MODELING VISUAL ATTENTION

Thesis of the Ph.D. dissertation

Anna Lazar

Supervisor:
Tamás Roska, D.Sc.
Ordinary Member of the Hungarian Academy of Science

Consultant:
Zoltán Vidnyánszky, D.Sc.
Doctor of the Hungarian Academy of Science

Péter Pázmány Catholic University
Faculty of Information Technology
Multidisciplinary Technical Sciences Doctoral School

Budapest, 2008.

1. Introduction

The attentional mechanism of a healthy person operates in such a natural manner and ease, that mainly we are not aware of how complex, in fact this mechanism is. With the help of our vision, we are informed about the objects and events in the surrounding world almost since our birth, and – in fortunate cases – it remains to be one of our most important sense-organs until the end of our life. Because of its' importance and *triviality* – meaning that, *as an experience* vision is a natural sense – many people try to understand it, and also to mimic it, to 'model' it, for a long while. Although, during that time, undeniably, enormous knowledge has been accumulated, we are still very far from the complete understanding of how the excitation of the photoreceptors in the retina by the photons transform into visual experience.

Usually, we *feel* that we continually sense, 'perceive' the outside world: we know what is under and above us, what is on the left and what is on the right; - in brief, we are constantly being informed of every small detail in our environment through our vision. This - although not completely baseless – after all, is essentially only a false sensation. According to different experiments, our environment can change even fundamentally (e.g. the walls turn from blue into red, or an object right in front of us disappears), but if this alteration is slow and smooth enough, we are unable to sense it. The explanation of this lies exactly in the fact, that we have only a restored 'picture' or 'representation' about the surrounding world.

However, this has more to do with memory, perhaps with stored knowledge, than with vision itself. And of course, this has its' good reason: if we *really*, permanently processed all the information being present in our visual environment, our brain would be extremely overloaded *needlessly*. For example, the shape and colour of the bookshelf aside, with all the accurate titles of every book on it, every small detail of the picture that can be seen through the half-open door from the other room, every tile on the neighboring house's roof, all these, simultaneously, surely are not important for us. Moreover, if we think a bit deeply on it, usually only a tiny little piece of our environment is important for us at a given moment, whereas the rest is redundant (for instance, the recurring motif of the wallpaper), or simply irrelevant (for instance the tiles on the next house's roof). Therefore the "stored representation method" is expressly beneficial. And, if anything varies compared to our stored data (for example, the lamp is being switched on in the adjacent room, the door is being opened, etc.), we will be informed through the appropriate, so called "bottom-up" attentional method.

At this point where we reach the concept of visual attention: this extremely important ability, which permits the (artificial or living) creature of finding the actually important data-fragment in the visual environment, without processing all the information, in real time.

It is hard to over-emphasize the evolutionary importance of this ability: a suddenly appearing predator, finding the mellow fruits in the bush, or maybe a warning of a tribe-mate.

Thus, from an engineering viewpoint, a system that is able to *attend*, is capable of finding that information-fragment in *real time*, which is *actually important* for the given system, right *there*, at that *given moment*. With this, enormous processing capacity can be saved, whereby the processing *quality* can be *increased,* whereas the *time* necessary for it can be *decreased* – since the redundant and/or unnecessary information-mass is not getting processed. Of course, the phrase "actually important" makes the problem extremely complex and difficult.

Our faculty is principally engaged in *neuromorphic modeling*, that is, the bases of our models are primarily living creatures. Human visual attention evolves from two different, but closely related, parallel working methods: an involuntary, reflex-like, so called "bottom-up", which, for example, directs our attention to a flickering red point in front of a grey background, and a voluntary one, so called "top-down", which, for example enables us to find our key in a packed drawer.

In the course of my Ph.D. studies, my main goal has been to understand and to model the bottom-up visual attentional method, as precisely as possible. During these years, a neuromorphic model has been created, in which the unknown parameters had been adjusted after human gaze direction measurements. Similarly, the accuracy of the model has been tested via comparing the models' predictions with human gaze direction measurements.

2. Experimental methods

My research area requires the joint application of different disciplines. Accordingly, as a first step, via neurobiological studies I have got acquainted with the basics of vision and with the mechanisms that form visual attention as well.

The substance of *modeling* lies in the proper selection of the elements forming a complex system (like an animals' visual system), more precisely, *the selection of those elements* which develop the features being important for us. Thus, if these elements are the same in different systems, then trespassing is possible among these systems. Accordingly, in a general sense, the vertebrate visual system can be considered as the basics of my model. (Present-day attentional models are by far not precise enough to detect the differences, for example between humans and primates.)

The main steps of the model are depicted on figure 1. As a first step, the input image (left hand-side, top) is being decomposed according to ten different retina channels (right hand-side, top). Next, each channel creates its' own saliency map, which is a two dimensional topographic map of the physical world in the brain (right hand-side, bottom). The weighted sum of these maps form the "final" or "master" saliency map (left hand side, bottom), which is a topographic map of the visual scene as well. The saliency map codes how striking, how obtrusive are the corresponding points in the physical world. The most intense point of this map attracts our attention the most, thus the corresponding location of the physical world is being mapped into the center of the sharp seeing, that is, to the fovea.

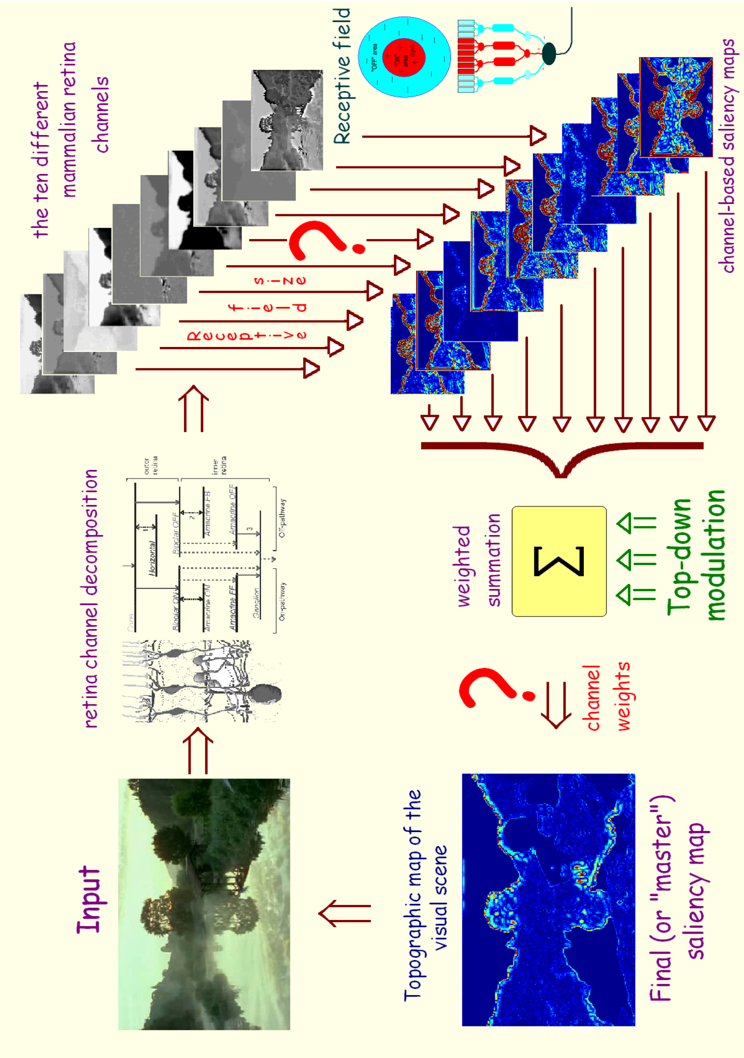
Input
retina channel decomposition
the ten different mammalian retina channels
Receptive field
Receptive field size
?
channel-based saliency maps
weighted summation
Σ
Top-down modulation
?
channel weights
Topographic map of the visual scene
Final (or "master") saliency map

Figure 1.: The functioning of the bottom-up attentional mechanism. As a first step, the input image splits up according to the different retina channels (ten different ganglion cell types), hereby forming the ten topographic maps of the input scene (right hand-side, top of the picture). Afterwards, each of these maps creates its' own saliency map (right hand-side, bottom) which will sum up into a "final" or "master" map (left hand-side, bottom). The most intense point of this final map attracts the attention; all the other locations are suppressed.

I have realized the above model in (Borland) C++.
The first main step has been the investigation and the completion of the retina channels. The model runs on a CNN (Cellular Neural/Nonlinear Network) simulator, which I have also prepared in Borland C++. I have got the proper parameters, which define the exact spatio-temporal behaviour of the different retina channels, from a previous work carried out by David Balya. Further developments of the model – of course, under the guidance of my supervisor and consultant – constitute my own work.

The principles underlying the retina-model are briefly the following: every retinal cell-layer (photo-receptors, horizontal, bipolar, amacrine and ganglion cell-layers) corresponds to a CNN-layer (figure 1, top, middle). The properties of the different cell-layers (average diameter of the dendritic tree, temporal properties of the cell responds, etc.) can be approximated with appropriate CNN templates and parameters. The connections *between* these CNN layers (excitations, inhibitions, temporal delays, diffusion parameters, etc.) have also been defined in a way, so that they approximate the output of the corresponding retinal layers, as close as possible. The *temporal* properties of the retina channels have been entrapped with the adaptation of a weighted, circular memory buffer: the newly processed frame overwrites always the oldest, and the overall output of the given channel is the weighted, pixel-wise summation of the buffer content (figure 1, top of the image, right hand side).

The next step is the creation of the saliency maps belonging to the individual retina channels (figure 1, right hand side). These maps are being formed by differently sized receptive fields (RF). In other words, every channel has a different "optimal" receptive field size, or else, a different receptive field distribution (figure 3, table 1). In the beginning of the cerebral vision-processing (that is, in the "low" brain areas), the RFs are relatively small, and also circle-shaped. The higher we get in the brain hierarchy, the biggest the RFs are, concerning their size, and the more complex they become, with respect to their shape. Since in the beginning of my studies, the RF-sizes belonging to the different channels were practically unknown, in the initial state of the model, these have been adjustable values via the keyboard. (On figure 1, the parameters for which no literature data has been existent up to now are highlighted with red question marks.) The other important, yet unknown parameters which determine the final saliency map are the *weights* of the channel-based saliency maps (figure 1, bottom, middle).

I have approximated these parameters via human gaze direction measurements. For this purpose, I have applied an equipment called "*iView X Hi-Speed System*" suitable for gaze direction measurements. The "training-set", that is, the video clip that the subjects have seen for the process of *estimating* the parameters, was a ~33 second flow, consisting of 267 frames, 8fps, where, each frame had a 512x298 pixel/frame resolution, 96 dpi. The stimulus did not contain any voice. It consisted of four shorter natural scenes, containing birds, mountains, lakes, horses, etc. The reason behind the usage of a *moving natural* input was justified by the fact that, according to literature-data, if the subject had no specific task to perform (e.g. into which continent the subject puts the scene, or, how many red and blue parrots the subject counts, etc.), then these conditions primarily trigger bottom-up visual attention.

During the measurements I have investigated the efficiency of 40 different receptive field sizes, for all the channels. This means, RF sizes spreading from 0.5° up to approximately 26°, expressed in terms of the viewing angle.

For the purpose of defining the *channel weights* I have applied different approaches addressing the following question: considering a given stimuli (frame), which channel(s) participate in triggering the saccade, and also, in what *extent* do these determine the new fixation position. (We call "saccades" those little eye-movements, "jumps", for which the center of the focus changes, that is, when one changes the fixation location.♣) During the measurements I have applied 240 Hz sampling frequency and I have only taken into account the saccades bigger then 1°.

For evaluating the measured data according to the above assumptions (differing regarding which channels are being considered as saccade-triggering ones with respect to given stimuli) I have developed programs under MatLab.

For the purpose of *validating* the received parameters, I have performed similar human measurements on a "test video set" with an

♣ In the literature we can find the word "saccade" in the sense of the shifting of the entire visual scene, but in the thesis I use this word in the sense given in the text.

analogous topic (that is: moving natural scenes) using the same equipment. The other settings had been the same, but for the sake of accuracy, I have used a longer-duration stimulus including 9 scenes, 477 frames, ~ 56 seconds.

During the validation process, I have measured the correspondence between the models' predictions and human gaze directions. For all the frames in the test video set, I have determined more points, as possible fixation locations (like: "on this frame, the coordinates of the most probable fixation location is the *x-y* pair, the coordinates of the second most probable position is *x'-y'* ", etc.). The results have shown a quite accurate correspondence: in ~70% of the cases, the *measured* location was among the first four *predicted* locations. The accidental chance of this is less then 20%.

During the generation of a visual attentional model, the goal is to reproduce the accomplishment of living creatures, namely, the capability of finding the actually important visual information in the redundant and/or irrelevant torrent, in real time. This is possible by using heuristic methods and ideas as well, but the final goal is – primarily in the case of *neuromorphic* modeling – to understand and mimic the neural structure of the creature that we have used as model, as proper as possible. The importance of this lies on the fact that during the development of such a model, we can learn a lot about the functioning of living systems. Moreover, problems of an engineering design create a correlation loop with biological measurements as well. Furthermore, regarding efficiency, heuristic systems are hardly up to the operational level of the corresponding mechanisms in living creatures.

3. New scientific results

Thesis #1: A new efficient method in the development of the bottom-up attentional model. The employment of the multi-channel mammalian retina model, which is based on the latest biological findings, instead of using the heuristic, low level visual feature filtering, and its' consequences.

In living creatures, the information processing starts already in the retina. Even more, the information leaves the retina in a highly filtered and organized way, and projects towards the higher brain areas for further processing. The first precise enough neurobiological descriptions of this information-classification – and thus also the retina-models built on them – have been appeared only in the last few years. Accordingly, this retinal process has been neglected in the earlier models, and instead of it, heuristic, different low level visual feature extraction algorithms have been applied.

The main novelties of the model I have implemented are the following: Firstly, the application of the methodology of the above mentioned multi-channel decomposition of the visual information, and thus the exploitation of the latest results of the retina research. Secondly, the estimation of the corresponding receptive field sizes in order to form the proper channel-based saliency maps by them.

I.1. I have improved the 'classical' visual attention model in a way that instead of using the generally applied 3-5 low level visual feature extraction (characterizing the 'classical' model), I am using the multi-channel mammalian retina decomposition method, which is based on the most recent neurobiological discoveries.

The first step in a neuromorphic visual attention model is the decomposition of the input image/video, according to the, so called, "low level visual features" (figure 1). Present-day models characteristically employ 3-5 of them, such as, edge-filtering, corner-filtering, color-filtering, etc.

Instead, in my model I have used the recently revealed and modeled mammalian retina network model, which differentiates ten channels (figure 2). In the case of five channels, the functions can be read of the output (edge-detection, motion-filtering, intensity and two color oppositions)*, while the function of the remaining five is unknown, in the sense that, the aim of their process could not be formulated explicitly, at least up to present. Consequently, none of the heuristic models can incorporate them.

* According to the latest researches, certain cells in the retina respond to motion direction-dependently, that is, in certain living creatures, another channel could exist, which filters motion in a direction selective manner. (Fried, S. I., Muench, T. A., & Werblin, F. S. (2002). Mechanisms and circuitry underlying direction selectivity, in the retina. Nature, 420, 411-414.)

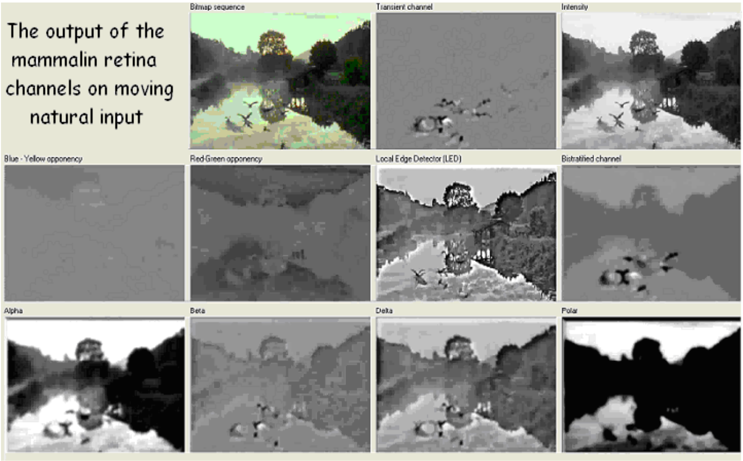

Figure 2.: The functioning of the mammalian retina channels on moving input. Uppermost row, left picture (input): birds flying over a lake. Next image on the right: the "Transient" channel, which in-filters everything that moves, and eliminates all the steady parts. Only the flying birds trigger answer on the given stimulus. Same row, right-most image: "Intensity" channel. Middle row from left to right: Blue-yellow contrast, Red-green contrast, LED (Local Edge Detector) and "Bistratified" channels. The function of the Bistratified channel, similarly to the channels depicted in the bottom-most row, had not been formulated explicitly before. The channels in the bottom-most row, from left to right are: Alpha, Beta, Delta and Polar.

Corollary: In my model, similarly to living systems, the saliency maps that are based on those retina channels having non-explicitly described functions, also take part in the allocation of the fixation location. I have investigated their role on moving visual input.

The so called "saliency maps" are two dimensional, topographic maps of the physical world in the brain, such that, the activity of certain neurons are proportional with the 'vividness', 'high-contrast' of the corresponding locations in the physical world.

Since the retina channels having non-explicitly described functions form saliency maps as well, and thus they take part in the formation of the final saliency map, neglecting them significantly modifies the final results. In my model, I have taken into account the saliency maps for *all* the retina channels, and I have determined the weights, the 'importance' of the saliency maps belonging to these channels by the same method, that I have used for the explicitly formulated ones.

Seven channels' response (Transient, LED, Bistratified, Alpha, Beta, Delta, Polar) out of the ten, depend not only on the actual stimulus, but also on its temporal behaviour. In other words, the response of these channels – and accordingly the saliency maps based on them – more or less react on *changes*, on *motion*. The effect of these saliency maps, during the formation of bottom up visual attention, for the first time has been investigated during my measurements.

Thesis #2: The estimation and optimization of the unknown parameters – namely, the receptive field sizes belonging to the different channels as well as the channel weights – based on human gaze-direction measurements. Additionally, the verification of the model, based also on human measurements.

The model includes two essential, but unknown parameters: firstly, *what sized receptive fields* form the saliency maps on the different retina channels, and secondly, what is the *weighting* with which the channel-based saliency maps form the final saliency map. (These are marked with red question marks on figure 1.) These parameters had been estimated via human gaze direction measurements, and I have checked the accuracy of the obtained model with similar measurements as well.

Directly, we can *not* measure the channel-based saliency maps (i.e. those belonging to a given retina channel) or their effects, but only the *fixation locations*, provided by the observers who have taken part in the experiments. We can only *infer*, deduce these immeasurable values by using different assumptions; that is, by using *indirect* methods. This is true for the *weighting* of the channels based maps as well. (It is quite difficult to design an experiment, a “stimulus”, which affects only one of the channels – it is enough to mention, that if the stimulus is for example *dynamic*, then it immediately affects the seven spatio-temporal channels and the Intensity one as well.)
Since, according to literature data, the gaze directions controlled by bottom-up mechanism are essentially determined by these saliency maps, I have estimated their *efficiency* through their most intensive points, namely, via the correspondence between the ‘keenest’ locations of these channel-based saliency maps and the measured fixation locations. Consequently, I have estimated the missing parameters via *inferences* – which are another reason why the *validation* (comparison with human gaze direction measurements) has been so important.

II.1. I have determined optimal receptive field (RF) sizes for all the ten retina channels in our model, via human measurements. These correspond to those receptive fields sizes that generate the corresponding saliency maps. This process involves the investigation of 40 different RF sizes, between ~0.5° and ~26°, expressed in terms of the viewing angle.

For the same input, receptive fields with different sizes result in different saliency maps. I consider a receptive field size as *optimal*, if the saliency map created by it is the most *effective*, that is, for which the most intense points of the corresponding saliency map give the most accurate concurrence with the *measured* fixation locations.

Different saccades are provoked by different channels. The open questions are the following: ***1)*** *how many channels* take part in the provocation of a given saccade, and ***2)*** *which are these* channels concretely. Addressing these questions, I have investigated two different assumptions:

1) The channels which trigger a saccade (determine the new fixation location), are those being the most "effective" according to *arbitrary* receptive field sizes.

2) The channels which trigger a saccade are those that are effective *in average*, that is, all the saliency maps according to all the 40 receptive field sizes participate in the averaging.

I have investigated the results if the first 1, 3 and 5 most effective channels take part in the generation of the final saliency map, according to both assumptions. During the evaluation of the different cases, I have obtained curves similar to those that can be seen on figure 3. This diagram shows the curves that belong to the most accurate estimation.

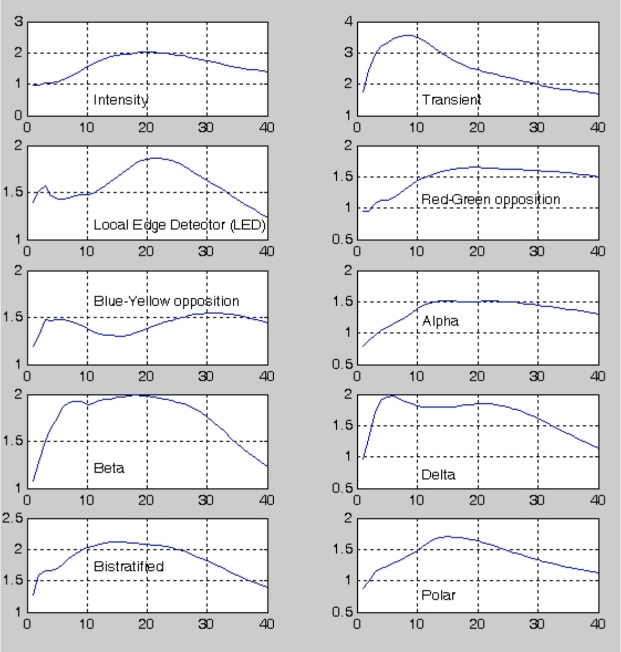

Figure 3: Each of the ten diagrams belongs to a retina channel. They show the average saliency values in the measured fixation locations, in the function of the 40 different receptive field sizes.
For all the input frames, and for all the ten channels, for every RF size, I have determined the average saliency values, and *assumed* that the channels which take part in triggering a saccade are those that achieve the highest *average* saliency value on the given stimulus (frame).

For the different channels, the '*optimal*' receptive field sizes are those, by which the corresponding curves reach their maximum (table I). The final, "tuned" model uses these RF sizes for creating the saliency maps.

In the 'real', *living* retina, the channels have an *interval* of RF sizes. In a biological viewpoint, the curves like figure 3 preferably show the *distribution* (density) of the different sized RFs, in the different retina channels. However, the explanation of the biological relevance of these curves was not the subject of my research. I emphasize that these investigations are based on a model level with aggregated functional tests and are not related to the neurobiological details.

Int	**Tr**	**LED**	**R-G**	**B-Y**	α	β	δ	**Bist**	**Pol**
20	9	21	20	31	22	19	4	15	15
12,9°	5,5°	13,6°	12,9°	20,2°	14,2°	12,2°	2,18°	9,6°	9,6°

Table I.: The optimal receptive field sizes belonging to the different retina channels. The first row indicates the abbreviations of the channels, which are sequentially the followings: 1)"Intensity", 2)"Transient", 3)LED (Local Edge Detector), 4)Red-green color-opposition, 5)Blue-yellow color-opposition, 6)"Alpha", 7)"Beta", 8)"Delta", 9)"Bistratified", 10)"Polar"

The middle row indicates the *indices* of the optimal RFs (see figure 3), while the bottom-most row shows the same, but in *viewing angle*. The connection between the indices and the viewing angle is: $\mathrm{tg}\frac{\alpha}{2}=\frac{4\mathrm{i}-3}{100}*0{,}147$, where *i* indicates the index, and α is the viewing angle.

II.2 I have investigated different hypotheses addressing the question: what is the *proportion* (“weight”) by which the different channels are responsible for provoking the saccades, that is, for determining the new fixation locations. Based on these, I have obtained different channel weightings.

I have analyzed assumptions, in which the channel weights had been kept constant, that is, the channel based-saliency maps had contributed in the formation of the final saliency map with always the same proportion. And also, I have investigated strategies, in which the channel weights had been constantly updated, according to the actual input.

- The assumptions for the fix channel-weighting strategies – which strongly build onto the previous point –, have been the following:

  The *channel-weights* are proportional to the relative *ratio* (percentage), by which they prove to be *saccade-triggering*:

  i. by *arbitrary* receptive field sizes
     (that is, how often do the highest saliency value(s) belong to the different channels – according to *any* RF)
     More concretely, a channel’s weight is proportional to the frequency that the channel-based saliency map contained one of the highest values.

  ii. by *average saliency value*
     (that is, how often do the different channels prove to be the most salient one *in average* – using all the RF sizes)

More concretely, a channel's weight is proportional to the frequency that the channel-based saliency map was one of the highest in average.

The results are depicted on figure 4, whereas the accuracy of the different hypotheses can be seen on figure 5.
The nominations *i)* and *ii)* refers to the previous points.

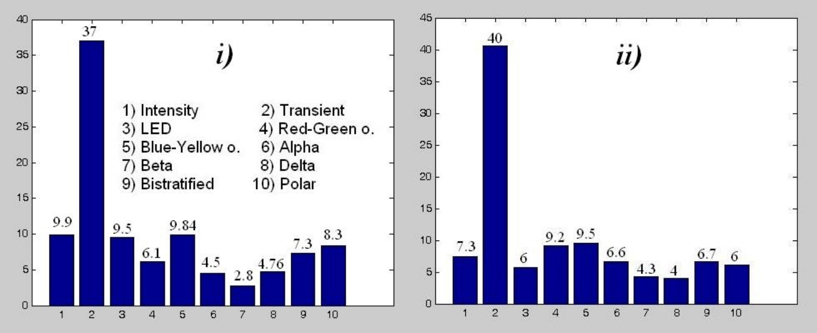

Figure 4: The estimated channel weights according to the two different hypotheses. The basics of the estimation: which channel how often (in what percentage) proved to be saccade-triggering according to the two assumptions. The exact values are indicated on the top of each bar. The accuracy is depicted on figure 5.

- The hypothesis for determining the channel weights in a continually updated manner is based on the assumption that the involvement of the different channels depend on the actual stimulus. In other words, the actual channel weights depend on the input, instead of being pre-defined.

The two basic assumptions are the same than previously: those channel(s) are responsible for triggering a saccade on the actual stimulus, which:

i. contains outstandingly high saliency values belonging to *any* RF size

ii. are the most salient *in average* on the given frame

The weighting is proportional to these maximal/average values.

Contrary to the expectations – although the differences were small – the *fix* channel weighting strategies proved to be better than the *continually updated* ones, in the sense that they gave more accurate predictions, compared to human gaze direction measurements. The former strategies have performed better by ~5% than the latter ones.

Validation. I have verified the model's accuracy via human gaze direction measurements, and I have shown that the model predicts the human fixation locations with high conformity on complex natural scenes.

With the model adjusted according to the results of the described measurements, I have made predictions of the expected fixation locations, and then I have compared them with measured human gaze directions. The *measured* locations were among the four most probable *predicted* locations in ~70% of the cases, on the given frames (- the accurate value varies slightly according to the different hypothesis.) The accidental chance for this, under the same conditions, is a bit less than 20%. I have defined "*hit*", if the distance between the predicted and the measured location was less then 5°.[♦]

Figure 5 indicates the accuracy of the fix channel weighting strategies. The two approaches discussed in the text have been completed with a third one, in which the saliency map based on the Transient channel (which in-filters everything that moves and eliminates all the steady part) forms the final saliency map, on its own. According to literature-data on dynamic input, this channel is outstandingly strong – which is intuitively not surprising, if we take into account how naturally we snap our head at cats, birds, etc., if they abruptly make a motion on the periphery of our sight. These results have been confirmed by my measurements as well (left-most columns in the bar-trios).

On figure 5, "*i*" and "*ii*" refers to the approaches denoted similarly throughout the booklet.

[♦] Counting with 10°, the hit ratio ameliorates significantly – although of course the accidental chance as well. The 5° 'threshold' seemed to be a reasonable choice, both from biological and from evaluational viewpoints.

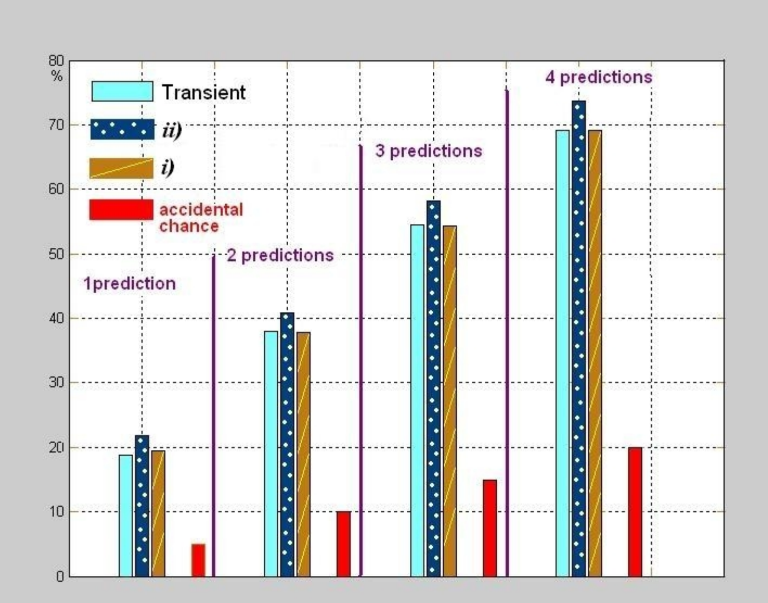

Figure 5. Accuracy of the fix channel weighting strategies.
The first column-trio indicates the percentage, when the *measured* fixation location overlapped with the *predicted* point, which was the most salient location defined by the model adjusted by the previously estimated RF sizes and channel weights. Similarly, the second column-trio indicates the results when the first 2 most salient locations have been determined as possible fixation location, whereas the third trio belongs to 3 predicted locations. Subsequently, the last bar-collection depicts the results belonging for 4 predicted locations. As it can be seen, in ~70% of the cases, the *measured* fixation location was among the *predicted* ones. The accidental chance for this is less than 20% (last, red column).
The different columns in each trio belong to the two different strategies discussed above, plus a third one (indicated with light blue), which shows the results if *only* the Transient channel has been taken into account (- that is, the master saliency map equals with the one belonging to the Transient channel). The fourth, red column shows the accidental chance under the same conditions.

Worthy of note is that two different people, with a good chance, will attend to different locations on the same frame, and also, the same person, observing a given frame in a video-flow at a second time will easily attend to another location than the he observed previously.

4. Applications of the results

Areas where attentional models can be applied are extremely wide, the subtasks and methods employed within them can be used in very many fields. Accordingly, during the last years, I have had the opportunity to test different parts of my model in real practical applications as well – namely in the "*Bionic Eyeglass Project*".

This project meant to help the everyday life of blind or visually impaired people with mobile equipment, via image-flow analysis and different recognition methods. The main lines, alike the subtasks, have been developed together with the expert of the "Hungarian National Association of Blind and Visually Impaired People". Within this project, I have successfully adapted different parts of the discussed model, or rather, of an expanded version of it. This version includes a *preprocessing part designed to stabilize the unstable input* that comes from a camera held by a blind walking person. These video-flows are usually extremely noisy and unstable, often accompanied by fast and unexpected camera motions. Additionally, often the picture's main objects shift significantly from one frame to another, e.g. during turning around. The goal of the image stabilization step is to keep the steady objects (e.g. buildings) in the same pixel positions, while the moving objects (for example the pedestrians) can change position.

The main idea in this step is to combine an optic flow algorithm with an affine transformation model, which can handle translation, scaling, rotation and shear. By using the optic flow algorithm we obtain estimation for the velocity of the pixels by measuring their time and spatial gradients (vertical and horizontal) apiece. Then, with the transformation model, the translation (in vertical and horizontal directions), scaling, rotation and shear *of the frame* can be estimated.

By the usage of the mammalian retina channel decomposition, the classical difficulty that image processing algorithms nowadays face (namely that the intensity or color values of the same object largely depend on the actual lighting conditions) can be avoided – at least partly. This observation has a fundamental importance in practical applications, and it is exploited in the methods aiming to solve the following problems raised within the Bionic Eyeglass Project:

- Locating LED indicators (in real-life indoor and outdoor scenes)
- Finding traffic signs in real-life street scenes.

  The main purpose of these two tasks is to realize a fast method that locates the areas which contain the traffic signs / LED indicators with high probability, on complex real-life outdoor scenes. Subsequently, a classifier algorithm has to analyze only the located ROIs ("Region of Interest") instead of the whole input, which can fasten up the whole process significantly. The main difficulties derive from the instability of the by-default bad-resolution input, the unconstrained lighting conditions, and from the *variety* of the possible inputs.

  The accuracy of the introduced methods is around 80%. The test database has been made out of complex real-life scenes, for all the different tasks.

- Finding light sources (lamps) – which task (although it seems to be a trivial ‘problem’ for a person with normal vision), could prevent annoyance for visually impaired people, for example, by preventing the lamps to remain switched-on for weeks after a guest.

  Here, the most important criterion is that the solution has to be independent from the input’s actual brightness, that is, the accuracy should be the same in the case of a sun-drenched room and a dark cell.

  The method I have introduced relies only on a single retina channel, the “Polar” channel, and achieves a very high accuracy: the ratio of the correct answers is around 99%.

The precise algorithms have been explained in separate publications.

Generally speaking, the possible application-field of a well functioning visual attentional system is extremely wide, starting from different monitoring systems via robot vision up to different ‘bionic’ applications. Nevertheless, a well functioning *bottom-up* system (which I have attempted to produce during my Ph.D. studies) is not a *complete* attentional system. It would be complete, if it had included the so called “top-down” method as well. However, our knowledge of this cortex-originated function is quite restricted for the time being, but at any rate, slimmer than necessary for a reliable and complete model.

At the same time, regarding the above task, *some* knowledge we already possess comes from well known data from the literature, for example, that this method is “fed-back” at the point of summing up the channel-based saliency maps, right before the creation of the final saliency map (figure 1, bottom, middle). On this schema, different practical applications can be constructed, for example via the task-dependent modification of these weights (e.g. finding traffic signs, from the above discussed applications).

5. Acknowledgement

First of all, I would like to thank Professor Tamás Roska, my supervisor, for the vast encouragement, guidance and help, for the stirring discussions and consultations, and, because he could always find time for me in his extremely dense time-schedule.
Similarly, I would like to say thanks for my consultant, Dr. Zoltán Vidnyánszky for his concern about me, for the pieces of advice, for the discussions, the careful reviewing of my publications, and also, that via him I could have a blink into the world and thinking of the biologists.

I would also like to thank my fellows at the Doctoral School for the discussions and assistance concerning a wide range of topics. To István Kóbor, László Havasi, Tamás Bárdi, Dániel Hillier and Gábor Vásárhelyi, with whom I have been sharing a common workplace, and moreover to György Cserey, Kristóf Iván, Gergő Soós, Barna Hegyi, Tamás Harczos, Csaba Benedek, Éva Bankó and Viktor Gál, without whom these years would have been much less exciting.

Also, special thanks for Dávid Bálya, for the help in the retina model I have used during my doctoral studies, and for the guidance he gave, primarily during my first grade.

Similarly, I would like to express my thanks to the “vision”, or rather “Bionic Eyeglass” group for the co-operation: to Dávid Bálya, Róbert Wagner and to Kristóf Karacs.

I am grateful for the help and good advice I have got from Professor György Karmos before my stay in Leuven: it would have been more difficult without them!

I am particularly indebted to Professor Marc Van Hulle as well, who has made possible an interesting and pleasant semester in his laboratory in Leuven. And also, from his group to Karl Pauwels, for his persistent help, and because he always had time for me, even for hours.

I keep beautiful memories of my university years which I had the luck to spend in Szeged. From there, I would like to thank András Kocsor, my former supervisor for guiding my first efforts, my first paper.

And last but not least, I am very-very grateful to my family, to David for being a great brother, and to my parents for their endless patience, love and support, which I have had the good luck to enjoy during all my studies. And also, to Elias, for his constant understanding and help in these ‘dense’ periods.

6. The author's publications

Journal papers

[1] **A. Lázár**, Z. Vidnyánszky, T. Roska, "Modeling stimulus-driven attentional selection in dynamic natural scenes," *International Journal of Circuit Theory and Applications*, (in print)

[2] **A. Lázár,** K. Pauwels, M. Van Hulle, T. Roska, "Scene analysis of unstable video flows – using multiple retina channels and attentional methods," *Integrated Circuits: Research, Technology and Applications*, (accepted)

Conference proceedings

[3] **A. K. Lázár**, R. Wagner, D. Bálya, T. Roska, " Functional representations of retina channels via the refineC retina simulator," *Cellular Neural Networks and their Applications. Proceedings of the 8th IEEE international workshop,* pp. 333-338, *2004*, Budapest

[4] Bálya D., **Lázár A.**, " Retinal processing", *XI. MITT Kongresszus*, Pécs (*2005*)

[5] Vidnyánszky Z., Kovács G., **Lázár A.**, "Active vision" , *XI. MITT Kongresszus*, Pécs (*2005*)

[6] **A. Lázár**, A. Kocsor, " An application of ranking methods: retrieving the importance order of decision factors," *IEEE International Workshop on Soft Computing Applications SOFA 2005*, Szeged, Hungary – Arad, Romania

[7] T. Roska, D. Bálya, **A. Lázár**, K. Karacs, R. Wágner, M. Szuhaj, "System aspects of a bionic eyeglass", *Proc. of International Symposium on Circuits and Systems ISCAS,* pp. 161-164, *2006*, Kos, Greece

[8] **A. Lázár**, T. Roska, "Human Tested Saliency Map Generation in the Bionic Eyeglass Project", *Proceedings of The 10th IEEE International Workshop on Cellular Neural Networks and their Applications,* pp.91-95, *2006*, Istanbul, Turkey

[9] K. Karacs, **A. Lázár**, R. Wagner, D. Bálya, T. Roska, "Bionic Eyeglass: an Audio Guide for Visually Impaired," *Proceedings of the 1st Biomedical Circuits and Systems Conference,* pp. 190-193, *2006,* London, UK

7. Literature related to the thesis

[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. 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

[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 modeling of visual attention,” *Nature Neuroscience,* Vol 2, 2001

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

[8] E. R. Kandel, J. H. Schwartz and Thomas M. Jessell, “Principles of Neural Science” Appleton & Lange, 3rd edition, 1996

[9] D. J. Parkhurst, E. Niebur „Stimulus-driven guidance of visual attention in natural scenes” **In** Neurobiology of attention, (L. Itti, G. Rees, J. K. Tsotsos **Ed.**), pp. 240-245, Elsevier, 2005

[10] R. Carmi, L. Itti „Visual causes versus correlates of attentional selection in dynamic scenes” **In** Vision Research, doi:10.1016/j.visres.2006.08.019 , 2006

[11] C. Koch and S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry”, *Hum. Neurobiol.* 4, 219-227, 1985

[12] H. Nothdurft, “Salience from feature contrast: additivity across dimensions”, *Vision Res.* 40, 1183-1201, 2000

[13] S. Shipp, “The brain circuitry of attention”, *Trends in Cognitive Sciences*, Vol.8 No.5, 2004

[14] L. Itti, “Models of Bottom-up Attention and Saliency”, **In:** Neurobiology of Attention, (L. Itti, G. Rees, J. K. Tsotsos **Ed.**), pp. 576-582, San Diego, CA:Elsevier, Jan 2005.

[15] R. Carmi and L. Itti, “The role of memory in guiding attention during natural vision”, *Journal of Vision*, Vol 6, No.9., pp. 898-914, 2006

[16] B. Zitova, J. Flusser, “Image registration: a survey”, *Image and Vision Comp*. Vol 21, 977-1000, 2003