR system is applied to recognize the text and then an off-the-shelf TTS [9] algorithm is utilized to convert the corresponding text into audio messages.

3.2. Estimating the Road Network Structure

For the estimation of the road network structure the system initially discards all street names from the primary map. Next connected operators [10], [11] and [12] are used to process the image as described in the sequel.

3.2.1. Connected Operators

Consider image I of Figure 4-a. In order to discard region B the general idea is to gradually reduce region B while retaining the rest regions of I. Therefore :

I) All regions of I are diminished (Figure 4-b) using the dilation operator $\delta_c(I)$.

II) Consider $\epsilon_c(I)$ as the erosion of I. To retain regions A and C but not region B the algorithm applies operator $max()$ on the images $\epsilon_c(I)$ and I.

Theoretically the second step is iteratively repeated. The intermediate images g_k of the k-th iteration step are recursively computed by equation:

$$g_k = max(\epsilon_c(g_{k-1}), I). \quad (1)$$

In the map case anti-extensive connected operators [11] are applied so as to enhance roads while diminishing street names.

Summarizing, the road sketch retrieval is modulated as follows (Figure 6) :

a) Consider the primary map M (Figure 5-a), $\delta_c(M)$ (Figure 5-b) as the dilation of M and the binary image of the street name's regions $T^{-1}[M]$(Figure 5-c) The image g_0 (Figure 5-d) is calculated according to equation:

$$g_0 = \delta_c(M) + T^{-1}[M]. \quad (2)$$

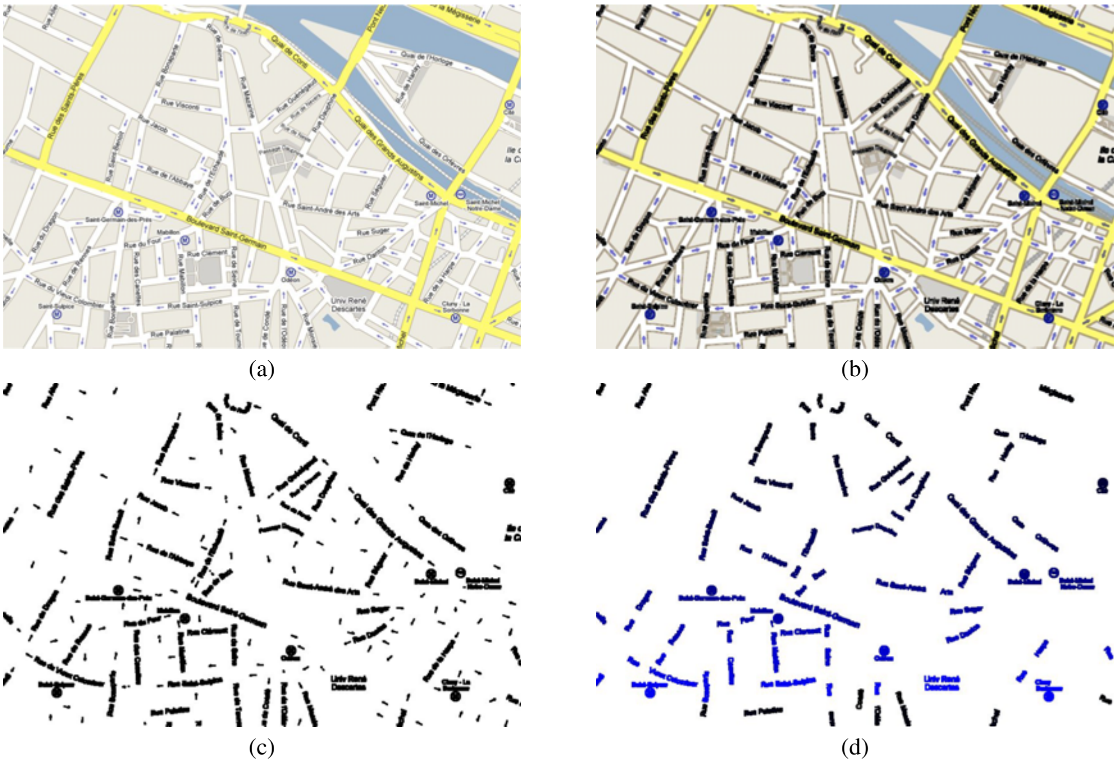
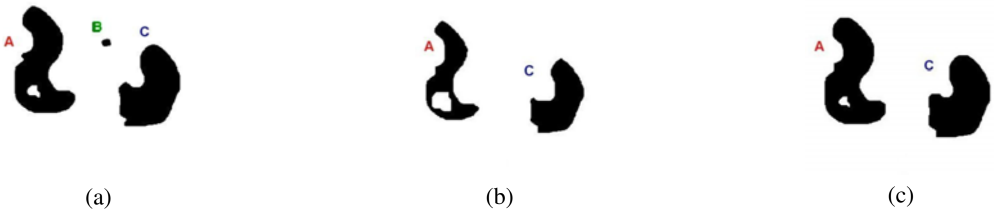

(a) (b) (c) (d)

Figure 2: *(a) Primary map image, (b) erosion, (c) dithering, (d) retrieved regions.*

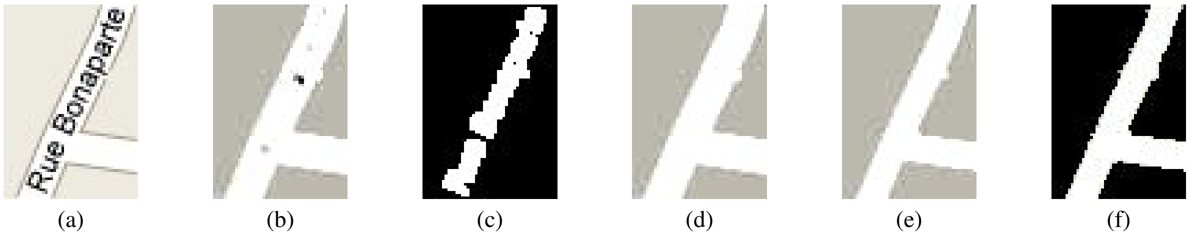
(a) (b) (c) (d)

Figure 3: *Symbol templates for : (a) parking, (b) hotel, (c) bank, and (d) church.*



(a) (b) (c)

Figure 4: *a) Primary image, b) region B is eliminated as desired. As a side effect, regions A and C have shrunk, c) desired areas are restored to their primary size.*



(a) (b) (c) (d) (e) (f)

Figure 5: *a) Primary image f, b) $\delta_c M$, c) $T^{-1}[M]$, d) g_0 term, e) street names have been removed while the road has been retained at his primary size and (f) the retrieved road network structure.*

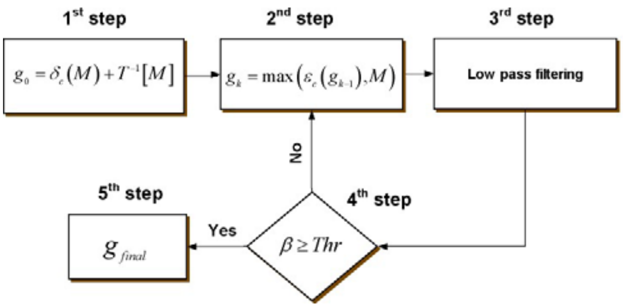


Figure 6: *Anti-extensive connected operators block diagram.*

b) If we represent as $\epsilon_c(M)$ the erosion of M, then each term g_k produced after the k-th iteration is recursively computed according to the equation:

$$g_k = max(\epsilon_c(g_{k-1}), M). \tag{3}$$

c) Next, the system examines the last two sequential terms g_{k-1} and g_k. If

$$\beta = \frac{N_V}{RC} = \begin{cases} \geq Thr, & \text{end} \\ < Thr, & \text{proceed in step b} \end{cases} \tag{4}$$

where, N_V is the amount of elements of set V,

$$V = \{(x, y) \in g_k \mid g_{k(x,y)} = g_{k-1(x,y)}\} \tag{5}$$

R, C are the image dimensions and Thr the relative threshold experimentally selected to be : $Thr = 0.98$.

Note that an additional step (c) is introduced in the algorithm so as to ensure convergence after finite number of iterations. Figure 5-e illustrates the result of the aforementioned iterative procedure.

3.2.2. Color-Based Clustering

After eliminating street names a color-based clustering algorithm classifies every pixel of the resulting image g_{final} to two sets, namely: 'S_R and S_B, where S_R represents map roads while S_B represents map buildings. Apparently $g_{final} = S_R \cup S_B$. Thus, we assume that for every type of road a mean color value is predefined. Let us assume that η random variables $\boldsymbol{z}_i$, where $i \in \{1, 2, \ldots, \eta\}$, represent road colors and follow 3D Gaussian distribution in the color space. More precisely :

$$f_i(\boldsymbol{r}) = \frac{1}{\sqrt{(2\pi)^3 |\boldsymbol{C}|}} e^{-\frac{1}{2}(\mathbf{r}-\mathbf{r}_i)^T \mathbf{C}(\mathbf{r}-\mathbf{r}_i)} \tag{6}$$

where $\boldsymbol{r}_i$ the mean value of the i-th road color, matrix $\boldsymbol{C}$ is the covariance matrix and is diagonal.

After evaluating all functions f_i, pixel p is classified in set S_R if and only if

$$f_i(\boldsymbol{r}_p) > t_i \tag{7}$$

for only one of the functions f_i , where t_i the i-th threshold. Otherwise p is classified in set S_B. Figure 5-f illustrates the retrieved road network structure.

Crossroads are simply detected as the areas that belong jointly to more streets. Each street segment that lies between two crossroads is then linked to the appropriate street name.

3.3. Generation of the Haptic Map

The final step is the construction of the Multimodal Map. The latter's model supports haptic interaction and provides navigational audio messages.

The haptic representation of the map is generated as a grooved line map (Figure 12-b), since it is reported in the literature that such a structure is better perceived using a haptic device, when compared to a raised line map [13]. The latter is used for interaction with the visually impaired using a haptic and an audio device.

Although exploration in egocentric reference frames is possible in our framework, exocentric reference frames were used since they seem to provide a better basis for the navigation and object recognition tasks that were performed in our experiments.

4. OPTICAL CHARACTER RECOGNITION OF STREET NAMES

Character segmentation and recognition have been performed for several decades, especially typewritten characters from scanner. Results are satisfying for this kind of characters and machine vision and character recognition may now be used for industrial purposes or assisting tools such as for visually impaired. A natural scene text reading system has been developed in [14] and the character recognition is coupled with an off-the-shelf text-to-speech algorithm to provide audio reading to the blind. Commercial OCR software perform well on "clean" documents with a minimum resolution of 300 dpi, which is barely obtained with tiny characters of maps, in this context, as shown in Figure 7. We chose to develop a generic OCR, in order to build a versatile system, where maps that come from different sources, are equally treated.

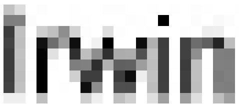


Figure 7: *Sample of tiny names of streets from maps.*

As described in section 3.1.2 the extracted images of the street names are properly oriented. However, several steps need to be performed: character segmentation as we use a character-based recognizer and character recognition.

4.1. Character segmentation

Names of streets, roads and so on are already binarized and well-oriented. First off all, a connected component-based analysis is performed with a 8-connectivity, as characters are quite thin. Hence, after binarization and rotation, more characters are broken or considered as broken with a 4-connectivity.

Two important problems in character segmentation still need to be addressed: broken characters and touching ones. To have a good segmentation, it is really important to consider these issues before the recognition step. Thanks to the average of characters width, computed after connected components analysis, all overlapping parts are grouped to be only one character, like for the letter "e" in Figure 8.

On the other hand, some touching characters may appear such as in Figure 9 and they may be cut with the Caliper distance. A Caliper histogram is formed plotting the distance between the uppermost and bottommost pixels in each column. A weak weight is applied for minima in strategic positions (which

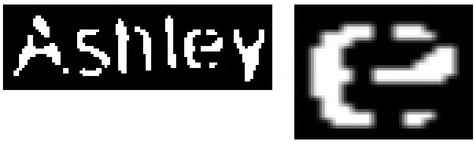


Figure 8: *Left: broken characters in a street name. Right: overlap of four parts of the character "e".*

is the middle for two assumed characters or one third and two thirds for three assumed characters) and a strong weight for the borders of characters. Hence with the touching character "ry", displayed in Figure 9, the right cut place is closer of the letter "y".

Characters with a ratio height/width inferior to 0.75 are considered to be more than one character and the Caliper distance is computed to find possible cut places.

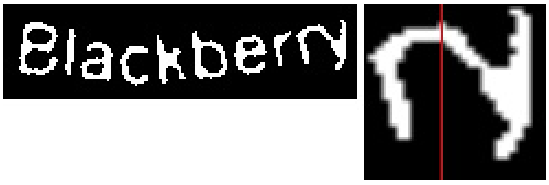


Figure 9: *Left: A word with touching characters. Right: Caliper algorithm: the chosen cut (in red) between "r" and "y" of the "Blackberry" word.*

4.2. Character recognition

For character recognition, we use a supervised classification using a multi-layer perceptron with features based on contours profiles [15].

In order to recognize many variations of the same character, features need to be robust against noise, distortions, style variation, translation, small rotation or shear. Invariants are features which have approximately the same value for samples of the same character, deformed or not. To be as invariant as possible, our input-characters are normalized into a $N \times N$ size with $N = 16$. However, not all variations among characters such as noise or degradations can be modelled by invariants, and the database used to train the neural network must have different variations of a same character.

The feature vector is based on the edges of characters and a probe is sent in each direction (horizontal, vertical and diagonal) and to get the information of holes like in the "B" character, some interior probes are sent from the center. Moreover, another feature is added: the ratio between original height and original width in order to very easily discriminate an "i" from an "m". Explanations of probes are given in Figure 10.

No feature selection is defined and the feature set is a vector of 57 values provided to an MLP with one hidden layer of 120 neurons and an output layer of size 36 for each Latin letter and digit. Due to few training samples for capital letters, uppercase and lowercase letters were grouped into the same class. The total number of training samples is 40614 divided into 80% for training only and 20% for cross-validation purpose in order to avoid overtraining.

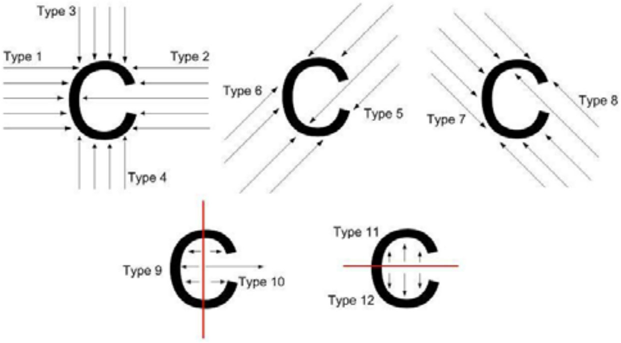


Figure 10: *The probes characteristics used to extract character contour.*

Due to orientation-free characters in maps, some names of streets may be upside down such as in Figure 11. Hence, based on confidence levels of OCR outputs, two character recognition are performed. The first one is done on the initial word and the second one on the transformed word with a 2Π-rotation. This recognition-based reorientation contains no errors, as well-oriented words have always the highest confidence level.

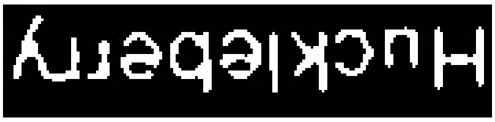


Figure 11: *Sample of a wrongly oriented street name.*

5. EXPERIMENTAL RESULTS

The proposed method was tested in generating Multimodal Map models of maps from various providers. Notice that all used samples fulfill all the prerequisites enumerated in Section 2.

The user explores the map in exocentric reference frames, where the viewpoint of the observer is extracted from the virtual environment. Although exploration in egocentric reference frames, where the viewpoint of the virtual environment corresponds to the viewpoint of the observer as if he/she was immersed within the scene is possible in our framework, the evaluation was performed with exocentric reference frames, because they provided a better basis for the navigation and object recognition tasks conducted in our experiments. In addition, our method transforms the pseudo-3D maps into grooved-line maps (Figure 12-b) that are more efficient for haptic interaction, when compared to the raised line-maps [13].

Figure 12-a shows a map image of the center of Thessaloniki acquired by [16]. The obtained Multimodal Map is illustrated in Figure 12b. Figure 13-a depicts a map image of Seattle acquired by [17]. Figure 13-b shows the extracted street names, while the detected crossroads are shown in Figure 13-c with red color. Finally, the produced Multimodal Map is illustrated in Figure 13-d.

As illustrated in Figure 13-c, the majority of crossroads are detected correctly. As also illustrated in Figures 12-b and 13-d the resulting pseudo-3D representation of the map is very clear and haptic rendering can be easily performed at interactive rates. Moreover, using the force shading method [18], the resulting force feedback is smooth and does not suffer from strong discontinuities. Moreover as Figures 13a and 13b illustrate the entirety of street names are identified despite their varying orientation and scale.

Figure [17] illustrates the obtained results for a map sample of New York City (Figure 14). Note that although the width of the streets varies, the obtained road network structure Figure 14-b can be efficiently obtained and the Multimodal Map is accurately generated (Figure 14-c).

Regarding the optical character recognition of street names, error rates are computed using the Levenshtein distance [19] between the ground truth and the resulting text. The Levenshtein distance or edit distance between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character. Equal weights for each operation are employed in our computation. Error rates are then computed by dividing with the number of characters. By using the Levenshtein distance, some error rates for a word may be superior to 1, but it is useful to penalize broken characters.

Tests have been computed on 87 words and 525 characters on two different partial town maps, Georgia and Michigan. The observed recognition rate was 81.3%.

Results may be still increased if a dedicated recognizer is used and more importantly, if a lexicon of streets, avenues and so on is exploited. It will lead to either good results or no answers when no matching may be performed between the recognized word and lexicon entries. Following that, additional steps may be then applied to turn the "no-answer" case into good results.

6. CONCLUSIONS

During the latest years cartography has been greatly advanced with the GPS being the tip of the iceberg. Unfortunately, up to date common maps are perceived using the visual modality. Therefore, special population categories like the visually impaired cannot access this information. In the present paper research is focused on transforming visual data to haptic representations.

A robust framework has been presented that generates a haptic-audio representation of the 2D map. The blind users are able to navigate in the generated pseudo-3D map using a haptic device, while audio feedback regarding the street names is also provided. Extensive tests have illustrated the efficiency of the approach.

7. ACKNOWLEDGMENTS

This work is supported by the SIMILAR IST Network of Excellence (FP6-507609).

8. REFERENCES

[1] "Disability INformation Resources (DINF)". http://www.dinf.ne.jp/doc/english/Us_Eu/conf/csun_98/. 13

[2] H. Tang and D. J. Beebe, "An Oral Tactile Interface for Blind Navigation", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, pp. 116–123, March 2006. 13

[3] V. Gaudissart, S. Ferreira, C. Thillou, and B. Gosselin, "SYPOLE: Mobile Reading Assistant for Blind People", in *Proceedings of European Signal Processing Conference (EUSIPCO 2005)*, (Antalya, Turkey), 2005. 13

[4] *Scalable Vector Graphics (SVG) 1.1 Specification W3C Recommendation*, January 2003. 13

[5] L. W. Ching and M. K. Leung, "SINVI : Smart Indoor Navigation for the Visually Impaired", in *8th International Conference on Control, Automation, Robotics and Vision Kunming, China*, December 6-9 2004. 13

[6] D. Tzovaras, G. Nikolakis, G. Fergadis, S. Malasiotis, and M. Stavrakis, "Design and Implementation of Haptic Virtual Enviroments for the Training of the Visually Impaired", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 12, pp. 266–278, June 2004. 13

[7] A. Chalechale, G. Naghdy, and A. Mertins, "Sketch-Based Image Matching Using Angular Partitioning", *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans*, vol. 35, pp. 28–41, January 2005. 14

[8] A. Chalechale, G. Naghdy, and A. Mertins, "Sketch-Based Image Retrieval Using Angular Partitioning", in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2003)*, pp. 668–671, December 2003. 14

[9] N. Grammalidis, N. Sarris, F. Deligianni, and M. G. Strintzis, "Three Dimensional Facial Adaptation for MPEG-4 Talking Heads", *EURASIP Journal on Applied Signal Processing, Special Issue on Signal Processing for 3D Imaging and Virtual Reality*, vol. 2002, pp. 1005–1020, October 2002. 14

[10] P. Salembier, L. Garrido, and A. Oliveras, "Region-based filtering of images and video sequences: a morphological viewpoint", in *UPC Barcelona SPAIN*, May 2001. 14

[11] P. Salembier, A. Oliveras, and L. Garrido, "Anti-extensive connected operators for image and sequence processing", *IEEE Transactions on Image Processing*, vol. 7, pp. 555–570, April 1998. 14

[12] P.Salembier and F.Marqués, "Region-based representations of image and video: Segmentation tools for multimedia services", *IEEE Transactions on circuits and systems for video technology*, vol. 9, pp. 1147–1169, December 1999. 14

[13] R. Ramloll, W. Yu, S. Brewster, B. Riedel, M. Burton, and G. Dimigen, "Constructing sonified haptic line graphs for the blind student: First steps.", in *ACM conference on Assistive technologies*, (Arlington, USA), 2000. 16, 17

[14] C. Thillou, S. Ferreira, and B. Gosselin, "An embedded application for degraded text recognition", *Eurasip Journal on Applied Signal Processing, Special Issue on Advances in Intelligent Vision Systems: methods and applications*, vol. 13, no. Number, pp. 2127–2135, 2005. 16

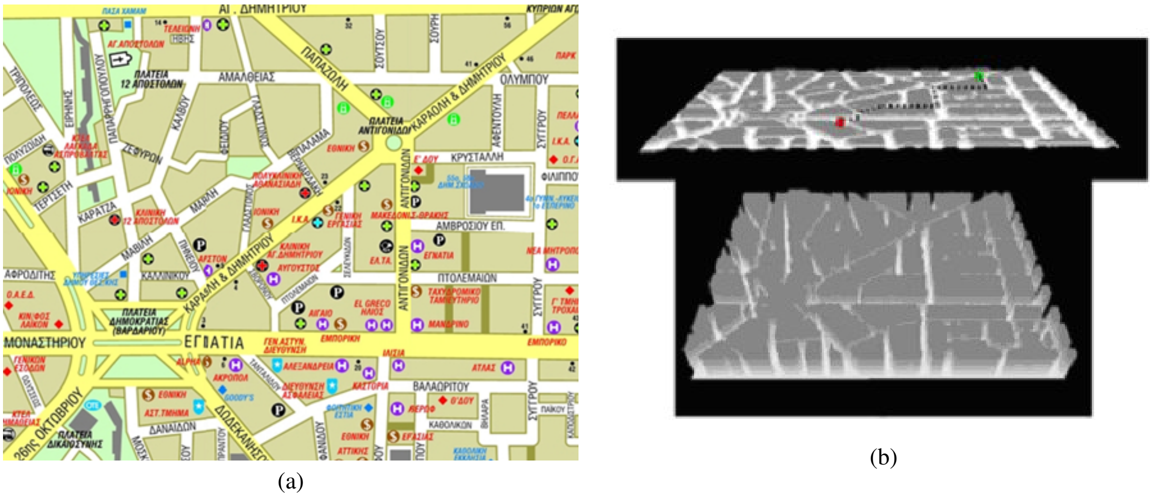
[15] B. Gosselin, *Application de réseaux de neurones artificiels à la reconnaissance automatique de caractères manuscrits*. PhD thesis, Faculté Polytechnique de Mons, 1996. 17

[16] "031.gr Desktop". http://www.031.gr. 17

[17] "Google Maps". http://maps.google.com/. 17, 18

[18] D. Ruspini, K. Kolarov, and O. Khatib, "The Haptic Display of Complex Graphical Environments", in *Computer Graphics (SIGGRAPH'97 Conf. Proc.)*, pp. 345–352, 1997. 17

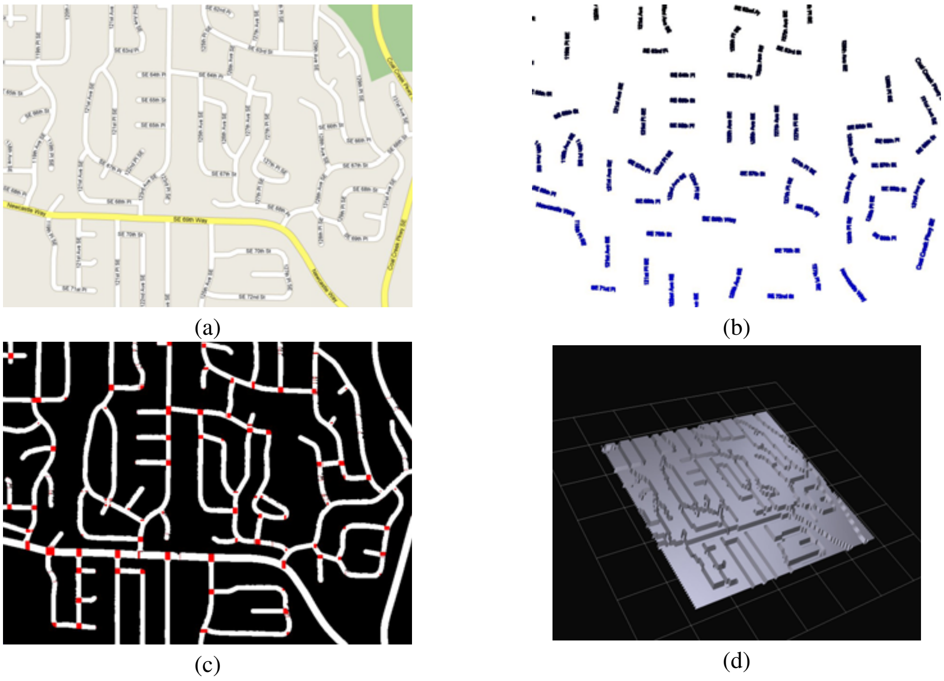
[19] V. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals", *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966. 18



(a)

(b)

Figure 12: *(a) Primary map image illustrating the center of Thessaloniki city Greece, (b) The obtained Multimodal Map model used for evaluation by visually impaired.*



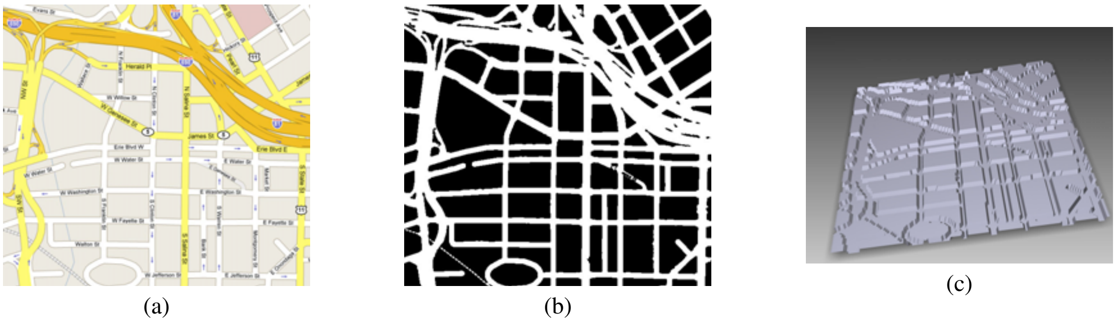
(a)

(b)

(c)

(d)

Figure 13: *(a) Primary map image, (b) extracted street names, (c)red spots indicate crossroads and (c) 3D map model.*



(a)

(b)

(c)

Figure 14: *(a) Primary map image, (b) obtained road network structure is street width invariant, (c) the obtained 3D map model.*