System aspects of a bionic eyeglass

Tamás Roska, Dávid Bálya, Anna Lázár, Kristóf Karacs, Robert Wagner
Faculty of Information Technology, Pázmány University and
Analogic and Neural Computing Laboratory, SZTAKI, Hungarian Academy of Sciences, and
Mihály Szuhaj
Hungarian National Association of Blind and Visually Impaired People,
Budapest, Hungary

Abstract—**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 navigating blind persons. In our new project, we are designing a Bionic Eyeglass that is providing a wearable TeraOps visual computing power to advise visually impaired people in their daily life. In this paper the system aspects are explained. There are three different types of situations (home, office, street) and a few standard image flows (with some auditory information). The basic tasks are indoor and outdoor events, defined by blind people. Two types of cellular wave computing algorithms are used: general purpose spatial-temporal event detection by analogic subroutines developed so far, and recently developed multi-channel mammalian retinal model followed by a classifier. Typical indoor and outdoor event detection processes are being considered.**

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 navigating blind or visually impaired persons in everyday needs. In our new project, we are designing a Bionic Eyeglass that is providing a wearable TeraOps visual computing power to give them support in their daily life. The presented system differs from existing topographic classification techniques in the intensive multi-channel retina-like preprocessing of the input flow, as well as the specific semantic embedding technique. The system is designed and implemented using the Cellular Wave Computing principle and the adaptive Cellular Nonlinear Network (CNN) Universal Machine architecture [1, 2, 3].

There is a strong biological motivation behind building a multi-channel adaptive algorithmic framework. It has been known since long that the mammalian visual system processes the world through a set of separate spatial-temporal channels and some outer retinal effects can be represented using the CNN Universal Machine [1]. However, the striking new result is that the organization of these channels begins already in the retina, where a vertical interaction across many parallel stack representations can be identified [4].

Our Bionic Eyeglass makes a major difference compared to any other devices made for visually impaired people since it is based on

- a cellular visual microprocessor family developed via the CNN Universal Machine principle with unprecedented computing power on a ~1 cm2 silicon chip with ~1W dissipated power,
- a dual visual input architecture (called the Bi-i [5]), and its software technology [6] and system implementation based on the above type of microprocessors ,
- a multi-channel mammalian retinal model [7] based on the recently discovered retinal operation and implemented real-time on the Bi-i., and
- the cellular wave computing algorithms combining topographic and non-topographic multimodal sensory flows [6].

A specific objective is to communicate the recognized objects and /or situations to the impaired persons by sound (speech). The research, design, and experimental implementation of the hardware and software tasks will be followed by the practical clinically supervised tests with the active participation of blind or visually severely impaired people as well as their ophthalmologists.

In this paper the system aspects of the Bionic Eyeglass are explained. The next section outlines the system requirements, design and architecture. The last section shows some details of the partially neuromorphic saliency and event recognition system.

II. System Design and Architecture

The Bionic Eyeglass provides a wearable TeraOps visual computing power to advise visually impaired people in their daily life. There are three different types of common situations: home, work, and on the way between them. A few standard image flows with some auditory information is used as benchmark. The basic tasks are indoor and outdoor events, defined by blind people.

ONR N00014-02-1-0884 , and
RET Szentágothai Tudásközpont 2005

TABLE I. TYPICAL SITUATIONS

<table>
<tr><td>Place</td><td>Home</td><td>Street</td><td>Office</td></tr>
<tr><td>Lightning</td><td>Controlled</td><td>Uncontrolled</td><td>Controlled</td></tr>
<tr><td>Events</td><td colspan="3">both emergency and conscious</td></tr>
<tr><td rowspan="6">requested</td><td>Color and pattern recognition of clothes</td><td>Recognition of marked and unmarked 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>Support in navigation in public offices and restrooms</td></tr>
<tr><td rowspan="4"></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 client displays (e.g: in banks)</td></tr>
<tr><td colspan="2">Recognition of messages on ATMs</td></tr>
<tr><td rowspan="2">autonomous warnings</td><td>light left switched on</td><td colspan="2" rowspan="2">obstacles at head and chest level (branches, signs, devices attached to the wall, etc.)</td></tr>
<tr><td>gas oven left turned on</td></tr>
</table>

Though we tried to restrict the task by selecting some typical places (home, street and office), the proposition is still very complex. The algorithmic development starts from the former analogic CNN algorithms for recognition of door handle and door sign [8], as well as object avoidance mechanisms [9], integrates the cellular wave computing algorithms for typical situations and blooms to a neuromorphic system with attention-selection and semantic embedding. The hardware implementation platform evolves from the present Bi-i self contained unit and mobile phone platform to a single, integrated eyeglass-mount unit using a SoC.

Two types of cellular wave computing algorithms will be used: (i) stand-alone templates and subroutines and (ii) bio-inspired neuromorphic spatial-temporal event detection. Examples for the former one are the door handle detection, corridor sign extraction, bank-note and letter extraction [10]. The second type of algorithm is a neuromorphic saliency system [11] using the recently developed multi-channel mammalian retinal model [7] followed by a classifier using the semantic embedding principle (e.g. [12]). The system architecture is shown in Figure 1.

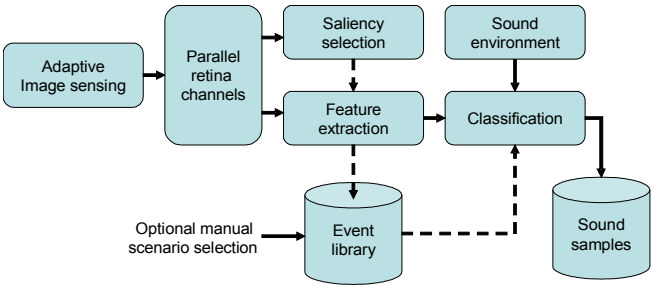

Figure 1. The partially neuromorphic system overview

III. COLOR PROCESSING

Despite the fact that visually impaired people do not perceive color the information about the color of an object is important in some cases e.g. color of clothes, Figure 2. Thus we include a function that informs the blinds about the color texture of the objects seen. For the computation of the perceived color we will use the CIE Luv color space, which has the property that the distances of stimuli is similar to the human perceived chromatic distance. The transformation between the color spaces occurs on an accompanying digital hardware, because this is a pixel-wise operation. The great advantage of the CNN-UM architecture comes with local processing. In the field of color processing such operation is needed at a chromatic preprocessing stage where an anisotropic diffusion is performed. This allows us the preservation of color boundaries but the elimination of smooth transitions along illumination changes [13].

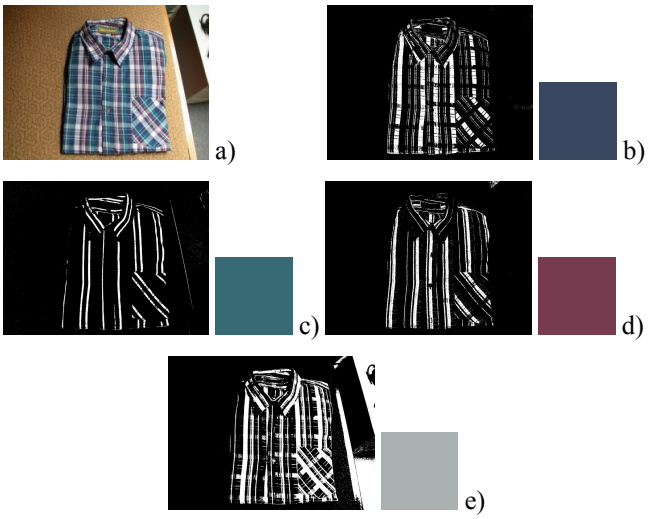

Figure 2. An example of color filtering: a) shows the original picture; b-e) show filtering of different colors. White areas show the location of the classified colors indicated in the small box beside the images.

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 [13] on adaptive image sensing 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.

B. Parallel image sensing- processing

The first and best-known part of the visual system is the retina, which is a sophisticated feature preprocessor with a continuous input and several parallel output channels [14]. 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. The outputs of these sub-channels are then combined in a programmable configuration to form new channel responses.

An example for this is recognition of signs of public transport vehicles. This is a controlled situation thus the user has to activate this function. The processing uses subsequent frames of the input video flow to recognize the sign. Algorithms modeling the channels first locate the sign on the scene, then extract features and classify the number (see [12]). The process is shown on Figure 3 and 4.

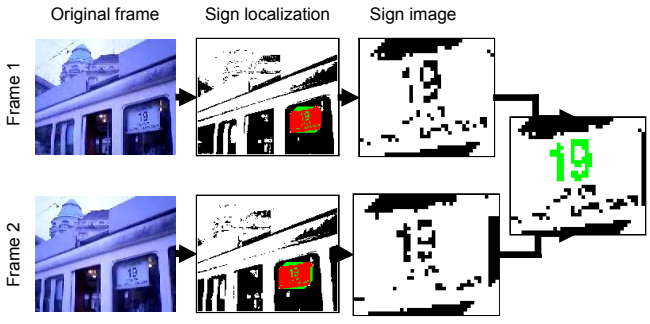

Figure 3. Localization of the sign and the number on a tram

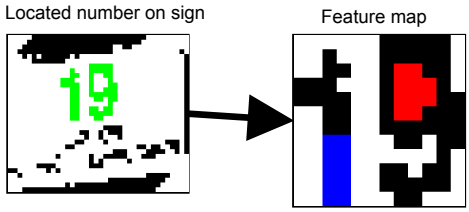

Figure 4. Localization of the sign and the number on a tram

C. Saliency selection

Visual attention is our ability to direct our gaze rapidly towards the objects of interest. This is a very complex mechanism, which includes two different, but tightly together, parallel working methods. These are the bottom-up (or image-based) and the top down (or task-driven) methods [11]. Bottom-up originates at the retina and goes towards higher brain areas. This is involuntary, fast (25-50 ms) and comes before getting aware of the scene. Top down originates in the high brain areas and projects towards the muscles of the eyes. This is voluntary, slower (200 ms) and task-dependent. Since the bottom-up method is bounded to lower brain areas in the sense of processing hierarchy, we know much more about this process. Bottom up method basically is for filtering out the salient, conspicuous, sudden and unexpected parts of the visual scene, therefore it helps increasing the efficiency of the algorithm of a bionic glass.

The flow diagram of the bottom-up process is depicted in Figure 5. 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.

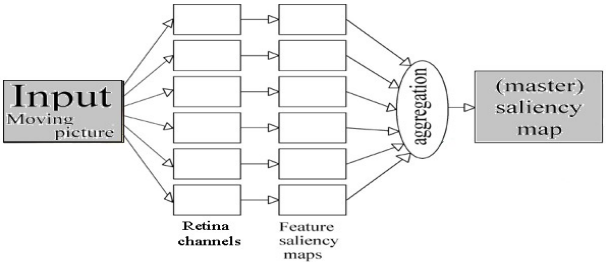

Figure 5. The flow diagram of the bottom-up attention mechanism.

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 not the same, they change according to the actual tasks.

One of the outdoor tasks that we would like to perform is to define the direction of the escalator. This is particularly important in those cases, when nobody or very few people is on the spot, so the blind person can not move with the crowd or can not ask. Figure 6 shows a potential solution for this task: looking for horizontal lines that can be filtered out from most of the retina channels. Even so we are using all of them –it would be unnecessary–: the transient (third picture in the first row), the local edge detector (beneath the transient) and the intensity channels are enough. If this the detected bar moves upwards, i.e. its vertical co-ordinate lessens, then the escalator shoves out, otherwise it draws near. If the bars are steady, then the device is out of order.

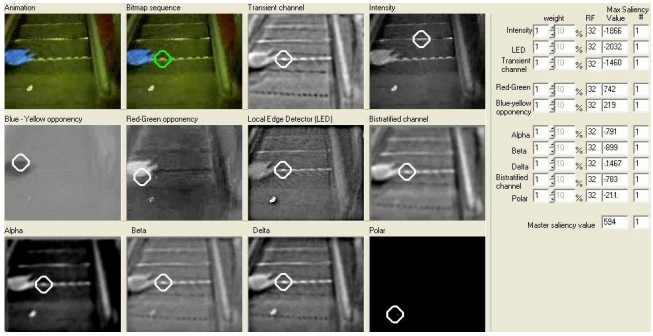

Figure 6. The most salient point in a moving staircase

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 categorisation, whereas the number of them is still treatable. We have chosen the Sequential Floating Forward Selection (SFFS) algorithm [15], which works as follows:

- **Input**: $\boldsymbol{Y}$, the whole feature set
- **Output**: $\boldsymbol{X}_m$, an m-sized feature-set
- **Termination**: when $\boldsymbol{k=m}$
- **Initialisation**: $\boldsymbol{X}_0=\emptyset$, $k=0$, where k is the number of the already selected features.

Step1: $x^+ := \underset{x \in (Y - Xk)}{\operatorname{argmax}} J(X_k + x) \quad X_{k+1} := X_k + x^+ \qquad k = k + 1$

Step2: $x^- := \underset{x \in Xk}{\operatorname{argmax}} J(X_k - x)$

if $J(X_k - x^-) > J(X_{k-1})$

then $X_{k-1} := X_k - x^- \quad k = k - 1$ and goto Step2

else goto Step1

Note that the k index in this algorithm denotes the number of the elements in the feature-set and not the step number. The function $\boldsymbol{J}$, which measures the accuracy of the selection, can be defined in several modes, for example it can be the Fisher-quotient. 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 recognising the number on a public transport vehicle, Figure 4. In addition to the visual input we plan to use auditory clues as well e.g. the noise of the arriving bus or tram, 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 [16]. The ART network has its inspiring roots in neurobiological modeling and has a mathematical background. A further advantage is that a modified version of ART can be implemented on existing CNN-UM architecture.

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

[2] T. Roska, "Computational and Computer Complexity of Analogic Cellular Wave Computers", *Journal of Circuits, Systems, and Computers*, Vol. 12, No. 4, pp. 539-562, 2003

[3] F. S. Werblin, T. Roska and L. O. Chua, "The analogic cellular neural network as a bionic eye," *Intl. J. of Circuit Theory and Applications*; Vol. 23, pp. 541-569, 1995

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

[5] 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.

[6] Cs. Rekeczky, I. Szatmári, D. Bálya, G. Tímár, and Á. Zarándy, “Cellular Multiadaptive Analogic Architecture: a Computational Framework for UAV Applications,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 51, pp.864-884, 2004

[7] 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

[8] M. Csapodi and T. Roska, “Dynamic analogic CNN algorithms for a complex recognition task - a first step towards a bionic eyeglass,” *Int. Journal ofCTA*, Vol. .24, No.1, pp.127-144, 1996

[9] V. Gál, T. Roska, “Collision Prediction via the CNN Universal Machine Int.", *Workshop on Cellular Neural Networks and Their Applications* (CNNA 2000), Catania, Italy, pp 105-110.

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

[12] K. Karacs and T. Roska, “Holistic Feature Extraction from Handwritten Words on Wave Computers”, *Proc. IEEE Int’l Workshop on Cellular Neural Networks and their Applications* (CNNA 2004), pp. 364-369, Budapest 2004

[13] 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

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

[15] 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.

[16] 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.