Bionic Eyeglass: The First Prototype

A Personal Navigation Device for Visually Impaired – A Review

Kristóf Karacs, Anna Lázár, Róbert Wagner, Bence Bálint, Tamás Roska
Faculty of Information Technology,
Pázmány Péter Catholic University
Budapest, Hungary

Mihály Szuhaj
'IT for the Visually Impaired' foundation,

Budapest, Hungary

***Abstract*—The first self-contained experimental prototype of a bionic eyeglass is presented in this paper, a device that helps blind and visually impaired people in basic tasks of their everyday life by converting visual information into speech. The indoor and outdoor situations and tasks have been selected by a technical committee consisting of blind and visually impaired persons, considering their most important needs and potential practical benefits that an audio guide can provide. The prototype system uses a cell phone as a frontend and embedded cellular visual computer as a computing device. Typical events have been collected in the Blind Mobile Navigation Database to validate the algorithms developed.**

Keywords-personal navigation, cellular computing, visual impairment

I. INTRODUCTION

In spite of the impressive advances related to retinal prostheses, there is no imminent promise to make them soon available with a realistic performance to help blind or visually impaired persons in everyday needs. We proposed a device that monitors the surrounding environment and retrieves information for the blind individual about it [1]. The device is based on the Cellular Neural/nonlinear Network – Universal Machine (CNN-UM) and the underlying Cellular Wave Computing principle [2]-[4]. This paper presents the first prototype of the Bionic Eyeglass.

The project has three main distinctive features respected to similar previous works ([5]-[9]):

1. Frequent communication with blind and visually impaired people. This helps us to identify the main situations that a blind person faces in everyday life and which tasks are to be solved in these situations.
2. Standardized clinical tests. Testing of the prototype system is carried out using standardized methods by doctors, taking into account the type of visual injury of the patients.
3. Neuromorphic solutions. The system is based on an intensive multi-channel retina-like preprocessing of the input flow with semantic embedding. The CNN-UM proved to be a suitable tool for modeling the processing in the retina, where each of the channels extract different spatio-temporal features of the flow. [3] The system also uses a neuromorphic attention model to focus on the information that is the most important for humans.

The next section outlines the system requirements and architecture of the prototype system. The third and fourth sections present color processing and some details of the partially neuromorphic saliency and event recognition system. Section five presents the video database we created for training and testing the algorithms.

II. SYSTEM ARCHITECTURE

The Bionic Eyeglass provides a wearable TeraOps visual computing power to advise visually impaired people in their daily life: at home, at work, and on the way between them. The basic tasks are summarized in Table I. Italicized tasks refer to those ones that we already developed algorithms for.

TABLE I TYPICAL TASKS CONSIDERED FOR THE BIONIC EYEGLASS

<table>
<tr><th></th><th>Home</th><th>Street</th><th>Office</th></tr>
<tr><td rowspan="6">User-initiated functions</td><td>Color and pattern recognition of clothes</td><td>Recognition of crosswalks</td><td>Recognition of control signs and displays in elevators</td></tr>
<tr><td>Bank note recognition</td><td>Escalator direction recognition</td><td>Navigation in public offices and restrooms</td></tr>
<tr><td rowspan="3"></td><td>Public transport sign recognition</td><td>Identification of restroom signs</td></tr>
<tr><td>Bus and tram stop identification</td><td>Recognition of signs on walkways</td></tr>
<tr><td colspan="2">Recognition of fluorescent displays</td></tr>
<tr><td colspan="3">Recognition of messages on LCD displays</td></tr>
<tr><td rowspan="2">Autonomous warnings</td><td>Light left on</td><td rowspan="2" colspan="2">Obstacles at head and chest level (i.e. boughs, signs, devices attached to the wall, trucks being loaded)</td></tr>
<tr><td>Gas oven left turned on</td></tr>
</table>

Two types of cellular wave computing algorithms are used: (i) standard templates and subroutines and (ii) bio-inspired neuromorphic spatial-temporal event detection. Examples for the former one are door handle detection, corridor sign extraction, banknote and letter recognition [10], [11]. The second type of algorithm is a neuromorphic saliency system [12] using the recently developed multi-

ONR N00014-02-1-0884
RET Szentágothai Tudásközpont

channel mammalian retinal model [13] followed by a classifier using the semantic embedding principle. The system architecture is shown in Fig. 1.

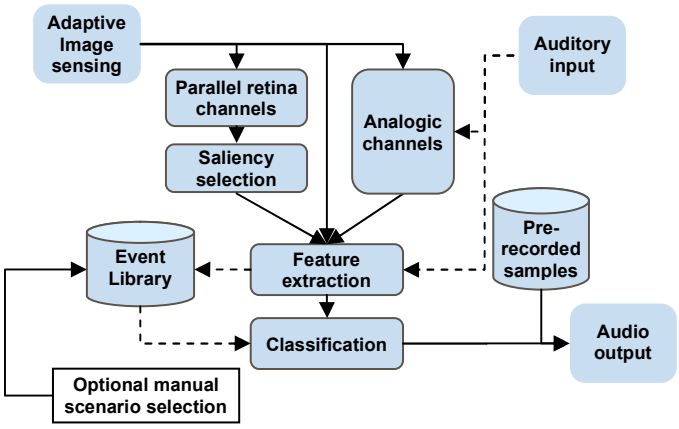

Figure 1. Overview of the system architecture

The prototype system has been built to show the feasibility of the system and to facilitate clinical and pilot tests. It consists of a cell phone with built-in camera and loudspeakers, the Bi-i visual computer [14] and a wireless adapter (Fig. 2.). Although the Bi-i has on-chip visual input, we do not make use of it in the present configuration since the form factor of the current Bi-i version makes it impractical.

Figure 2. The prototype system consists of a cell phone (bottom), the Bi-i visual computer (left) and the wireless adapter (right). Typical usage

We use a WiFi-enabled Nokia N95 cell phone in the prototype system. We implemented a communication framework for it that receives the image flow recorded by the cell-phone camera and streams it through the wireless connection to the Bi-i. The Bi-i visual computer runs the algorithms on its parallel architecture and returns the result via the wireless connection to the cell phone. The cell phone plays a pre-recorded audio sample based on the result received from the Bi-i.

The prototype system is portable. The Bi-i and the wireless adapter are powered from 12V and 5V batteries respectively. These devices of the prototype system weigh less than 3 lbs and can be carried in a backpack.

The hardware implementation platform will evolve from the present Bi-i and cell phone platform to a single, integrated eyeglass-mount unit using cellular wave computing, neuromorphic hardware. The first step in designing a wearable or handheld hardware is the integration of the cellular architecture and the cell phone.

III. COLOR PROCESSING

Visually impaired people need color information in some cases, such as to be able to choose clothes of matching colors. Thus we include a function that informs the user about the color and texture of the objects seen.

In the system we extract the colors of the scene and retrieve their location. The extraction of the colors can be interpreted as a color segmentation problem. Color segmentation algorithms have two main aspects. One is the type of representation of the color space, the other is the method used to group the pixels [15].

Color space representations are linear or nonlinear transformations of the RGB color space. For the computation of the perceived color we use the nonlinear CIE Luv color space. A great advantage of it is that distances of the stimuli are similar to the human perceived chromatic distance. Hence the transformation of the RGB channels corresponds to the human perception.

The methods used for the grouping of pixels can be classified as followings:

1) Pixel based classification. These algorithms try to find groups of pixels in the 3D color histogram. Common techniques for this are thresholding or clustering.

2) Region based methods. These consist of growing regions from an initial seed until color boundaries are reached.

3) Edge based methods. Similar to the region based methods, but at first the boundaries of color objects are specified and a region based method is applied afterwards.

We use the k-means clustering (pixel based method) with k=16. Steps of the algorithm are as follows:

i. RGB → Luv color space conversion
ii. Luminance adaptation
iii. Clustering
iv. Merging similar clusters
v. White correction
vi. Identification of color names for all regions

Steps on the CNN-UM are shown in bold, the other operations like clustering and color space conversion require less regional processing and they are performed on the accompanying digital processor.

The luminance adaptation reduces the distorting effects of the illumination by eliminating the differences in the spatial low-pass component of the luminance of the perceived image (see Fig. 2 a. b.). Since the "L" channel of the Luv space represents the luminance we performed our local luminance adaptation on it. The advantage of the CNN-UM architecture comes with local processing, which enables the easy computation of the local average.

A common problem of k-means clustering is oversegmentation, because we do not know the proper number of initial cluster centers. We overcame this problem with a postprocessing step, in which we merged similar clusters, whose centers are closer to each other then a given threshold. The result of clustering and merging can be seen on Fig. 3.

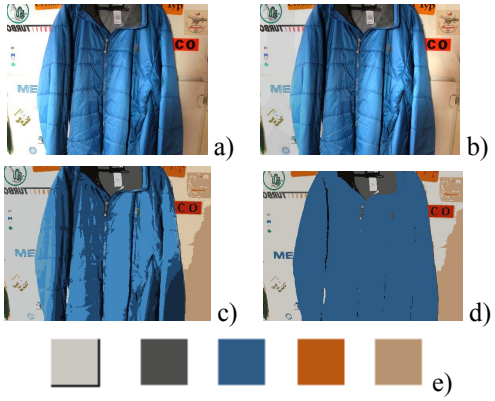

Figure 3. Clustering and merging of clusters. a.) shows the original picture taken by a standard digital camera. b.) shows the effect of the luminance adaptation. On c.) we can see the result of clustering. On this picture the pixels have the color of the cluster they assigned to. Regions of a given color represent a cluster. d.) shows the clusters after merging of similar clusters. The main cluster colors can be seen on e.)

The postprocessing step "White Correction" (see steps of the algorithm) exploits the fact that the image is clustered. If there is a cluster that lies close to saturated white color we multiply its channels to become white. Channels of the other clusters are also scaled with the same multiplication factors. This correction reduces chromatic distortion (Fig. 3).

The last stage determines the location of the clusters and gives a verbal classification of their location using speech samples (Fig 4).

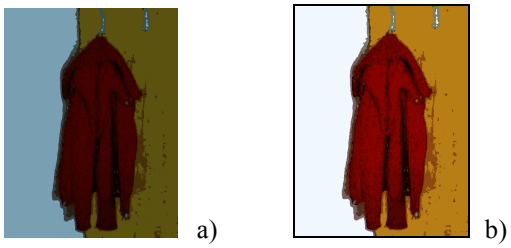

Figure 4. a) shows image taken by a mobile phone, which has extra sensitivity for blue colors. b) Shows the corrected clusters, we can see the white color of the wall.

IV. Some Details of the Neuromorphic Saliency and Event Recognition System

A. *Adaptive image sensing*

Adaptive image sensing is important if we deal with scenes that have large intra-scene dynamic range, like in real-world street image flows. Recent works on adaptive image sensing [16] using CNN-UM are developed using locally adaptable sensor array. A retina-like adaptation can be achieved by adjusting the integration time so, that the local average of an image region becomes the half of the maximum value. This eliminates the intra scene DC differences. In outdoor scenes where the variations of illumination might be large – both in time and in space – the adaptation is a useful property that enables the operation of the recognition steps. Single chip vision systems with adaptive image sensors are going to be available soon [17].

B. *Parallel image sensing - processing*

The first and best-known part of the visual system is the retina that is a sophisticated feature preprocessor with a continuous input and several parallel output channels [18]. These interacting channels represent the visual scene by extracting several features. These features are filtered and considered as components of a vector that is classified.

Beyond reflecting the biological motivations, our main goal was to create an efficient algorithmic framework for real-life experiments, thus the enhanced image flow is analyzed via temporal, spatial and spatio-temporal processing channels that are derived from retinal processing and semantic understanding of the task. The outputs of these sub-channels are then combined in a programmable configuration to form new channel responses.

An example for this is the detection and recognition of signs of public transport vehicles. This is a user activated function. The processing uses subsequent frames of the input video flow to detect and recognize the sign. Eventually determining the location of the sign is much more difficult than locating it. Different algorithms are needed for black and white signs, and color displays. [11], [19]

A black and white sign is defined as a rectangular shaped, almost white spot inside a big dark area (window) in the lower part of the image. This semantic definition is represented by algorithms modeling some channels. An example is shown on Fig. 5.

Figure 5. Sign localization on a tram

Detecting color displays is even more difficult. We use a method combining colors and morphology: two color filters are applied to each of the three color channels and binary morphology is used to select locations with shaped that are script like.

Binary images of the numbers often become vague due to low resolution input and the high level of noise present on them. To overcome this problem we make use of the a priori knowledge that signs normally do not change in time, which means we can superpose subsequent sign images to achieve better image quality. For this purpose we use a fading memory calculated as a weighted average of a new frame and the previous memory image.

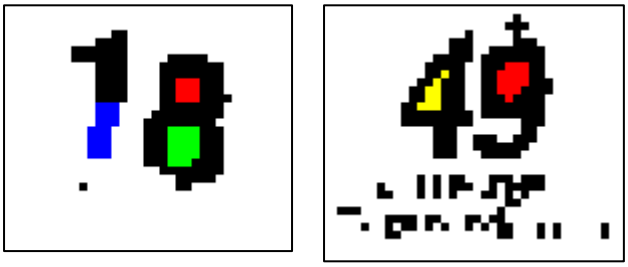

Figure 6. Feature maps of route numbers: vertical line (blue), upper hole (red), lower hole (green), triangular hole (yellow)

For number recognition we use topographic shape features that can be extracted by cellular wave algorithms. In the first step the number of figures of the number is determined by counting connected objects on the image that are bigger than a threshold. Features used include holes and lines, classified based on shape, size, position and orientation. Sample feature maps can be seen on Fig 6. Recognition is carried out via classification through a feature conversion table that assigns a number to each feature combination.

TABLE II FEATURE CONVERSION TABLE

Fig.	Hole			Line	
	Big	*Round*	*Triang.*	*Horiz.*	*Vert.*
0	+.				
1	LTO				M
2	LO		DRO	D	
3	LO				
4			U	M	
5		URO, DLO		U	
6		URO, D			
7				U	
8		U, D			
9		U, DLO			

+: Present D: Down U: Up M: Middle
L: Left R: Right O: Open T: Tight

C. *Saliency selection*

Visual attention is an ability to direct our gaze rapidly towards the objects of interest. This is a complex mechanism, which includes two different, but tightly coupled, parallel working systems. These are the bottom-up (or image-based) and the top down (or task-driven) methods [12]. Bottom-up originates at the retina and goes towards higher brain areas (involuntary), while top down originates in the high brain areas and projects towards the muscles of the eyes (voluntary). We know much more about the bottom-up method, which basically filters out the salient, conspicuous, sudden and unexpected parts of the visual scene.

In nature, saliency is "calculated" with receptive fields (RF), where neurons are organized into concentric circles: a central- and a peripheral part that respond antagonistically. Thereafter, in higher brain areas the structure of RFs gets more complex and their size increases too.

The flow diagram of the bottom-up process is shown in Fig. 7. In the first step the incoming vision is dissolved into several parallel retina channels, which are topographic maps of the visual scene. These channels code different low-level visual features, like motion, edges, colour antagonisms etc. In our model we use real retina channel emulations.

Once these channels are drawn up, each creates its own saliency map, which indicates that how salient, how 'loud' the different points are according to the appropriate low-level visual feature. These are also topographic maps, like the final (or master) saliency map, which is produced by aggregating the former ones and the most salient point wins the attention. The weights of the different retina channels are set according to the different tasks.

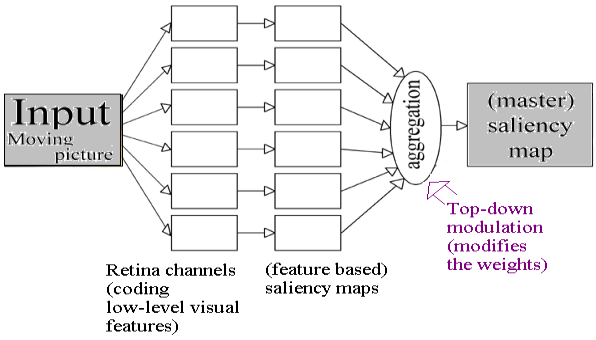

Figure 7. The flow diagram of the bottom-up attention mechanism.

D. *Autonome feature extraction and selection*

The retina-like spatial-temporal feature channels are further analyzed to extract low-level features. These binary maps describe the density of edges, irregularity, rough/fine structures, connected structures etc. of the input. The image around the most salient point is processed in detail. Local features are extracted, based on the assumption that the black patches are objects. These objects as entities are collected in a list and their features such as area or eccentricity are computed. Descriptive statistics is used to aggregate the same feature of the different objects such as min or mean.

The number of the features that can be extracted is enormous. We have to find those attributes that are informative enough for proper object categorization, whereas the number of them is still treatable. We have chosen the Sequential Floating Forward Selection (SFFS) algorithm

[20], and the Fisher-quotient as an accuracy function. We have picked this algorithm because in practical adaptations this proved to be the best.

E. Spatio-temporal event library

The Event Library contains descriptions of events in the expected scenarios; see Table I. Parallel scenarios are activated by salient features extracted from the scene. If a scenario is active it has an influence on the attention direction. The scenarios are weighted by a priori information and by the identified events, and the more weight a scenario is assigned the bigger the influence it will have on decisions and attention direction.

F. Multimodal classification with semantic embedding

The classification task can be greatly enhanced by using semantic embedding. This is the way formally and systematically evaluating the sensory context. These can be location based autonomous tasks, as listed in Table I, or restricted set of objects for example in recognizing the number on a public transport vehicle. (Fig. 6.) In addition to the visual input we plan to use auditory clues as well e.g. the noise of the arriving bus, the rustle of the escalator.

There are several classifiers that could have been used. We have applied an adaptive resonance theory (ART) based module, capable of learning on pre-selected training image flows [21]. The ART network has its roots in neurobiology. Its great advantage is that its modified version can be implemented on existing CNN-UM chips.

V. BLIND MOBILE NAVIGATION DATABASE

We established a database for training and testing the algorithms developed for the Bionic Eyeglass. The database is continuously thriving, more than 200 video flows of lengths between 10 and 90 seconds have been already recorded by a blind person in different situations mentioned in Table I. Indoor and outdoor recordings were taken under different light conditions.

We used commercial cellular phones and digital cameras to record videos. The resolution of the videos is either QCIF or QVGA. Presently no visual microprocessors are available with a resolution higher than QCIF, these recordings were taken for performance comparison purposes. Phones appropriate for this task must have a camera capable of video recording with at least QCIF resolution, and there must be a hard-button by which recording can be started and stopped (soft buttons on a touch-screen are too vague for a visually impaired user).

ACKNOWLEDGMENT

The support of the Hungarian Academy of Sciences, the P. Pázmány Catholic University, the Office of Naval Research as well as the Szentágothai Knowledge Center are kindly acknowledged.

REFERENCES

[1] T. Roska, D. Bálya, A. Lázár, K. Karacs and R. Wagner, "System aspects of a bionic eyeglass," in *Proc. of the 2006 IEEE International Symposium on Circuits and Systems (ISCAS 2006)*, Island of Kos, Greece, May 2006, pp. 161–164.

[2] L. O. Chua, T. Roska, Cellular Neural Networks and Visual Computing, Cambridge University Press, Cambridge, UK, 2002

[3] B. Roska and F. S. Werblin, "Vertical interactions across ten parallel, stacked representations in the mammalian retina," *Nature*, Vol. 410, pp. 583-587, 2001

[4] A Zarándy, Cs. Rekeczky, P. Földesy, I. Szatmári, "The new framework of applications – The Aladdin system," *J. Circuits Systems Computers* Vol. 12, pp. 769-782, 2003

[5] M. Mattar, A. Hanson, and E. Learned-Miller. "Sign Classification for the Visually Impaired", *Technical Report*, 2005-014 University of Massachusetts Amherst, 2005

[6] P. B. L. Meijer, "An Experimental System for Auditory Image Representations," *IEEE Transactions on Biomedical Engineering*, Vol. 39, No. 2, pp. 112-121, Feb 1992. Reprinted in the 1993 IMIA Yearbook of Medical Informatics, pp. 291-300.

[7] Amir Amedi, Felix Bermpohl, Joan Camprodon, Lotfi Merabet, Peter Meijer and Alvaro Pascual-Leone, "LO is a meta-modal operator for shape: an fMRI study using auditory-to-visual sensory substitution", 12th Annual Meeting of the Organization for Human Brain Mapping (HBM 2006) in Florence, Italy, June 11-15, 2006

[8] Chiu-Hung Cheng; Chung-Yu Wu; Bing Sheu; Li-Ju Lin; Kuan-Hsun Huang; Hsin-Chin Jiang; Wen-Cheng Yen; Chieao-Wei Hsiao: "In the blink of a silicon eye", Circuits and Devices Magazine, IEEE Vol. 17, No. 3, May 2001, pp. 20–32.

[9] A. Dollberg, H.G. Graf, B. Höfflinger, W. Nisch, J.D. Schulze Spuentrup, K. Schumacher: "A Fully Testable Retinal Implant", *Proceeding Biomedical Engineering*, 2003

[10] Á. Zarándy, F. Werblin, T. Roska and L. O. Chua, "Novel Types of Analogic CNN Algorithms for Recognizing Bank-notes," *Proceedings of IEEE Int. Workshop on Cellular Neural Networks and their Applications*, pp. 273–278, 1994

[11] K. Karacs and T. Roska, "Route number recognition via the Bionic Eyeglass," in *Proc. of 10th IEEE Int. Workshop on Cellular Neural Networks and their Applications*, Istanbul, Aug. 2006, pp. 79–84.

[12] L. Itti, Modeling Primate Visual Attention, in**:** *Computational Neuroscience: A Comprehensive Approach,* (J. Feng Ed.), pp. 635-655, Boca Raton: CRC Press, 2003

[13] D. Bálya, B. Roska, T. Roska, F. S. Werblin, "A CNN Framework for Modeling Parallel Processing in a Mammalian Retina," *Int'l Journal on Circuit Theory and Applications*, Vol. 30, pp. 363-393, 2002

[14] Á. Zarándy, Cs. Rekeczky: "Bi-i: a standalone ultra high speed cellular vision system", *IEEE Circuits and Systems Magazine*, 2005; 5(2):36 -45

[15] H.D. Cheng, X.H. Jiang, Y. Sun, Jingli Wang "Color image segmentation: advances and prospects," *Pattern Recognition* Vol. 34, (2001), pp 2259-2281

[16] R. Wagner, Á. Zarándy and T. Roska, "Adaptive Perception with Locally-Adaptable Sensor Array", *IEEE Transactions on Circuits and Systems I, : Regular Papers*, Vol. 51, No.5, pp. 1014-1023, 2004

[17] Anafocus Ltd, "AnaFocus Chooses Altera's Embedded Solutions for Single Chip Vision System" Press release, http://www.anafocus.com/home.php?opcion=3&subopcion=0

[18] J. E. Dowling, The Retina: An Approachable Part of the Brain, The Belknap Press of Harvard University Press, Cambridge, 1987

[19] K. Karacs, T. Roska, "Locating and Reading Color Displays with the Bionic Eyeglass," in *Proc. of the 18th European Conference on Circuit Theory and Design (ECCTD 2007)*, Seville, Spain, Aug. 2007, pp. 515–518.

[20] P. Pudil, F. J. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," *Proc. Inter. Conf. on Pattern Recognition*, 1994, vol. 1, pp. 279–283.

[21] G. Carpenter and S. Grossberg "A massively parallel architecture for a selforganizing neural pattern recognition machine," *Computer Vision, Graphics, and Image Processing*, Vol. 37, 1987, pp 54-115.