Human tested saliency map generation in the Bionic Eyeglass Project

Anna Lázár, Tamás Roska

***Abstract*—In this paper we introduce a bio-inspired approach that we apply in the "Bionic Eyeglass Project" which meant to help the everyday life of blind people. The basic idea is to mimic the nervous system, how it filters out the currently relevant information from the irrelevant mass – namely realize an attention model.**

The framework is a complete bottom-up based attention model where the parameters are adjusted via human tests. The principles are firstly the recently discovered ten different mammalian retina channels [3, 4], secondly the saliency map generation and finally how saliency depends on the receptive field types.

In the first part of the paper we introduce the theoretical background while the second part contains the empirical results.

I. Introduction

Emerging processing capacity and more and more sophisticated image processing algorithms gives the opportunity to prepare applications that can help blind or almost sightless people in their every-day life. In the Bionic Eyeglass Project we work together with the "Hungarian National Association for the Blind and Visually Impaired" thus these people can define what is important to them. Accordingly they appointed three main fields: home, workplace and the route between these two locations. Each one of these contains several sub-tasks, for example on the street the system should detect the direction of a moving staircase or the incidental obstacles in head-level (e.g. a loafing limb) [11].

In this paper we examine how a general bottom-up attention mechanism can be adopted in such a task.

II. Attention as Information Filtration

A. The two basic methods

Selective visual attention consists of two different but closely parallel-working methods: top-down (TD) and bottom-up (BU) [5]. They got their name after the direction of projection in the brain hierarchy. Accordingly, BU originates in the low areas - namely in the retina -, and projects towards the high areas, such as the ventro- and dorso-lateral prefrontal cortex. Correspondingly the origin of the TD is mostly bounded to the fronto-parietal cortex and drifts towards the eye. For a comparison of the main characteristics see table 1.

TABLE I
THE TWO ATTENTIONAL MECHANISM: THE BOTTOM-UP AND THE TOP-DOWN

Bottom-up	Top-down
Image based	Task dependent
Originated at the low levels of the brain hierarchy (retina) and goes towards the high areas (prefrontal cortex)	Originates at the high levels of the brain hierarchy (prefrontal cortex) and goes towards the low areas (retina)
Unconscious	Intentional
Takes 25~50 ms	Takes ~200 ms
It comes before getting aware of the scenery	Visual features can be adjusted voluntary according to the given task.
e.g.: evoked by a flickering red point in front of a grey background	e. g.: has a dominant effect when someone is searching for a key in a crowded drawer

Although in this project we have specific propositions to solve, we use bottom-up method. This is because top-down mechanism "sits" on the bottom-up method by using the same system, 'just' modifying the variant parameters. Thus if the tasks are predetermined, these parameters can be adjusted fittingly. In other word, task-dependency hangs on the weighting of the individual saliency maps, which are channel- and receptive field dependent. (See section III)

B. The Bottom-up algorithm in our model

"Bottom-up" is often called "image-based", indicating that this mechanism is based on saliency values that the different points in the visual scene reach. Most of the models that work out BU mechanism use more or less the same principles [6]. First, that a point's garishness is composed of several conspicuous-values – each of these belongs to different low level visual features. Second, that a location's saliency-value basically depends on the surrounding context. This means that a point's conspicuous-value is not equal with its garishness in an 'absolute value', but it is proportional with the contrast that it composes with it’s near surrounding. Third, the final saliency map aggregates the conspicuous-values that belong to the different low-level visual features with different weights – this weighting vitally depends on the top-down modulation (see Fig. 1). Fourth, scene understanding and object recognition tightly interplays in gaze-direction [5, 10] (nevertheless these are not parts of the BU method, in living nervous systems they play very important role).

To sum up, the main steps are usually the followings

- Dissolve the incoming picture according to low level visual features: colours, intensity (on, off), motion, junctions, etc. Usually the certain models employ only a few of these. In our model we use real retina channels instead of the heuristic ones. In the frame model we have ten channels, but the certain practical applications usually do not need all of them, typically one or two is enough.
- Create the saliency maps to each (used) channel. There are several strategies, the relevant precept is to measure the contrast between a point and it’s surrounding. We handle this task with receptive fields (RF) – as the neuromorph approach. Saliency maps and RFs play such a fundamental role in our model that we deal with them in a separate section.
- Feature combination. Unify the feature-based saliency maps into one final one, which is thus already feature independent. The weighting of the different channels are usually not equal, this is generally under some kind of top-down (task-dependent) modulation.

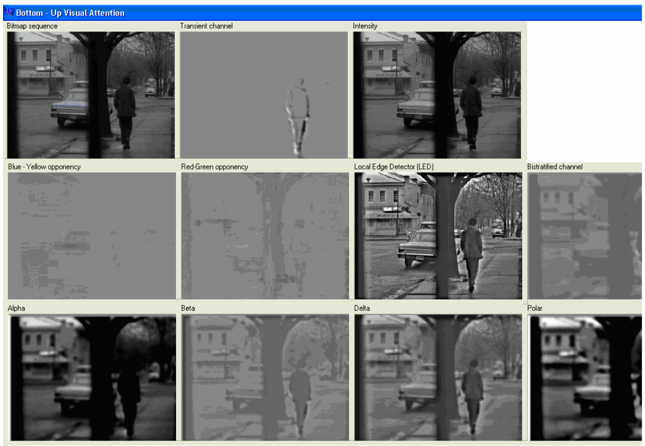

Fig. 2. The function of the retina. The input image (first picture) is processed by ten different ganglion-cell types which form the ten retina channels. [3]
The second picture in the first row (next to the input image) is the output of the “transient” channel which filters out the mobile parts of the visual scene and removes all the steady sections: at this moment it is only the girl that is moving. Normally this is one of the strongest channels.
The last image in the first row depicts the output of the “intensity” channel. In the second row we can see the blue-yellow- and the red-green contrast channels (these are the color channels), the LED (local edge detector) and the “bistratified” channels.
The functions of the channels depicted in the third row (alpha-, beta-, delta- and polar) are unknown for the present.

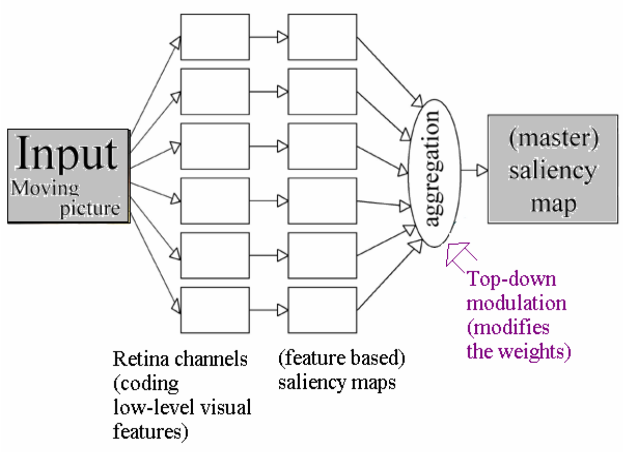

Fig. 1. The flow chart of the BU mechanism. In the first step the incoming picture (or frame in a video flow) is decomposed according to the low level visual features. - these are the different retina channels. The second step is to create the saliency maps – one for each channel. These maps are feature-based. In the next step they are aggregated into one final or “master” saliency map, which is thus already feature-independent, thus codes a general saliency value for each point of the input.

- Determine the most salient point or points (find the location that has the highest saliency value). In the theoretical model this is a winner-tale-all mechanism, which means that the whole process is for locating a single point, which will be the attended location. In practical applications we usually work with more locations and then make some processes on them – distance measuring, target tracking, etc.
- On still images the last (theoretical) step is creating a mechanism called “inhibition of return” which is for preventing attention to rut into a point. This forbids the attended locations for a while, thus attention can move to the next most salient point, then to the third one, etc. This process can differ in several items in the certain models.

Figure 1depicts the main steps of the bottom-up method.

C. *Low level visual features and retina channels*

Attention models usually work with heuristic low level visual features such as colors or color oppositions (red-green, blue-yellow), orientations, junctions, intensity, etc. [6] Our approach is to use real retina channels. [1, 2, 4] The framework of the retina channel model we used is based on biological measurements that were performed on rabbits [3]. (Rabbit retina resemble to the human retina very much.) Neuromorph processing differs from the heuristic assimilation as much as in half of the channels we can not even phrase the individual channel-function. (polar, bistratified, alpha, beta, delta). We can draw up the functions for the “LED” channel (which is an acronym for Local Edge Detector), the two color opposition channels, the intensity channel (brightness) and finally the transient channel which filters out everything that moves and eliminates everything that is motionless [9]. These channels should be treated as topographic maps: if two points are adjacent in the visual scene then the neurons coding their features will be adjacent in the given brain area as well. Topographic organization is a very common set-up in developed nervous systems. (For more details on retina please refer to [7, 8] and on channels and feature extraction to [2, 10].) The response of the individual channels can be seen on fig. 2. The original (input) image is an everyday scene, where a girl is walking on the street. No response (of the neuron in the topographic map that is located on the corresponding spot) is indicated by mean grey in the picture, brighter colors show “on” responses, darker colors show “off” responses. The brightness/darkness is proportional with the neurons reply.

As we mentioned before, in practical applications we usually need only one or two channels – which ones exactly, wildly depends on the task.

III. Saliency and Attention

Since in this system we attend to salient locations, the calculation of these values are a key step in this model. A point’s saliency represents how jarring it is compared to its surroundings. In nature, saliency is “calculated” with receptive fields, where neurons are organized into concentric circles: a central- and a peripheral part which response antagonistically. (See fig. 3) If the central part of an ON-center - OFF surrounding RF is stimulated with light, then it will increase the RF’s response, while if the light falls onto the surrounding part, then the reply will decrease. If both part is exposed, then there will be no change in the ganglion cells response. [10].

Receptive field organization characterizes the vision system from very low level to high levels: the higher we get in the hierarchy the more complex the RFs are. From one hand their size increases and from the other hand they compose different shapes, like horizontal or vertical lines. The ideal stimuli for a receptive field arrangement like this will be the horizontal (or vertical, etc.) line. In other words the salient points with such receptive fields will be the horizontal lines.

This means, that if we know that (in a given task) what kind of stimulus are we interested in, then we can design appropriate RF for it. For example, if we want to define the direction of a moving staircase, then we should attend to (moving) horizontal lines.

IV. Practical Applications and Observations

During the design of an application that lies on the above principles, we basically have to return two key verdicts: first, which retina channels do we want to use, and second: what shaped and sized receptive fields do we want to apply on the chosen channels.

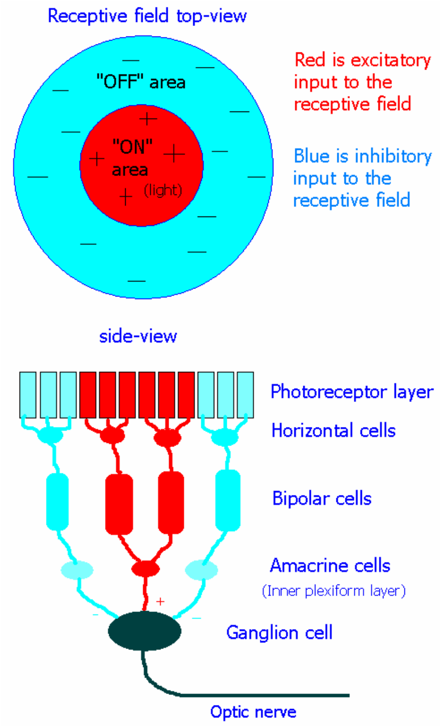

Fig. 3 The schematic organization of a ganglion cells receptive field (RF). This arrangement gives optimal response when the center and the surround is stimulated antagonistically. ON center - OFF surround cells are excited when the center is stimulated by light and inhibited when the surrounding part is stimulated with light. The size and the shape of the FR dramatically influences how salient a given location is valued. In the retina there are circle-shaped RFs, while in the higher brain areas these fields conglomerate into more complex structures (like oriented bars). In practical applications we can select the RF type that fits the best to the given task.

Different channels filter different low level visual features. In half of the cases we can not phrase the function (nature seems to judge different peculiarities important that are suggested by the engineering approach), but we can see and test whether the desired features emphasized in that channel or even neglected.

At the same time, nothing obstructs us to simultaneously utilize information gained from different channels (- probably the nervous system works the same way.) For example, on outdoor video-flows, it is often very useful to distinguish the sky and the foreground. For instance when we are searching for moving bars (e.g. looking for trams or busses) current collectors or transmission lines can disturb quite badly. Such problems can be by-passed by identifying those parts of the frame that shows the sky. (See fig. 4). Practical observations show that if the saliency values on the alpha- and polar channels are in a certain range, that region belongs to the sky with a quite high chance. (Of course these ranges vary by channels.) The red dashed region on the right most picture depicts the section that was identified as sky.

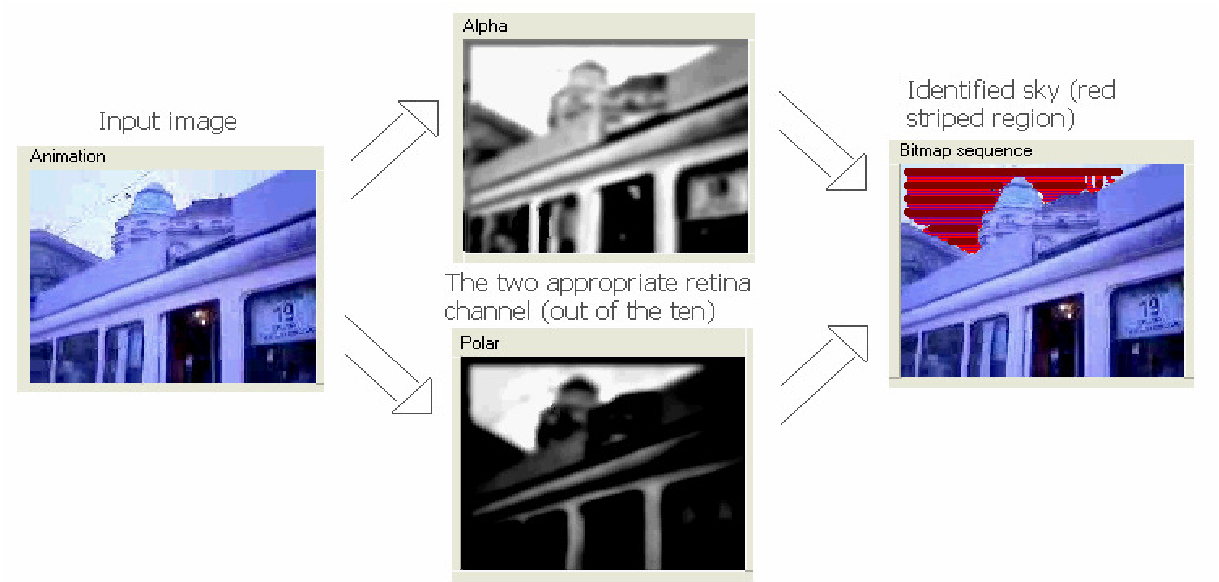
Input image
Animation
Alpha
The two appropriate retina channel (out of the ten)
Polar
Identified sky (red striped region)
Bitmap sequence

Fig. 4. Identifying the sky. More complex processing can be done by using data from different channels *simultaneously*. In this case we want to identify the background (namely the sky). If the values are between an appropriate range in the Alpha- and in the Polar channels a satisfying filtering can be retrieved. The ranges and the channels are task dependent and usually determined by experimental data.

The idea of creating more complex receptive fields arises artlessly. For example one can suspect that number-shaped receptive fields could find numbers independently from the tram or bus recognition. Experiences show, that the efficiency of the usage of RFs above a certain complexity, decays fast. This is an outgrowth of many facts, like too complex receptive fields can not solve the problem of size – the same object (or number on a tram) looks smaller from a bigger distance. However figure 5 also shows, that with luck this approach can also work - but for practical use this conception turned out to be too doubtful.

Consequently remains the simple shaped (dots, lines) receptive field types which differ only in size and orientations (-probably it is not a coincidence that RFs in the visual system neither exceed that certain complexity).

At the same time combining the information gained from these simple RFs can bear important data: e.g. in the situation depicted in figure 4 and 5 (tram), vertical bars that are moving in the same direction with the same speed, belong to the same vehicle with a good chance.

Additionally, motionless camera could increase the processing quality dramatically (- for example the walls of the surrounding buildings wouldn't appear to be mobile) and one channel could help in quite much tasks. This channel is the “transient” channel which filters out everything that is moving (Figure 2, second picture in the top row). In many cases the fixation of the camera seems to be resolvable at least for a short time.

Another challenge is to define the direction of a moving staircase. Here we have to implement this task when nobody (or only very few people) is around, otherwise the sightless person could move with the crowd or could ask. In this case this condition is an advantage, because the horizontal lines are not masked. The algorithm appeared to be quite robust on this task. Mostly all the channels can filter out the moving horizontal line, but from the viewpoint of the task the “bistratified” channel appeared to be the best choice. Thus taking other channels data into count would only increase the data to process and in the same time would also decrease the quality. Once we have the moving bars, the direction can be told by the alteration of the vertical coordinate: if it grows then the stair draws near, otherwise it shoves out. Finally, if the salient horizontal lines are motionless, then the escalator is out of order. (See Fig. 6)

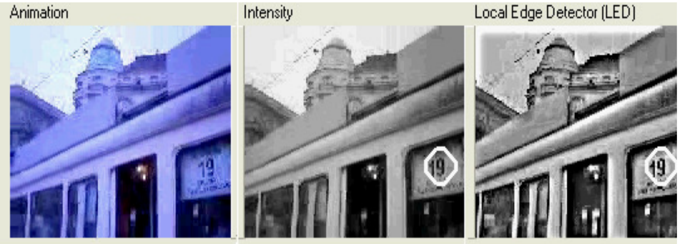

Fig. 5. Nevertheless number – shaped receptive fields can sometimes give good results (white circles on the Intensity- and the Local Edge Detector channels show the identified numbers), the usage of too complex FRs gives too unreliable data for practical applications. RFs above a certain complexity loose their generality, e.g. size-invariancy.

V. Acknowledgement

This work is a part of the “Bionic Eyeglass project” under the support of Szentágothai Knowledge Center and supervision of Tamás Roska.. We would also like to thank to Robert Wagner, Kristóf Karacs and Dávid Bálya, for advices in attention to Zoltán Vidnyánszky.

References

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

[2] 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

[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] D. Bályа, 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

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

[6] L. Itti and Christof Koch, “Computational modelling of visual attention,” *Nature Neuroscience,* Vol 2, 2001

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

[8] Richard H. Mashland “The fundamental plan of the retina”, *Nature neuroscience* Vol 4 No. 9, 2001

[9] A. K. Lazar, R. Wagner, D. Balya and T. Roska “Functional representations of retina channels via the RefineC retina simulator” Proc. of the 8th IEEE International Biannual Workshop on Cellular Neural Networks and their Applications, 2004

[10] E. R. Kandel and J. H. Schwartz, “Principles of Neuroscience” Elsevier, New York, 1985

[11] T. Roska, D. Bálya, A. K. Lázár, K. Karacs, R. Wagner „System aspects of a bionic eyeglass” Proc. of the IEEE International Symposium on Circuits and Systems, 2006

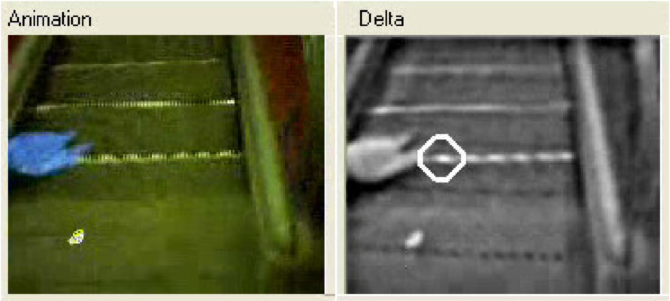

Fig. 6. One of the outdoor tasks is to define the direction of a moving staircase. The above picture shows a potential solution for this function: we are looking for horizontal lines (with horizontal bar-shaped FRs), which can be filtered out from almost all of the retina channels – even so one (or a few) channel is enough. In this case we choose the “delta” channel.
If this bar moves upwards (its y co-ordinate lessens) then the escalator shoves out, otherwise it draws near – or if the bars are steady, then the staircase is out of order.