Modeling stimulus-driven attentional selection in dynamic natural scenes

Anna Lázár[1], Zoltán Vidnyánszky[2], Tamás Roska[3]

1) Péter Pázmány Catholic University, Faculty of Information Technology, Budapest, Hungary, 1083 Budapest, Práter utca 50/a
2) Neurobionics Research Group, Hungarian Academy of Sciences - Peter Pazmany Catholic University - Semmelweis University, 1083 Budapest, Hungary
3) MTA SZTAKI, Computer and Automation Research Institute, Hungarian Academy of Sciences Hungary, 1111 Budapest, Kende u. 13-17

Abstract: In this paper we have developed a neuromorphic model of bottom-up visual attentional selection. The output of a recently developed neuromorphic multi-channel retina model has represented the input of our model. As a first step, a saliency map has been calculated for each retinal channel which, next, has been integrated into a master saliency map. Model parameters have been optimized based on human eye movement data measured during viewing dynamic natural scenes. We have tested two different strategies for weighting the channel-specific saliency maps during integration into a master map. In the first case, channel weights have been kept constant throughout the verification measurements, whereas, in the other case, they have been updated on each frame, according to the specific properties of the visual input. Surprisingly, the constant channel weighting strategies have performed better then the continually updated ones.
We have measured the model's accuracy by defining the hit ratio (concurrence) between the first few predicted locations (the most salient locations) and the measured fixation locations. Constant weighting methods have achieved ~74% hit ratio on 4 predictions. For a comparison, the accidental chance for this case has been less than 20%. This pure bottom-up approach has performed surprisingly well on dynamic natural input. Some practical applications have already been made with task-dependent simplifications.

Keywords*:* visual attention, neuromorphic modeling, eye movements, retina channels, saliency, receptive fields.

1) Introduction

In everyday situations, a large part of the information appearing in the visual scene is redundant and/or uninformative; meanwhile, the processing capacity is limited. This stands for both living organisms and artificial visual systems, so a selection mechanism is required that focuses to the processing capacity onto the immediately important, relevant or conspicuous information. From an engineering viewpoint, a system that is able to *attend*, can save enormous processing capacity. That is: increasing the quality meanwhile decreasing the time necessary for the process.

In the mammals, selection processes involve two main components: eye movement and attentional selection. Eye movement is the mechanism, which determines which part of the visual scene is going to be processed at high resolution by the central, foveal region of the visual system. However, eye movements are preceded by attentional selection processes, which determine the most salient, conspicuous part of the visual scene, where gaze is going to be directed. There are two different types of attentional selection: volitional (top-down) and stimulus-driven (bottom up). Top-down attentional selection is determined by the current goals of the organisms and is mediated by the top-down modulator projections from the front-parietal areas to the visual cortex. For example, searching for a red pen in a crowded drawer will result in a top-down attentional facilitation of the visual cortical neurons coding the red color and suppressing those which are selective for other colors. On the other hand, bottom-up attentional selection is determined by the physical properties of the visual input. In case of abundant visual input - consisting of many different visual objects - there is a competition between the neural representations of different objects that are simultaneously present in the visual scene. Bottom-up attentional selection refers to those mechanisms as a result of which, the most salient visual objects of the scene – according to its physical properties - gain processing advantage and "capture our attention" evoking an eye movement towards it.

Of course, an unmitigated, neuromorph attentional method, which, for example, characterizes mammals, is not only very effective, but also very complex. As mentioned above, it consists not only of the bottom-up, but also, of the top-down method. At the same time, the top-down method builds upon the bottom-up, in the sense that, it biases certain weightings in the bottom-up process (see also section 2.3 and figure 4) Thus, creating an attentional method that approaches the mammalian system in efficiency would have a lot of benefits and could be applied in many areas.

The main purpose of this work has been the development of a model of a bottom-up attentional selection that is able to determine the most conspicuous, salient part of dynamic natural-scene input. At the first stage, visual information has been processed by a recently developed neuromorphic retina model [1] and its output represented the input of our bottom-up selection model.

The usage of CNN (Cellular Neural/Nonlinear Network [2]) based algorithms in handling different visual problems is common: from robot navigation [3] to motion analysis [4] the range is entire. The retina model we use is also CNN-based and has been developed by keeping the main structure of the retina in a manageably simple form [1]. In this model 10 channels are realized from which 7 operate on a CNN simulator. The responses of these channels depend not only on the instantaneous input, but also, on the preceding stimulus, effectively exhibiting a kind of 'memory'. In contrast, the outputs of the other three channels (intensity, red-green- and blue-yellow opposites) depend only on the actual stimuli.

In the CNN simulator, each CNN layer corresponds for the retinal cone, horizontal, bipolar, amacrine and ganglion layers, respectively. In these channel-models there are inhibitory connections between the cone-horizontal and the bipolar-amacrine pairs, as well as

excitatory linking between the cone-bipolar and amacrine-ganglion layer-pairs. The differences lie in the spatial and temporal parameters, which determine the characteristics of the individual channels. For more details see section 2.1 and [1].

The first models of visual attention have been developed in the 1980's, after Treisman and Gelade have proposed their feature integration theory [5], wherein, they have suggested that, only the basic visual dimensions (such as color and orientation), the so called 'low level visual features', are processed throughout the visual field in a parallel way. Afterwards, it is the visual attention which binds together the low-level features belonging to the same object into coherent object representation. This later, attention-based process, takes place in a serial way; attention is allocated to one or at most a few objects at a time.

A detailed bottom-up, stimulus-driven visual attentional model has been proposed by Koch and Ullmann in 1985 [6]. In this model feature-specific "saliency maps" have been calculated for the different visual features (color, orientation, etc.). Saliency maps are scalar, two-dimensional topographic maps, representing feature contrasts, rather than, a given feature's absolute value at each location of the visual field. As a next step, feature-specific saliency maps have been integrated into a so called "master" or "final" saliency map. In the master map the saliency representation was already feature-independent. Lastly, due to a "winner-take-all" mechanism the most salient part of the master map (which has the highest salience value) gains processing advantage and captures attention, while, other salient parts of the map are suppressed.

In the last two decades, several models of visual attentional selection have been developed [7], most of them sharing the main components of the original Koch and Ullmann model [8, 9]. There are some important characteristics of these models: 1. the choice of the low-level visual features was heuristic and depended primarily on the purpose of the given model [10]; 2. weighting of the individual feature-specific saliency maps during integration into a master map was based on top-down approximations, mixing biological findings with heuristic methods to achieve higher efficiency; 3. With a few recent exceptions [11, 12], the models have been tested on static, non-dynamic visual input.

It is also instructive to mention that, in a related paper [13] Osberger and Rohaly have identified some factors on complex scenes, which have strong influence on visual attention. Based on these, they have created a model, that is able to make predictions for human gaze directions. Most of these features were driving the bottom-up process (motion, contrast, etc.), some of these were related with the top-down (people, context), meanwhile some were in "between" (shape, foreground/background distinguishment). They also highlighted the difficulty of the weighting of these features.

In comparison, in our model, we primarily focus on the elaboration of the bottom-up process, taking carefully into account *all* the features that might have any effect on the bottom-up process. This is being achived by including all the retina channels, both these, whose, their function is well understood, and also, those, whose their function is not sufficiently illuminated up to present. Moreover, we manage to give a satisfactory approximation on the weightings of all these features.

In our model, as an input we have used the output of a multi-channel neuromorphic retina model, instead of using heuristic feature extraction. Furthermore, channel weights have been adjusted and the model has been validated based on data obtained from measurements of human eye movements, while viewing dynamic natural scenes. As visual input, we have used short movies containing mountains, birds, clouds, seas, fields and rivers taken from the film "Le Peuple migrateur".

2) The model

2.1) Overview of the model

To get the algorithm complete, first we have made a general bio-inspired framework, in which we have inevitably had some unknown parameters. To define these values we have made human gaze direction measurements from which we have conjectured the missing values, according to different surmises. Once we've had this, we have been able to adjust the framework and make predictions with it. Having at our disposal these forecasts (where we think that a human observer would attend to) and other human gaze-direction measurements, we have been able to determine the accuracy of the gained model. The flow-diagram of the main steps is depicted on figure 1.

Although the description is mathematical, the model itself is not a mathematical input-output approach, but a biological one, which, is partly neuromorphic and partly biologically inspired. It is "neuromorphic" in the sense that, it applies a CNN-based retina simulator (to perform low-level visual feature extraction). Thus, in this step not only the *approach*, but also, the *structure* is neuromorphic, since, the CNN architecture itself is similar to the individual retina layers [1,2]: it has several layers of processing units ('neurons') locally connected to each other ('synapses'), and also, the adjacent layers exchange information between each other. [1]

From then on we have used strategies snooped from biological systems: topographic saliency maps, which are frequent in living visual systems from low until high brain levels, receptive field (RF) structures for saliency detection and competition between neurons.

As it will be explained in detail, two parameters (the weighting of the individual channel-based saliency maps during the creation of the final map and the receptive field sizes) have been calculated from human gaze direction measurements, so that the question, whether they are subject-dependent or not, arises only for these values. The curves recording the efficiency of the different receptive field sizes for the individual channels look similar between the individual subjects, in the sense that, they reach their maximum near to each other, and also, they have the same 'shape'. (The averaged curves are depicted on figures 8, 9 and 10). Thus, this value can be treated as subject-independent.

For the channel-weight parameters the situation is not that clear. In order to estimate, the number of times that each channel-based saliency map is considered to take part in the generation of the master map, we have used the consolidated data of the subjects. The surprising result we have obtained (namely that, constant channel weighting strategies perform better then continually updated ones) might be a result of the differences between the subjects.

(We have measured 21 naïve subjects in the first series from which we have retrieved the missing parameters, and 14 subjects in the second series of measurements, which has been used for the validation process.)

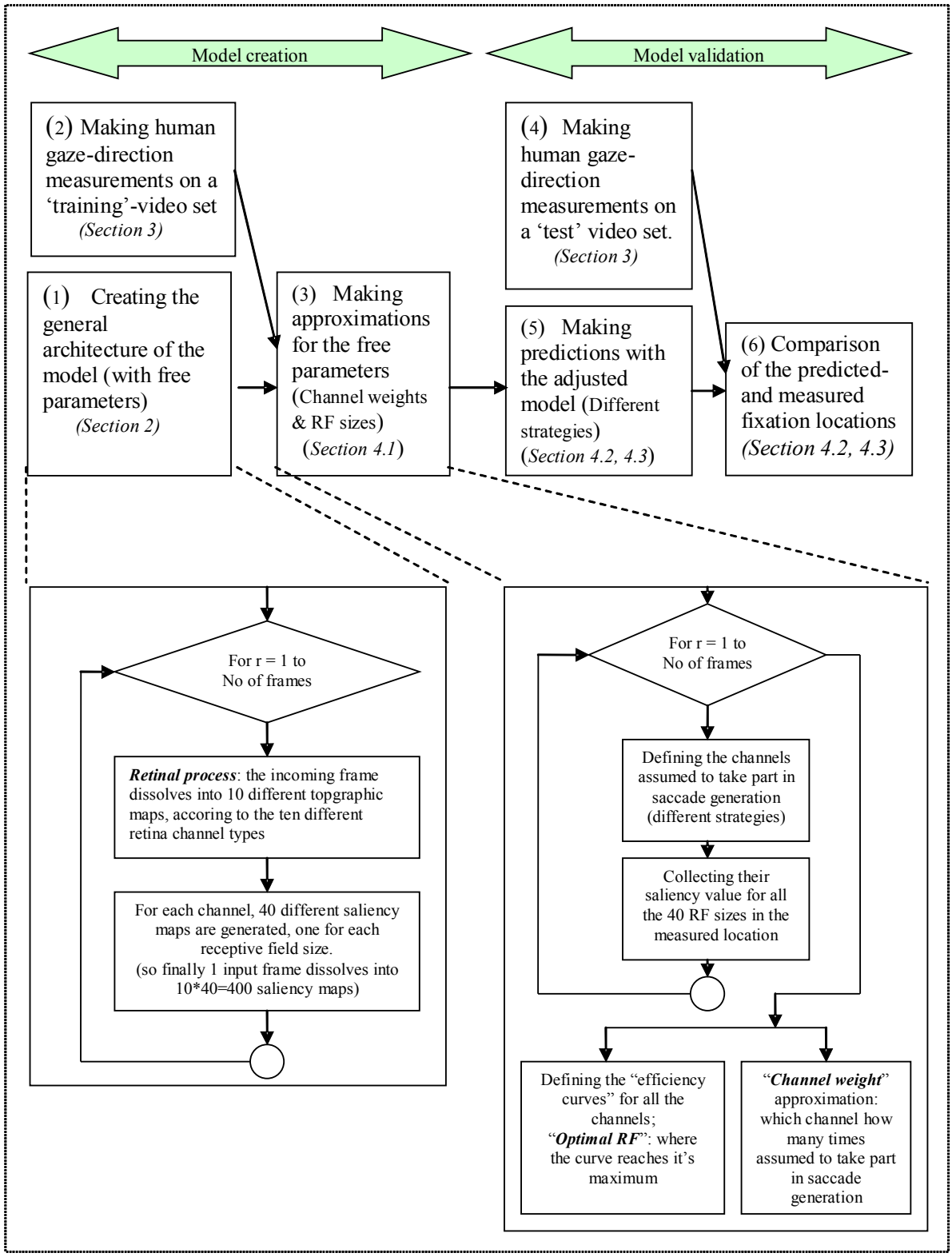

Figure 1: The overview of the model's creation and validation.
Creation: A general bottom-up attentional architecture has been created with free parameters (channel weightings and receptive field (RF) sizes; step 1). Then, we have made human gaze direction measurements on a training-set video (step 2) for approximating the missing data (step 3).
Validation: with the adjusted model, we made predictions (step 5) and human gaze direction measurements (on a different video, on the 'test-set' step 4). Finally we have compared the predicted and measured fixation locations (step 5).
Step 1 and 3 are detailed in the bottom boxes. For every step, we indicate where it can be found in the article.

2.2) The retina channels

Information processing in the mammalian visual system takes place simultaneously on several specialized channels. These spatio-temporal channels arise in the retina and persist to the high brain areas – while performing several processing steps. One of the biggest difficulties which, image processing algorithms nowadays face is that, the intensity (or color) values of the same object largely depend on the actual lighting conditions, reflections, and so on. From an engineering viewpoint, a stimuli-decomposition procedure (like the mammalian retina channels), can be an important step in resolving this classical problem, since, the outcome of most of these channels will not depend on the actual lighting conditions anymore (which is an outcome of the retina channel set-up, see below, and in particular, figure 3).

In the living retina, between the photoreceptors (which intercept the photons) and the ganglion cells (the axons of which form the eyes' "output", the optic nerve) there are several layers and cell-types which already start to process the information here, in the retina. The retina has ten histological layers. The information flows through the vertical pathway composed by the photoreceptors, bipolar cells and ganglion cells. Among these layers lie the two synaptic strata: the Outer Plexiform Layer (OPL) between the photoreceptors and the bipolar cells, and the Inner Plexiform Layer (IPL) between the bipolar cells and the ganglion cells. These strata primarily do not *convey* the information but they *modify* it. [14,15]

Recently, it has been discovered that a mammalian retina has ten parallel channels, and also, the neuromorphic structure of these channels have been found [16]. These channels give qualitatively different answers to the same input. The main differences lay both in spatial and temporal properties. For more biological details we refer to [16].

By these measurements, that having been made on the rabbit retina, only a first and rough approximation has reached completion, and not, the detailed circuitry. Using these findings, this is the first time that we have the possibility to consider this multi-channel pre-processing step. Thus, although the approximation for humans can only be 'bio-inspired' (instead of being 'neuromorphic'), still, we have found the adaptation of this multi-channel pre-processing step in attention modeling fundamental (even without the exact circuitry). As far as we know, this has been the first attempt to use this bio-inspired channel decomposition in attention-modeling. Moreover, using functional spatial temporal models instead of input-output models in the first part enhances the success of model identification.

On the above basis, using a neuromorphic retina-channel model, via Cellular Neural/Nonlinear Network (CNN) [2, 17], we intended to keep the basic processing principles of the retina in a simplified form.

The circuit structure of the mammalian retina, and the multilayer spatial temporal retina model are the same [17]. The receptive filed organizations are represented by the feedback and feedforward templates.

Once a template (or templates: feedback, feedforward, etc.) is defined for a CNN-layer, every cell in this layer operates in the same way, in the sense of their spatial and temporal behaviour. This is also true for the retina layers: each layer consists of the same *type* of cells (photoreceptors, bipolar cells, horizontal cells, etc.), which have similar behaviour. These layers are then connected (both in the retina and the simulator), serially, or create diffusion-layer pairs, which, can also be simulated with CNN (e.g. cone-horizontal layer-pair on figure 2). Due to the above mentioned similarity between the CNN structure and the living retina layers, the CNN simulator seemed to be a proper choice.

Three channels out of the ten (Intensity, Red-green opponency and Blue-yellow opponency) use *only the actual* data for producing the output. Namely, if "R" is the actual 'red' value in a given pixel-position (from the 'RGB' triplet), "G" is the green and "B" is the blue, then [15]

- Intensity has been calculated as: 0.812*G + 0.177*R + 0.1*B,
- Red-green opponency as : R-G
- Blue-yellow opponency as: B – (R+G)/2

These channels do not require simulations as complex as the remaining seven channels, the functioning of which can be described as follows. More concretely, in contrast to the previous three channels, those seven other channels perform spatio-*temporal* filtering (namely the 1: Transient, 2: Local Edge Detector, 3: Bistratified, 4: Alpha, 5: Beta, 6: Delta, 7: Polar), so the output of them can be simulated on a CNN-based retina simulator. The name of this simulator is 'RefineC', and its' functioning has been described in full detail in paper [1]. Briefly, the functioning of these channels is summarized below:

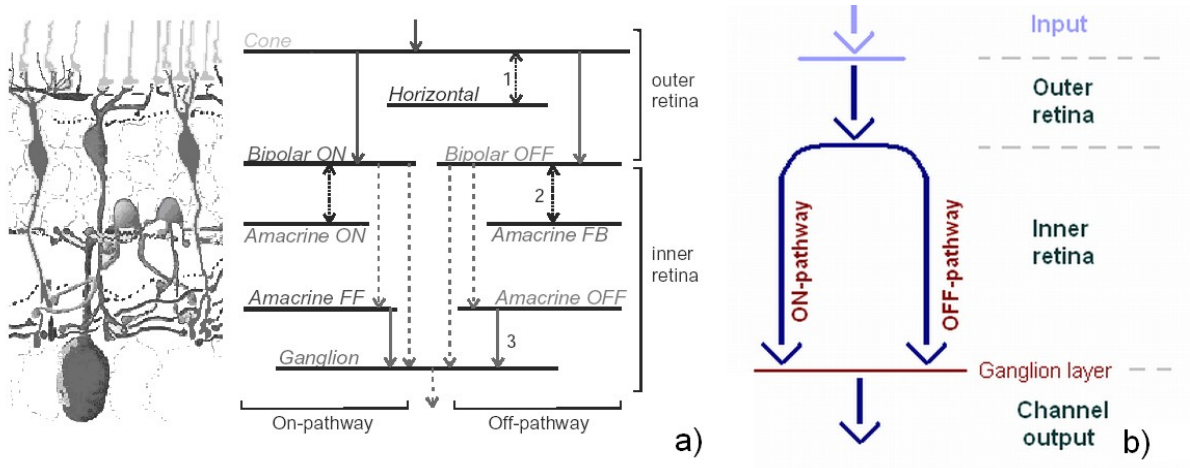

Figure 2: The scheme of a general retina channel b) roughly and a) with CNN layers. In our model we have seven of these, one for each ganglion-output. The interacting diffusion layers are numbered. The dashed lines show the inhibitory connections while the continual ones nominate the excitatory ones. Figure a) is cited from [1].

The sketch of a (general) spatio-temporal channel is depicted on figure 2 a). Each horizontal line on the right hand side is a CNN layer witch corresponds to a retina-layer (depicted on the left hand side of the picture). The outer retina, which is the same for all the channels, consists of the cone and the horizontal layer. The horizontal layer feeds back to the cone layer through an inhibitory connection, thus, the output of the cone layer includes the effect of the horizontal cells as well.

The bipolar cells connect the inner- and the outer retina. From an engineering viewpoint the inner retina can be divided into an On- and an Off-pathway. (Fig 2 b) "On" cells respond during illumination, "Off" cells respond when the light disappears, whereas "On-Off" cells react on both cases.

Each channel consists of three layer-pairs, which are serially connected. The first one is the cone-horizontal, which composes the outer retina. The second one is the amacrine-bipolar, where, the connection is also inhibitory similar to the previous one. The third connection is excitatory between the amacrine and the ganglion layers. The output of the retina is the output of the ganglion layer. Ganglion cells typically have two qualitatively different inputs: an excitatory and an inhibitory one. Excitation comes from the amacrine layer while inhibition derives from the bipolar cells.

We have prepared all the ten, biologically measured [16] and artificially modeled [1] channels. These include two colors-opponency, and one intensity channel, as well as, seven

other spatio-temporal channels. The seven parallel pathways (figure 2) have the same design as described above, and furthermore, they only differ in the parameters that determine their spatio-temporal characteristics.

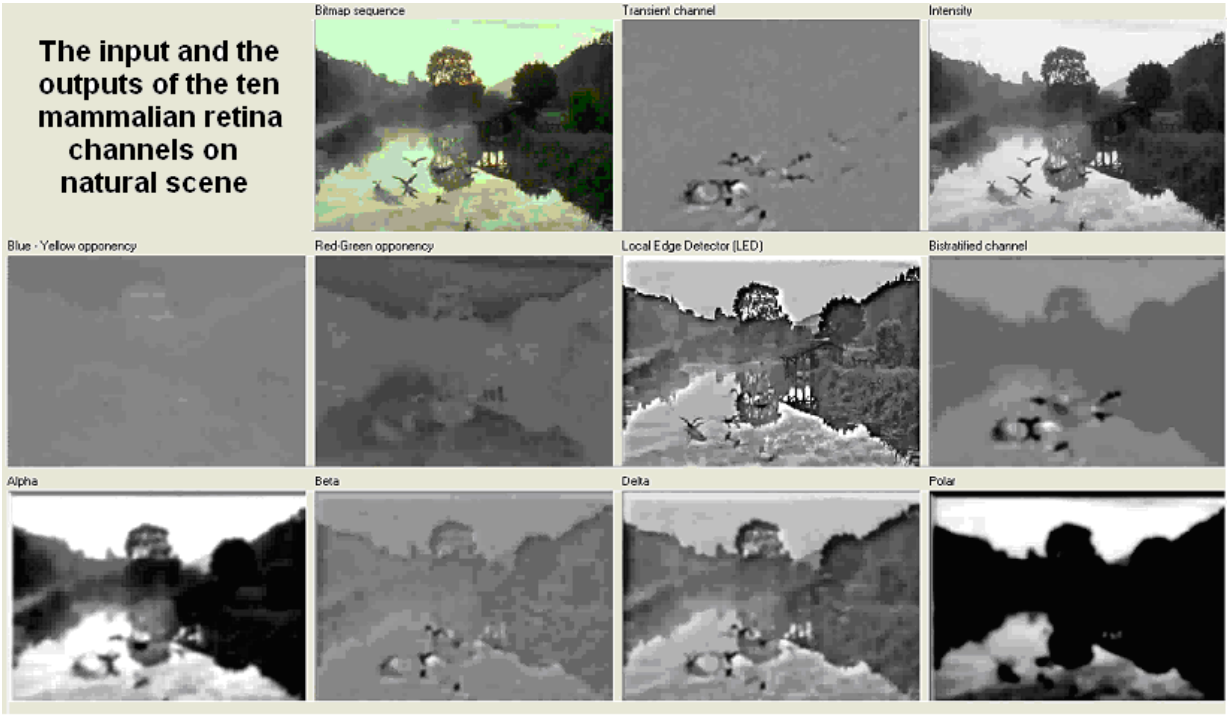

Figure 3: An example for the function of the retina. The input image (first picture) is processed by ten different pathways resulting in 10 ganglion-cell types which form the ten retina channels. [16] The second picture in the first row (next to the input image) is the output of the 'transient' channel which filters out the mobile parts of the visual scene and removes all the steady sections: at this moment the birds flight triggers the biggest response. 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, as well as the bistratified channel's task.

Firstly, we have performed the temporal processing. For this purpose we have used a buffer for the images, which preserved the recently processed sceneries - in the biological equivalent this corresponds to the information which is still under processing in deeper layers of the retina. Practically, this is a fixed-sized buffer, where, the certain positions indicate the time elapsed since the input reached the sensor. Each of these positions has different weights. (It is important to note that working with image frames is a corollary of working with simulators that run on PCs; this is because of the fact that, the retina has no frame-rate or any similar category: it works on a totally analog way, in the sense that the input image flow is continuous in time and value, and there is no time discretization, the only discretization is in space).

Once a new frame is being read, it gets diffused with the former images: that is, the signals which reach the retina beforehand, subsequently reside on different levels of the vertical pathway. The different layers of the retina have different diffusion characteristics: accordingly, the individual positions of the circular buffer have different weights that characterize the diffusion being made on the image restored there. For proper values please refer to [1]. Once this process has been completed, the result overwrites the oldest image. This is the outcome of the specific retina-channel.

Spatial processing is the effect of the diffusions that occur inside the certain layers. From an engineering viewpoint this is the outcome of the subtraction being made between two different diffusions engaged on the last (temporally already processed) frame.

Although some basics are known [15], the precise method explaining how the colors are processed is mostly undiscovered. We have used two color-opposition channels and one intensity. Figure 3 shows a snapshot of the ten retina channels for a natural scene. For detailed description of these channels please refer to [1, 16, 18].

2.3) The bottom-up mechanism

Selective visual attention consists of two different, but nevertheless closely related parallel-working processes: top-down (TD) and bottom-up (BU). TD is voluntary, originates in the higher areas, and testifies complex functions, such as, finding a key on a crowded table [10, 19]. This is strongly influenced by the observer's expectations, memory and purpose. It modifies the BU method via changing the weightings of the different saliency maps (see figure 4). In contrast, BU is fast, unconscious, and comes before getting aware of the scene. This happens, for example, when a flickering point is present in front of an idle background.

The "bottom-up" process is also called "image-based" or "stimulus-driven", indicating the fact that, the corresponding mechanism is based on the saliency values that the different points of the outside world, in their internal representations, reach. Most of the models that work out the BU mechanism use more or less the same principles [10]. Firstly, that a point's saliency is composed of several conspicuous-values – each of which belonging to different low level visual feature-channels. Secondly, that a location's saliency-value basically depends on the surroundings. This means that, a point's conspicuous-value is not equal with its garishness as an 'absolute value', but, it is proportional to the *contrast* that it forms with respect to its surrounding. Thirdly, the final saliency map aggregates the conspicuous-values that belong to the different low-level visual features with different weights. Fourthly, scene understanding and object recognition tightly interplay in gaze-direction. (Nevertheless, numerous data in the literature suggest that human gaze direction on natural scene closely follows the bottom-up mechanism if the subject has no specific task to perform [20] – a finding which we will exploit.)

To sum up, the main steps of the bottom-up method are the followings [10, 19] (Fig 4.):

- Dissolve the incoming picture according to low level visual features: Instead of the mostly heuristic ones (color oppositions, intensity, orientations, junctions, etc), we build our model on real retina channels, see previous section [1, 16]. In half of the channels we can denominate their function (such as: Local Edge Detector (LED) detects edges, or Transient channel detects motion), whereas, the other five channel's function is still unknown – at least we can not phrase it. Therefore these channels (Polar-, Alpha-, Beta-, Delta- and the Bistratified channels, see Fig. 3) have never appeared in heuristic artificial models.
- Create the saliency maps referring to each channel. There are several strategies in order to achieve this; the relevant precept is the measurement of the *contrast* between a point and its surroundings. For the purpose of defining these values, we have used different sized, circle-shaped receptive fields (RF), on and off (section 2.3). (Since different receptive field sizes generate different saliency maps on the same input, and also, the

extent of these RFs are unknown for the certain channels, we have made an *optimization* step.)

- Feature combination:. Unify the feature-based saliency maps into one final one, which is thus already feature independent. In other words, the final (or "master") saliency map is a *combination* of the feature (or channel-) based maps, thus, it does not depend on only one or a few features, but on *all* of them. The weighting of the different channels are not equal. We have used different approaches to estimate these weights: "constant" and "continually updated":
    - *Continually updated channels weighting strategies*: for every frame we have approximated the average and the maximal saliency values appearing on the individual channels, and we supposed that only the first few most salient channels participate in the generation of the master map with weightings that are proportional to their approximated saliency values. The effect of the other channels – on this specific stimulus – is negligible.
    - On the contrary, by *constant channel weighting strategies*, we have presumed that the different channels participate in the formation of the master map with a pre-defined, invariant ratio.
- Determine the most salient point (find the location that has the highest saliency value). This is a winner-tale-all mechanism, which means that, the whole process aims at locating this single point, which will be the attendant location.
- Particularly for still images: create a mechanism, called "inhibition of return", which aims at preventing attention to get stuck 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, and so on. Since we are working with moving pictures, the saliency maps change permanently, so that, this mechanism comes to fruition spontaneously.

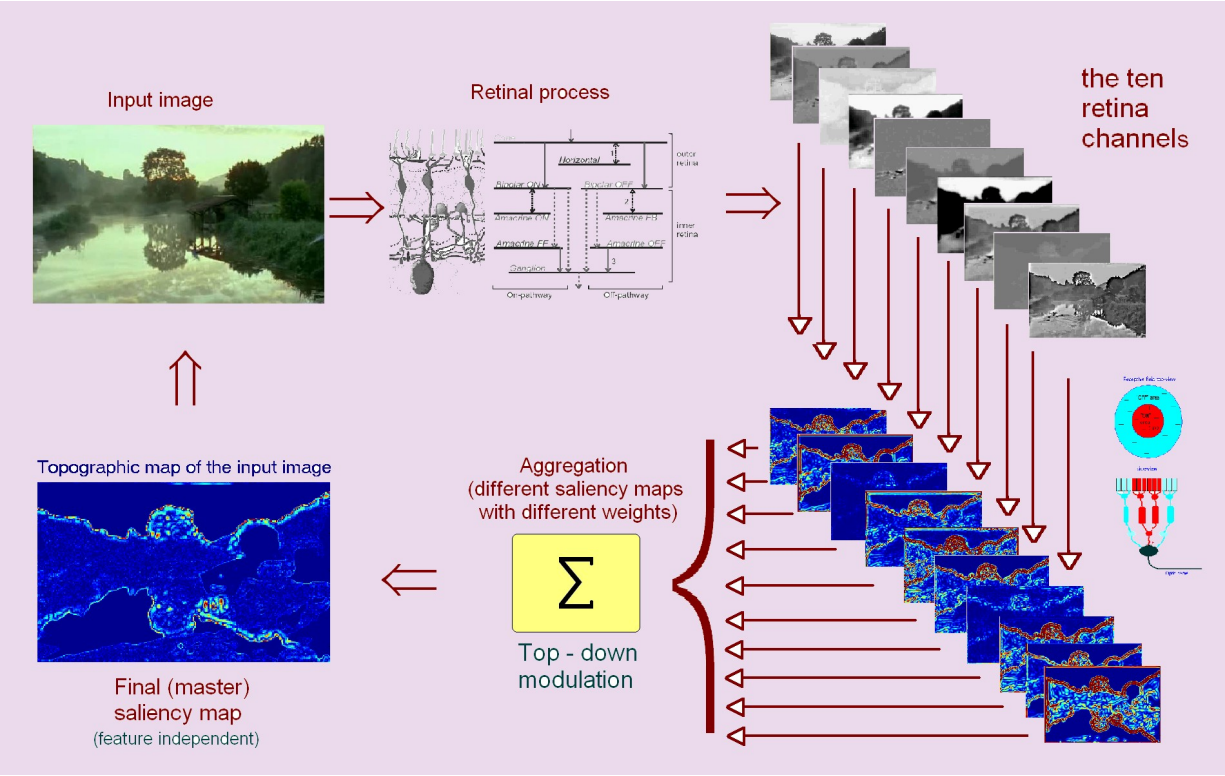

Figure 4: The diagram of the bottom-up mechanism. In the first step the input image (top of the picture, left hand-side) is decomposed into ten different retina channels (-topographical maps in different brain areas: the higher activity a neuron shows, the darker/lighter colour on the monitor appears. This is because we visualized the ON channels- and OFF channels response on the same picture.) In living beings this is a pre-attentive feature extraction mechanism which operates over the entire visual scene in a highly parallel way. Ones the input vision is decomposed, each retina-channel creates its own saliency map. For defining the individual point's saliency value, we used different sized, circle-shaped receptive fields (RF), on and off. The next step is the aggregation. The final (or master) saliency map is practically a weighted sum of the feature-based saliency maps. The weighting of these feature-dependent maps are under top-down modulation, if it is present. (Bottom of the picture) Then the winner-take-all mechanism chooses the final saliency map's most salient point: this point wins the attention, the others are suppressed. The corresponding picture-portion 'appears in the fovea', this is the small part of the visual scene that is processed in detail and the rest is processed only roughly.

2.4) Saliency calculation

For the saliency calculation we have used receptive fields (RFs). Their main structure can be seen on figure 5, which we have approximated according to figure 6. The calculation of the proper values has taken place as follows: ("***x***" is a simple index which corresponds to the size. The index ***x*** takes values from 1 to 40.)

D_t (= d_x) = 4*x*-3, the length of the outer square-side in pixels, see fig. 6
D_b = 2*x*-1, the length of the inner square-side in pixels, that is the central parts square-side
D_k = *x*-1, the width of the surrounding's ring in pixels
$S_b = [D_d/6]$ the length of the square-side that was cut off from the outer square in pixels
$S_k = [D_t/6]$ the length of the square-side that was cut off from the inner square in pixels
$N_b = D_b^2-4S_b^2$ the number of the pixels in the central part
$N_k = D_t^2-4S_k^2-N_b$ the number of the pixels in the surrounding region
$W_k = -A/(255*N_k)$ the weight of the surrounding part; this is necessary for the saliency calculation
$W_b = A/(255*N_b)$ the weight of the inner part; also necessary for the saliency calculation

'A' is an arbitrarily chosen value; corresponding to the maximum saliency value that an RF can return if it receives its optimal stimulus. We call a stimulus 'optimal' if both the central- and the surrounding part of the RF get the stimuli, they respond with the higher intensity. For example, light appears in the central part and disappears in the surrounding area.

For the sake of accuracy, we note that the value x=1 is an exception from the prior formulas. It corresponds to a one pixel centered, eight pixels surrounding receptive field.

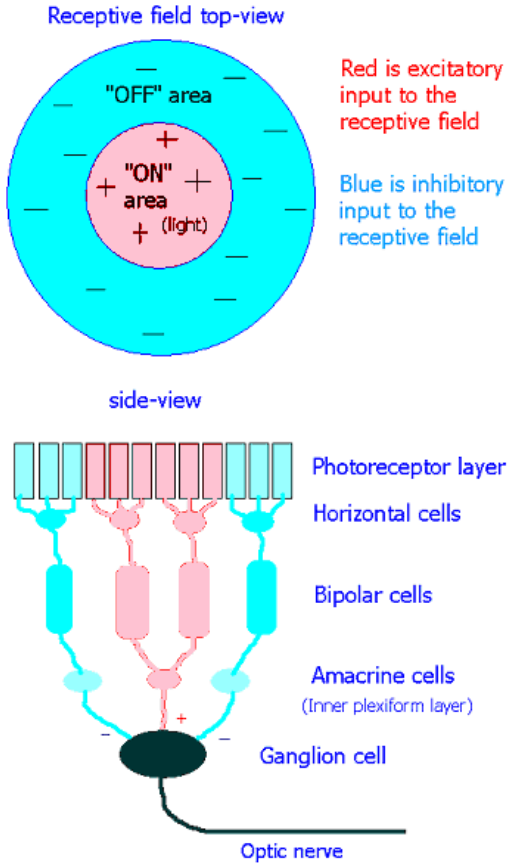

Figure 5: The simplified structure of a saliency receptive field (RF). Neurons are organized into concentric circles: a central- and a peripheral part which respond antagonistically. If the central part of an ON-center - OFF surrounding RF is stimulated with light, then, it will increase the RF's response, 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.

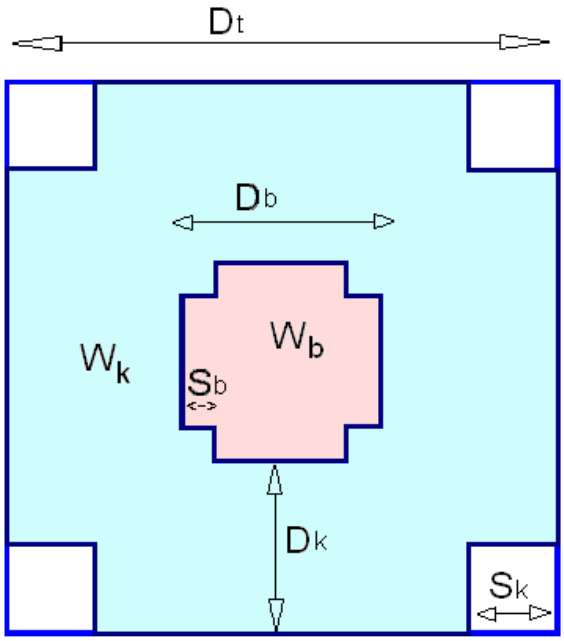

Figure 6: .The structure of a general RF approximation. Circles have been approximated with squares chopped down on their corners. The main principles have been: 1) to keep the neuromorphic ratios: In degree: half of the RF should belong to the central part and half to the surroundings. 2) different RF sizes should return the same saliency value if they receive their optimal stimulus. For precise values see text.

We have prepared receptive fields in 40 different sizes, in order to measure their effectiveness, estimated in terms of the appropriate size providing the maximal saliency value at the attendant location. The smallest was one pixel for the central part surrounded by a one-pixel-width belt. This matched for 0.5°. The largest (the 40th) has had a 79 pixel caliber central region surrounded by a 39 pixel belt. In the measurements this corresponded to 26°. The largest RFs in the human *retina* are about 10° which corresponds approximately to the 20th RF. The receptive field sizes increase linearly with respect to the index '***x***'. (See fig. 7)

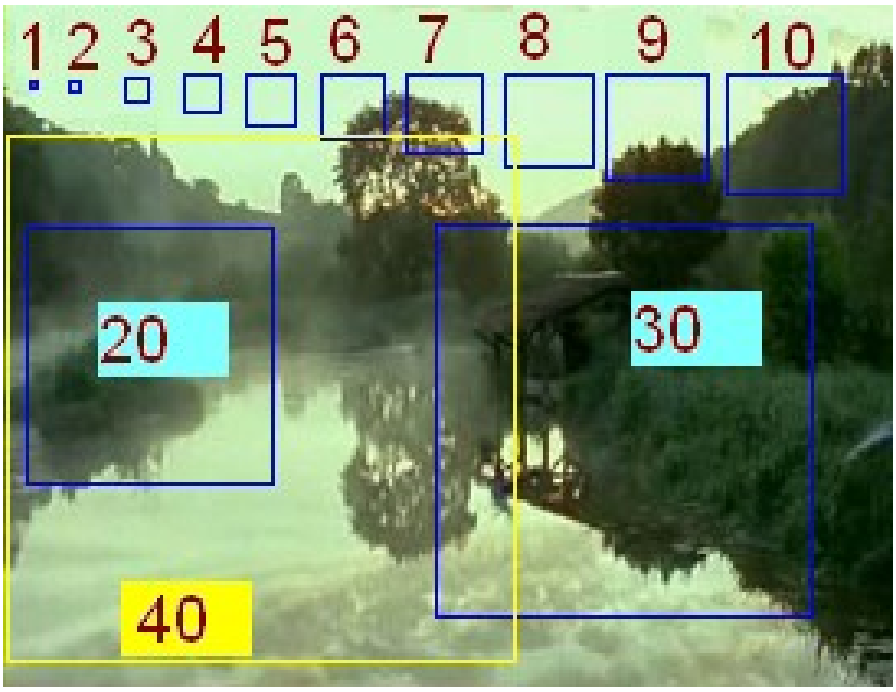

Figure 7: The investigated RF sizes compared to a frame of the stimulus. (The depicted numbers are the '***x***' indices in the above equations.) 1 is the smallest and 40 is the biggest RF.
x = 1 corresponds to ~ 0.5° viewing angle and ***x*** = 40 corresponds to ~ 26°.

From an engineering viewpoint, receptive fields (RF) correspond to filters applied on the different retina channels. The RF_r receptive field is determined by a [d_r by d_r] matrix and the images are defined as $IM_{k,c}$, [M by N] sized matrixes (these are the outputs of the retina channels) for every frame of the input video. The notations we have used are the following:

R	the number of the receptive field types (sizes), R=40
r	(actual) receptive field type, $r \in \{1, 2, ..., R\}$
d_r	the size of the 'r' receptive field, d=4*r-3.
(x, y)	the coordinate of the measured gaze direction (every coordinate-pair belongs to a k frame on which it was measured, where the saccade ended)
M,N	the size of the input video; M: width, N: height; M=273, N=201
C	the number of the channels, C=10
c	(actual) channel number, $c \in \{1, 2, ..., C\}$
K	the number of the frames, K=267
k	(actual) frame number, $k \in \{1, 2, ..., K\}$

The *IM* image matrices contain values between -127 and +128. -127 is black, 128 is white and 0 is middle-grey. This shifting (compared to the conventional bitmap valuing) is due to the antagonistic behavior of the receptive field weights: the outer part of the receptive fields has negative weight, whereas, the inner part has positive. Thus a stimulus, which "fits" to a receptive field, gives a maximal (absolute) value: a big positive, if it fits an ON-centered OFF-surrounded RF, and a big negative ("big" in absolute value) in the other case.

Before the measurements, we have calculated the *SM* saliency matrices for all the *IM* image matrices by all the R=40 receptive fields: this means, K*C*R saliency matrixes, namely 400 for every input frame.

Thus the saliency matrix is different for every frame, every channel and RF type:

$$SM=SM_{(k,\,c,\,r)}= IM_{k,c}.* RF_{r,} \tag{1}$$

where * denotes convolution. With CNN terminology the RFs are the 'A' templates.
We have defined the *SM* saliency map as an *M* by *N* matrix, so we cut down the outer 'ring' of the matrix that has come into existence because of the convolution.

Let us define the (*i,j*) coordinates of the (*x, y*) centered, *d* sized region in an arbitrary matrix as follows (these are simply those matrix-indices that belong to the *d* sized region of the *x-y* point):

$$S_{(x,y,d)} = \{(i, j) \,|\, \max_{\substack{1\le i\le M \\ 1\le j\le N}} \{|\, i - x \,|, |\, j - y \,|\} \le d/2 \} \tag{2}$$

During the measurements we have recorded the (*x*, *y*) coordinate pairs (the fixation locations) and the corresponding *k* frame-number for each (*x*, *y*) fixation location.
Then, we have allocated the proper *SVM* saliency matrix-segments for all these data-triplets:

$SVM^{k,c,r}_{(x,y)}$ is the [d_r by d_r] sized, (x,y) centered segment of the $SM^{k,c,r}$ saliency matrix.

(This means $C*R$=10*40 =400 matrix segments for every measured [(*x,y*), *k*] triplet.)
In order to define the final saliency value in a given location, we've put a discrete Gauss-filter with the same size and position with the receptive field. The next step has been the creation of these filters in all the *R*=40 sizes, the discrete form of (3), where *t* is the radius and σ is the standard deviation.

$$G_t = \frac{1}{\sqrt{2\pi\sigma}} e^{-t^2/2\sigma^2} \tag{3}$$

E.g. the 3 by 3 *G* discrete Gauss filter is: $G_1 = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} /16$ (4)

(We note that the bigger ones can be obtained, for example, with repeated convolution from G_1.)
In CNN terminology these filters also act as 'A' templates.
With these arrangements we can assign a scalar value for every measured [(*x,y*), *k*] data-triplets as follows:

$$SalVal^{k,c,r}_{(x,y)} = \frac{\sum^{M}\sum^{N}(SVM^{k,c,r}_{(x,y)} * G_r)}{\overline{SM^{k,c,r}}} \tag{5}$$

where $\overline{SM^{k,c,r}}$ is the average value of the $SM^{k,c,r}$ matrix and G_r is the discrete Gaussian matrix whose size is also d_r, like as, in the r^{th} receptive fields.

‘*’ can be interpreted as a filter or convolution. In the latter case, again, we “cut down” the outer ring of the result matrix, which happens to be there because of the convolution.

So the normalized saliency value of the c^{th} channel, r^{th} receptive field size, belonging to the k^{th} frame, where, the measured fixation location was (x,y), is defined by equation 5. For adjusting the model’s free parameters these values were essential.

Let P denote the number of *all* the measured (and used) fixations. (Some of the measured fixation locations “fell out” from the processing, for example because they followed a saccade less then 1 degree.) With these, the average saliency value arising on channel c by the r^{th} receptive field is:

$$\overline{SalVal_{c,r}} = \frac{\sum^{P} SalVal^{k,c,r}_{(x,y)}}{P} \tag{6}$$

To find those r^* RF sizes (the “optimal” receptive fields) for every c channel where the relative saliency values reach their maximum we have defined:

$$r^*_c = \arg\max_r \left\{ \overline{SalVal_{c,r}} \right\} \tag{7}$$

These are the x-values (which are receptive field indices) on figure 8, 9 and 10, where, these curves reach their maximums.

3) The measurements

3.1) Stimuli

We have used two video sets. One of them has been used for the basic measurements, the final purpose of which is to find the unknown parameters, namely, the most effective receptive field sizes for the different strategies and the channels weights. The second video has been used for the validation and assessment of the model.
Both video-sets have contained moving natural scenes, each, without any humans or any artificial environment: birds, horses, rivers flowers, sees, mountains, etc... Our goal with this restriction has been to keep the top-down attentional influence as low as possible.

The stimulus was 8 frame/second video, 512x298 pixel/frame, 96 dpi each. No audio was added. The first (“basic”) video-set included 4 clippets, 267 frames, ~33 seconds. The validation set contained 9 clippets with a sum of 447 frames, ~56 seconds.

Participants have been asked to watch both videos 4 times in the following order: 2 for the “basic” video, then, 2 for the validation video, then 2 for the “basic” video again, and finally 2 for the validation one.

3.2) Participants

21 naïve human observers have participated in the first "basic-" video-set measurements (and 2 non-naïve) and 14 naïve (plus 1 non-naïve) in the second series. Non-naïve participant's data has not been included in the evaluation. Each subject has had normal or corrected-to-normal vision.

3.3) Experimental design

The distance between the subject's eyes and the monitor was 50 cm; the inner part of the monitor was 40 cm x 30 cm. We have recorded saccade-end locations and in the first series we have processed the data belonging to saccades bigger then 1 degree, in order to find out the most BU-modified fixations. During 66 (naive) trials in the first case we have recorded 3995 fixations, from which, 2560 saccades were bigger then 1 degree.
The second run, for validation, has included 54 trials with 6430 saccade end-location recordings.

4) Results

4.1) Estimating the most effective receptive field sizes and channel weights

There are two "key-parameters" in the model: the *size* of the receptive fields (which might differ on each channel) and the *weights* of the feature-dependent saliency maps during the creation of the master saliency map. In the same time, when we measure human eye-movements, our usable information is the position and the saliency of the saccade-end location.

To estimate these parameters, our first step was to determine the effectiveness of the different receptive field (RF) sizes from ~0.5° up to ~26° with a step of ~0.6° (40 different sizes), for all the ten channels. (fig. 7, results on fig. 8-10). We have defined 'effectiveness' as the saliency value that the given RF provides at the attendant locations – attendant by humans on the first video set (see section 3.1).

Since it is very likely that the evoking a saccade is not a result of the ten channel's equal contribution, firstly, we have made assumptions regarding which channel, when, and, in what measure, participate in the determination of the new fixation location, thus, evoking a saccade.

Essentially we have used two approaches in order to filter out the saccade-triggering channels. Figures 8 and 9 belong to the first approach, whereas, figure 10 to the second one. In more detail, it is instructive to describe them as follows:

I. The first approach is based on the phenomenon that characterizes attentional mechanism in almost all levels, namely that stimuli *compete* with each other for attention. To realize this, we have treated those channels as saccade-generators, which have created the highest saliency at the attendant location by *any* RF size. In other words, just one outstandingly high saliency value being created on a given channel by a given RF size, is enough for the

characterization of this channel as a saccade-triggering one. Figures 8 and 9 depict the results for the two strategies based on this phenomenon. The corresponding curves look very similar, according to the expectations. With the same object, and for the purpose of comparison *among channels*, on figure 8, we have applied the same range on the y axis for all the channels, while, figures 9 and 10 have different scale for the certain channels for the purpose of showing the fine details among the curves. In one of these approaches (figure 8), we have set a threshold for channel filtering. The threshold was a given percent of the maximum saliency value that has come into existence on the given frame: 95% on fig *8. a)* and 70% on *b)*. In the other strategy (fig 9) we have graded the channels according to their *highest* saliency value and have taken out the first few channels (1 and 3 on figure *9 a)* and *b)*, respectively).

II. In the context of the second approach, we have assumed that those channels take part in the provocation of a saccade, being salient, in *average,* on the actual stimuli. The biological background of this approach is that every channel has a big range of different sized receptive fields, but their distribution could differ strongly. To define the mean saliency, we have determined the saliency values according to all the receptive field sizes (40) at the measured attendant location and have averaged it. Figure 10 shows our results for this approach; *a)* if we regard only one channel as effector for the certain saccades and *b)* if we take the first 3 channels. The bottom-most diagram on these figures shows the channel-distribution in percent, that is: for each channel how often it has been interpreted as the triggering one. These values have been used to approximate constant channel weights. For results see figure 11: In comparison, the second approach has proved to be more efficient.

These values depicted on the next figures have been determined on the frame where the saccade presumably evoked in the position where the given saccade ended. Accordingly, we have used the preceding frame compared to the one that we have measured the saccade end-location on. Since we have applied 8 fps video, meaning 125 ms retrace in time, which we interpret as the period between the saccade initialization and the fixation on the saccade end-location.

The bottom most graphs on each ensemble show the occurrences of the different channels (in percent) for the different strategies and parameters. Nicely appears the important role of the Transient channel - which filters out *changes* [16, 18] - on the second bar of each graph. We have also used these frequency distributions as an approximation for constant channel weights (section 4.2).

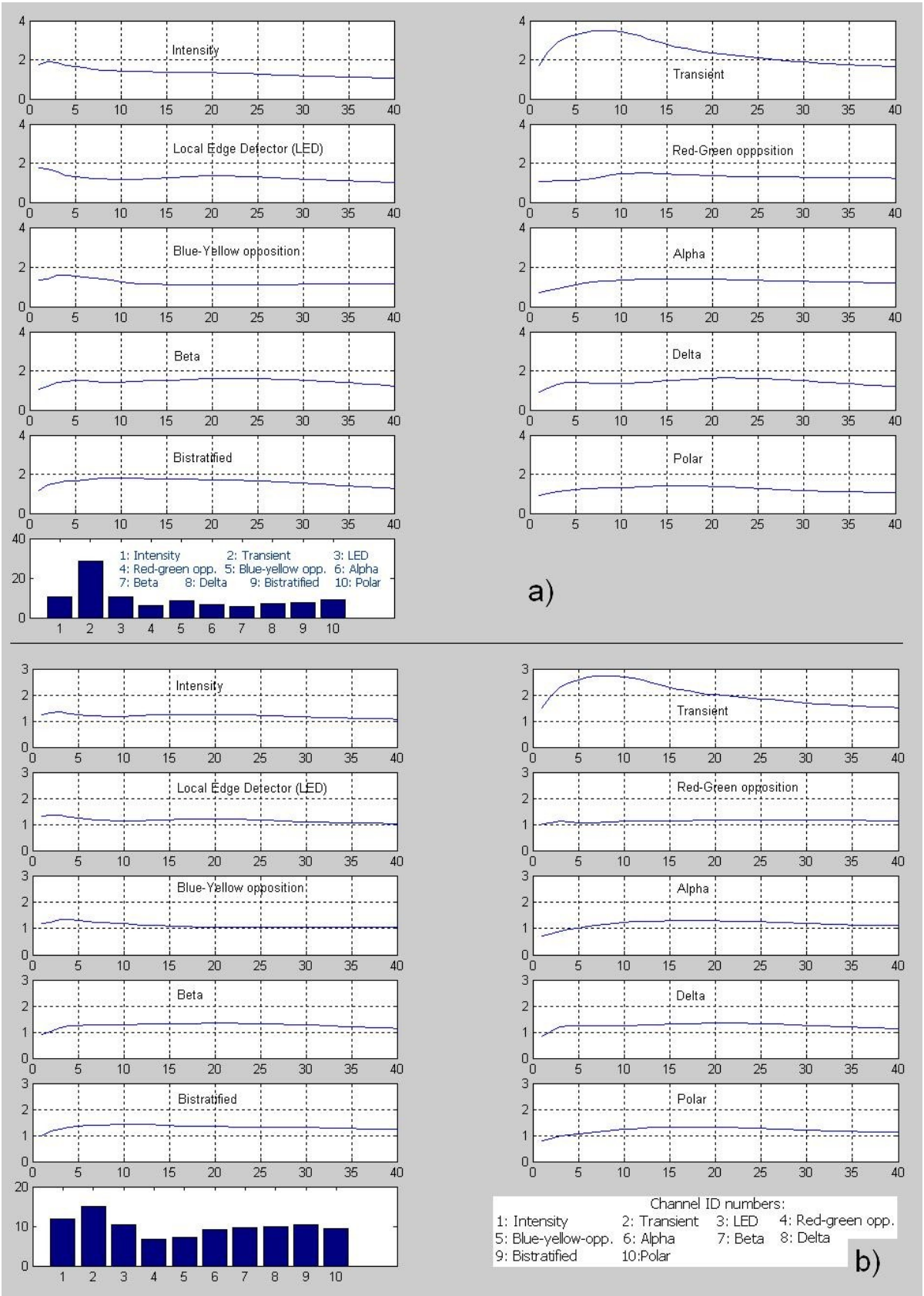

Figure 8: Average saliency values in the attendant locations in the function of the receptive field sizes. For each frame we calculated the saliency values for each channel, each RF size and defined the maximum among all the channels, all the RFs. Here the surmise was that those channels generate the saccade, which has at least one RF which by the saliency value reaches ***a***) 95%, ***b***) 70% of the maximum.
This approach gives very similar outcome then the previous one (see also fig 9.) For better comparison between channels here we applied the same scale on the y axis for each channel.

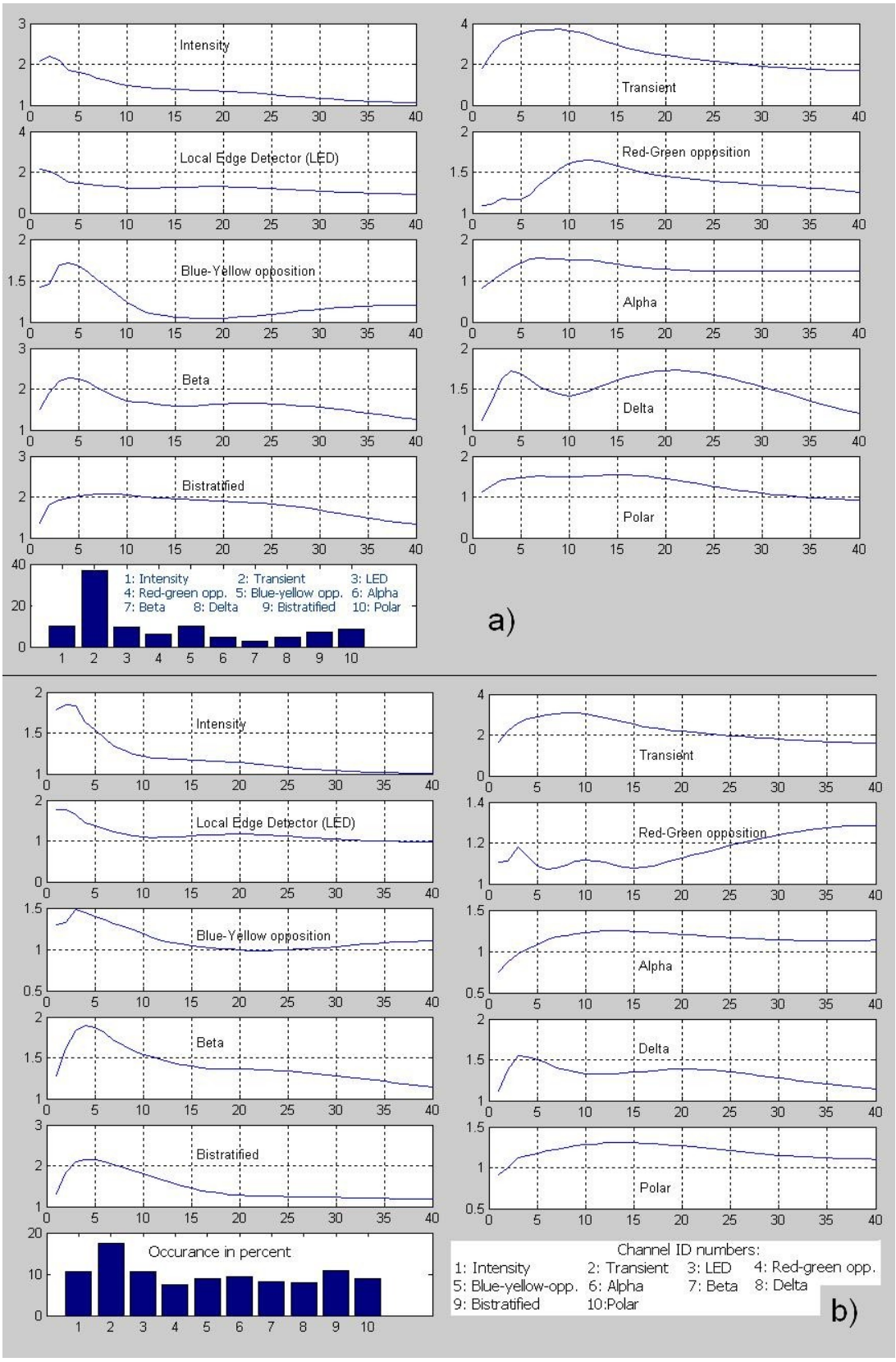

Figure 9: Average saliency values in the attendant locations in the function of the receptive field sizes. For each frame we calculated the saliency values for each channel, each RF size, and assumed, that those channels evoke the saccade, that have *at least one* RF that gives prominently high saliency value: biggest, than other channels can create with any receptive field size. Figure ***a***) shows the results if we defined only one channel as 'saccade-triggering channel', while the lower one (***b***) indicate the outcome when we counted the first 3 most salient channels for each saccade.

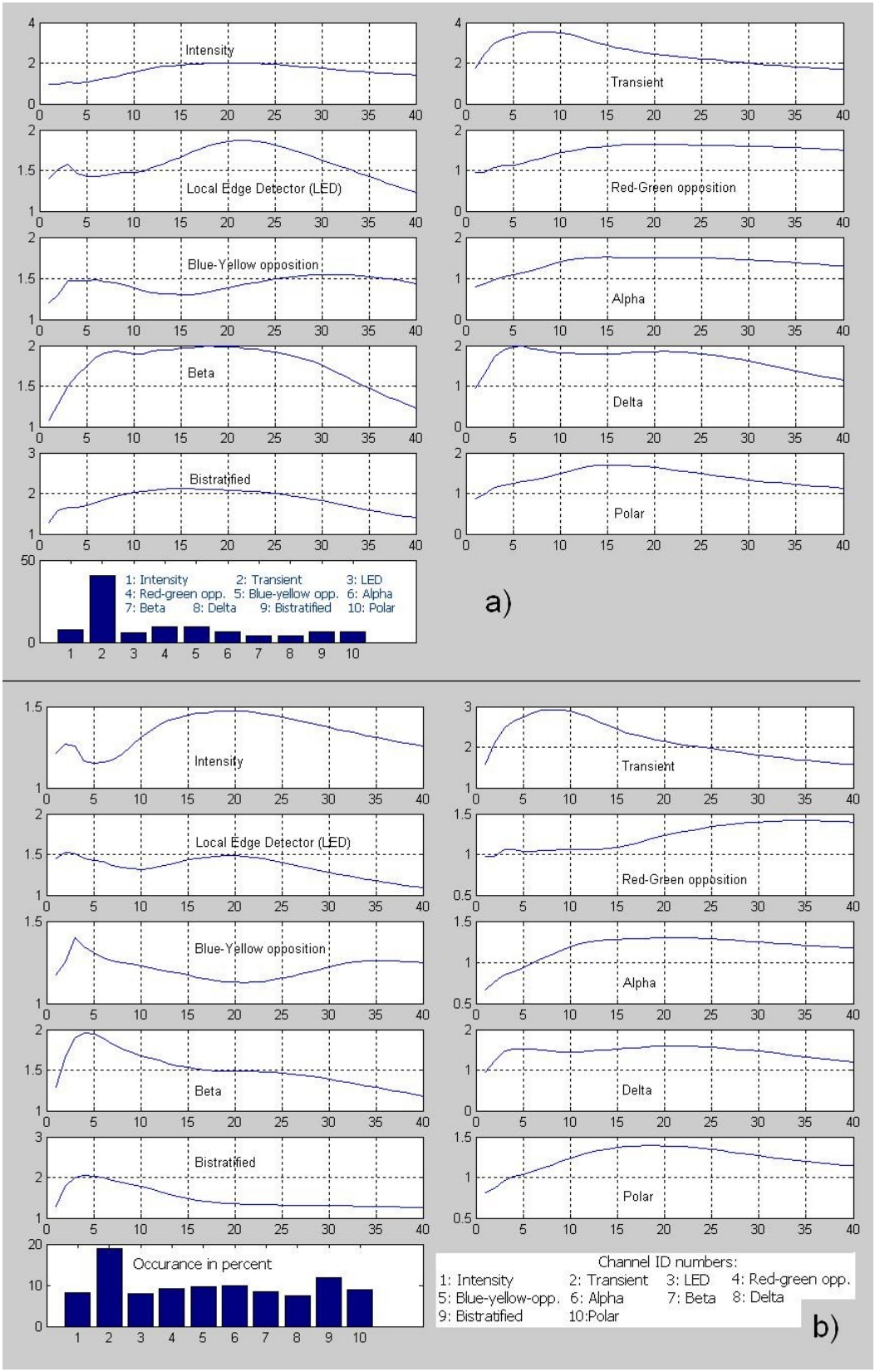

Figure 10: Average saliency values in the attendant locations in the function of the receptive field sizes. For each frame we defined the average saliency value for each channel (between all the 40 RF sizes) and assumed that those channels evoke the saccade, which gives the biggest *average* saliency on that particular stimulus (frame). The left picture (***a***) shows the result if we define the one most salient channel as saccade triggering, while figure ***b***) shows the outcome if we treat the first three most salient channels as effectors.

4.2) Verification results on constant channel weights

To test how close the models predictions are to human fixations, we have made the following:

- First we have calculated the master saliency maps using the channel-distributions yield from the above strategies as channel weights. More precisely, the individual channel weights were the percentage of how many times could the given channel appear as saccade-triggering…
    - …according to *any* receptive field size, ("aRf" on the figures). For rough values see figure *9 a*, bottom-most picture, for exact numbers please refer to the appendix 6.2, Table I, W1.
    - …according to its *optimal* receptive field, where, "optimal" is where the above curves reach their maximum. (Nominated with "OptRf" on the figures.) The exact saliency-map weights for this strategy can be found in appendix 6.2, as "W2" in Table I, and appendix 6.3 contains the indexes of the optimal receptive fields in Table II.
    - …and which channel how often has given the biggest *average* saliency. ("Avg" on the figures) For rough values see figure *10 a*, bottom-most picture, for exact numbers see appendix 6.2, W3.
- Then we have made "predictions" for the gaze directions. These were *locations* (*x*-*y* coordinate pairs) which, the model has calculated as the most probable fixation locations. This means that, if the model and the used assumptions are correct, a human will attend these locations with a higher probability than to other points. There were *more* of these predicted locations to every frame, ordered by decreasing probability: the first location has been calculated as the most likely fixation location, the second one as the second most probable, and so on. (Figures 11-14 depict the first four.) Practically, these probability values were *saliency values* calculated according to the different approaches (- different approaches mean different saliency map weighting strategies, explained above).
- After this we have made human gaze direction measurements, with the purpose of comparing the predicted locations with the 'real', human gaze directions.
- We have defined "hit", as if the distance between the predicted and measured fixation location was less than 5 degrees. (Accordingly, accidental chance was the product of an area of a 5 degrees-radius circle and the number of predictions (1, 2, 3 or 4), divided by the area of the monitor.)

Mathematically: Accidental chance (C_a) has been calculated as follows:

$$C_a = \frac{T_{rh} \times N_{pred}}{T_m} \qquad \text{(e1)}$$

where, T_{rh} denotes the area of a (in this case) 5 degree-radius circle, N_{pred} is the number of the predictions (from 1 to 4) and T_m indicates the sphere of the monitor.

Figure 11 shows the results for the above strategies.

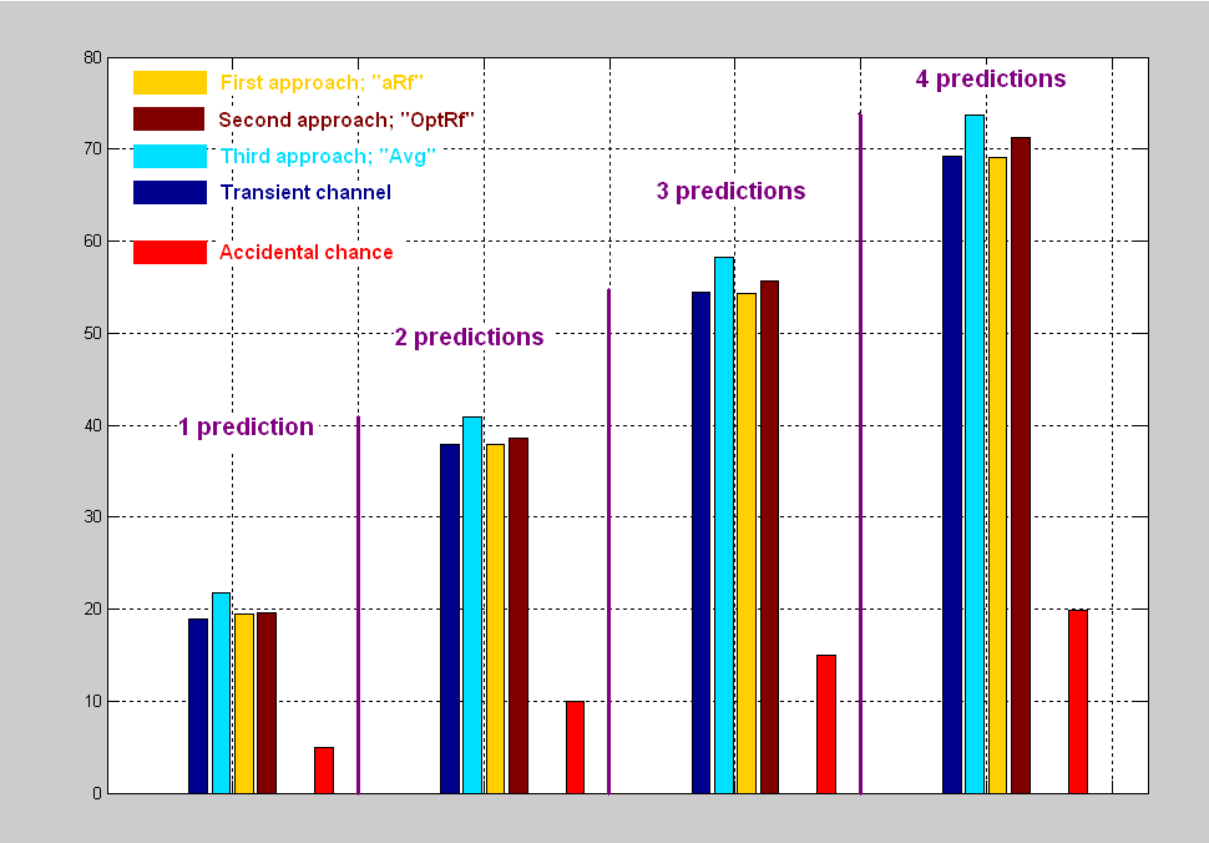

Figure 11: The validation results of the different constant strategies. The left-most bars in each quaternary indicated the outcome when only the 'Transient' channel is adverted and all the others are neglected. The next three bars show the results of the 'Avg', 'aRf' and 'OprRf' strategies, in this order. In these cases the master saliency maps apply all the channel-dependent maps, the differences lie in their weights. (For proper values see Appendix)

It is remarkable that, the transient channel's saliency map alone is almost as effective as the other strategies. This highlights the channel's important role on *dynamic* stimuli, which has been detected by other models as well [11]. On the whole, these approximations are quite effective: the first four predictions contain the measured fixation location with ~70%, for arbitrary subject.

This shows, that on *moving* stimuli, in the involuntary attentional method (bottom-up), the commanding role of the Transient channel seems to be undoubted. Nevertheless, in a general attention model, probably all the channels have their own role; it is enough to think of *motionless* stimulus, where, the Transient channel does not give any response, thus, the whole attentional method has to be under the control of other channels as well. Moreover, under top-down conditions, during which, search being based on complex visual features, comes to the front, probably these channels do have their important role as well.

4.3) Verification results on dynamic (continually updated) channel weights

Dynamic (continually updated) strategies presume that the triggering channels and their weights depend on the stimuli as well. This is not trivial, since channel weighting is primarily under top-down effects, which we have tried to minimize with our stimuli type. Still, we had expected higher efficiency from the continually updated channel weighting strategies, than from the constant ones. Still, the investigated approaches were efficient, with 3 predictions we

could yield a hit rate around 50%, and the first 4 prediction overlaps with the measured fixation location with ~60% chance. The results are depicted on figure 12, 13 and 14. In contrast with our expectations the constant strategies turned out to be more effective than the continually updated ones – this is not anticipated even if the differences are quite small. (see fig. 11, 14)

Each picture contains bar-quintuplets for every prediction-number and a sixth bar for comparison, which indicates the accidental chance. For the sake of a better contrast, we have elongated their height towards the quintuplet they belong to. The first bar always indicates the results for the case when, only the first most salient channel has been regarded as saccade-triggering; the second one, when the first two most salient channels have been taken into account, etc. Accordingly, the fifth bar shows the accuracy of the prediction, which is based on the five most salient channels. Although we have calculated all the ten cases, for the sake of simplicity, we have only depicted half of them – which already show the tendencies.

The two investigated strategies have been the same then before: In the first one, for which the results can be seen on figure 12, the surmise was that the channels that trigger the saccades are those which are salient on the given stimuli *in average*. To *estimate* this average saliency value for the certain channels we have calculated their saliency maps with the "*r*" receptive field size, with which, they proved to be the most effective in average – that is, where the curves on figure 10 reach their maximum. For the proper RF sizes see Appendix 6.3. As it can be seen, while on constant channel weighting cases this approximation has proved to be the most effective, on the continually updated ones this one lags behind the others. (See figure 11, 12 and 14, where 14 is a summary for the continually updated results.) Once we've had the channel-dependent saliency maps, we have defined their mean values. As next step, we have taken out the first *i* saliency maps to which the highest mean value belonged to (for all $i \in \{1, 2, \ldots, 10\}$, one after the other), and have created the master saliency map with proportional weighting to the average saliency values. (See equation 2) The model's predictions were the locations with the highest saliency values in the master saliency map, with at least 4.2° (viewing angle) distance between them. We have defined "hit", if the distance between the predicted and the measured location was less then 5 degree.

The other approach (results on fig. 13) differed from the above one, on the one hand, in the definition of "the most salient channel", and on the other hand, in the "*r*" receptive field sizes with which the channels-based saliency maps have been created. (See equation 3) For RFs we have used those ones, where the curves on figure 8 and 9 reach their maximum. (They represent similar approaches, where by similar RF sizes proved to be the most effective ones; for proper values see Appendix.) According to this approach, "channel-saliency" has been defined as the *maximum* channel-dependent saliency value – instead of the *average*. In other words, we have determined the biggest values in all the channel-based saliency maps and created the master map with the first *i* saliency maps containing the highest values, for all $i \in \{1, 2, \ldots, 10\}$, one after the other. The weighting was proportional with these maximum values.

That is, for the first "Avg" approach the final saliency map was calculated as follows:

$$FinalSM_k^{Avg,i} = \sum_{c=1}^{10} w_c \overline{SM_{k,c}} \qquad \text{where} \qquad \text{(e2)}$$

$$w_c = \begin{cases} mean(SM_{k,c}), \text{if } mean(SM_{k,c}) \text{ is in the first } i \text{ biggest values among } mean(SM_{k,c})\text{s}, \\ 0, otherwise \end{cases}$$

$c \in \{1, 2, \ldots, 10\}$

and

$\overline{SM_{k,c}} = ((IM_{k,c} * RF_r) * G_r) / mean(SM_{k,c})$, the normalized saliency matrix.

where "*" denotes convolution, and the inner brackets contain the saliency map before Gauss-filtering.

k	is the frame number
c	is the channel identifier, $c \in \{1, 2, \ ,10\}$, 1: Intensity, 2: Transient, etc.
i	is the number of channels we regard as saccade-triggering
w_c	the weight of the c^{th} channel in the final saliency map
$SM_{k,c}$	saliency map, belonging to the channel c on frame k.
$IM_{k,c}$	the activation map of channel c on frame k.
RF_r	Receptive field with the r size-index.
G_r	Discrete Gauss-filter with the r size-index.

The "r" indexes for the two approaches ("Avg" and "Mrf") differ; for proper values see Appendix.

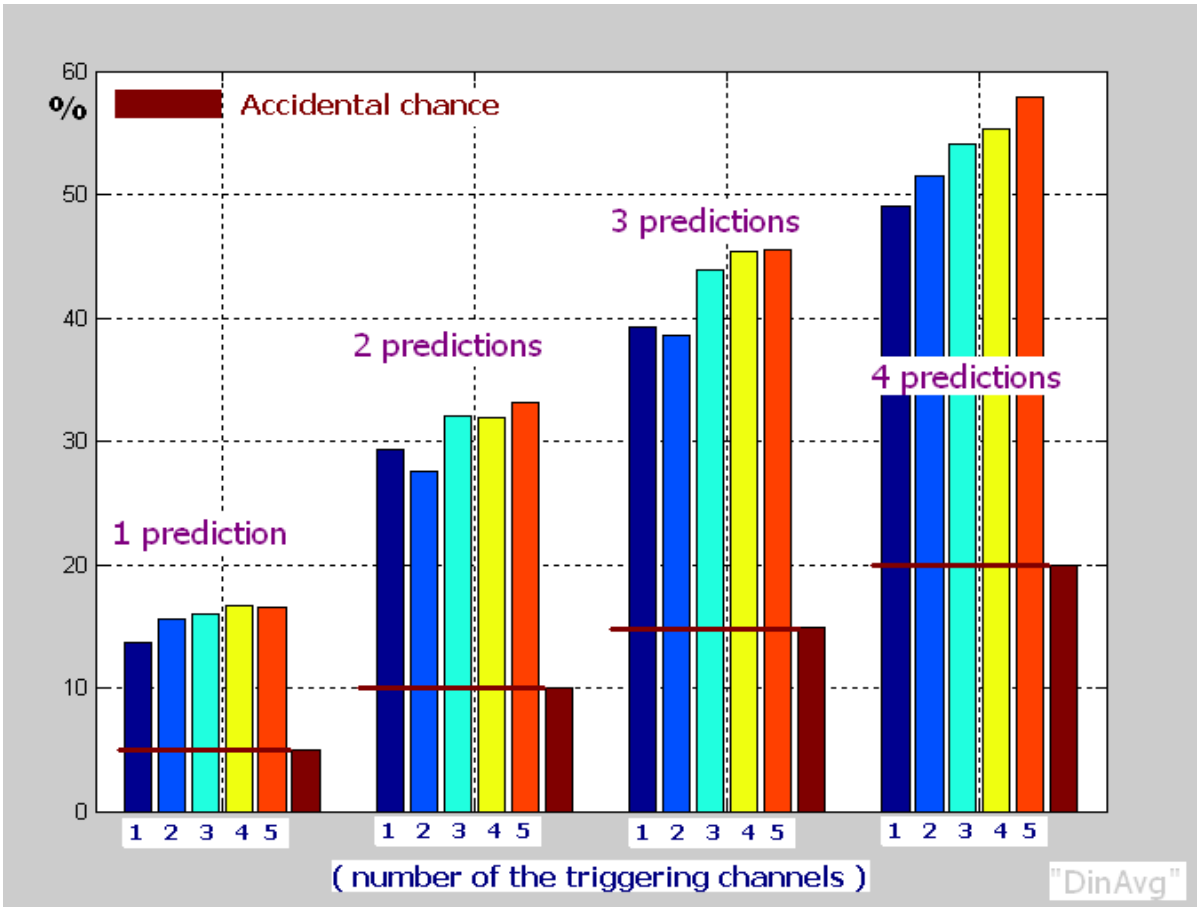

Figure 12: Verification data with dynamic channel choice, showing the approach denoted with 'Avg'. Here we assumed that a channel participates in triggering a saccade, if it is salient on the given stimuli (frame) *in average*. The first bar in every quintuplets shows the results for the case when only one channel creates the final saliency map, the second bar if two channels, etc. For the sake of simplicity we only depict the first 5 bar instead of the 10. (This already shows tendencies.) The horizontal bars indicate the accidental chance for making a 'hit'.
Once the saliency maps were ready, we made our predictions as the locations with the highest saliency values in the master map, with at least 4.2° distance between them. We defined 'hit', if the distance between the predicted and the measured location was less then 5 degree. We defined two predictions 'different', if their distance was at least 2.9 degree (agrees with the receptive field size index: 5)

Similarly, for the "Mrf" approach the final saliency map was calculated as:

$$FinalSM_k^{Mrf,i} = \sum_{c=1}^{10} w_c \overline{SM_{k,c}} \qquad \text{where} \qquad \text{(e3)}$$

$$w_c = \begin{cases} \max(SM_{k,c}), & \text{if } max(SM_{k,c}) \text{ is in the first } i \text{ biggest values among } max(SM_{k,c})\text{s, } c \in \{1, \\ 0, otherwise \end{cases}$$

2, , 10}

and

$$\overline{SM_{k,c}} = (IM_{k,c} * RF_r) * G_r / mean(SM_{k,c})$$,the normalized saliency map.

"*" denotes convolution again.

For the same strategy, if we choose the w_c saliency-map weights for the *normalized* saliency map instead if the un-normalized $SM_{k,c}$ map, we get a more effective method, for which the results can be seen on figure **13 b**.

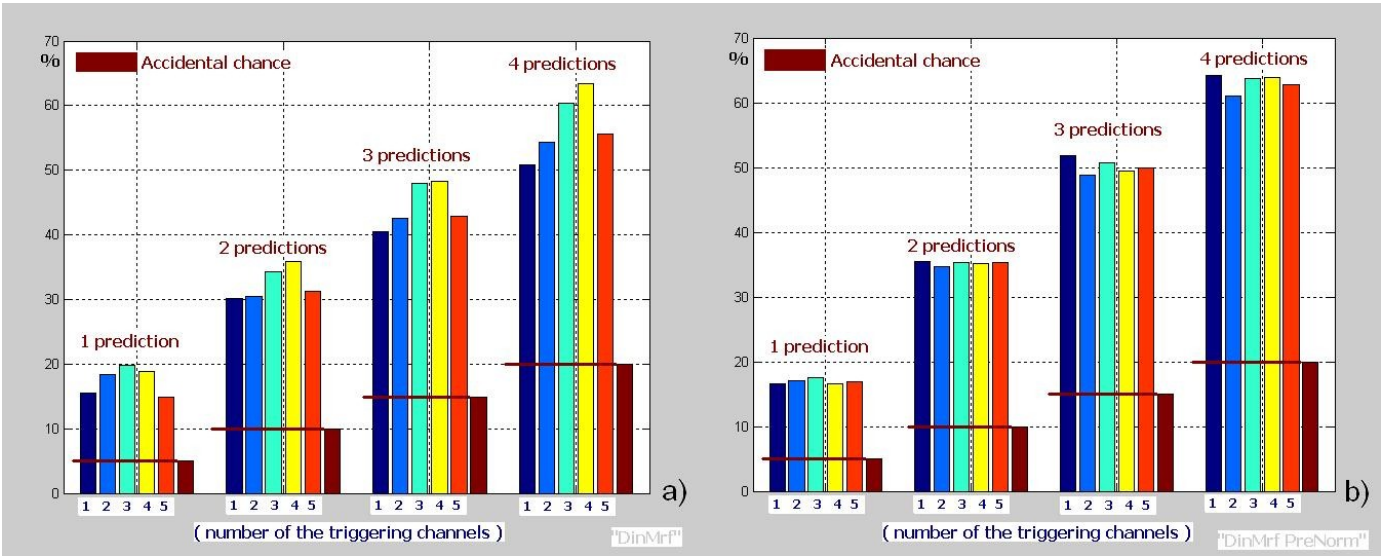

Figure 13: Verification data with dynamic channel choice, showing the approaches denoted with 'Mrf'. Here we surmised that a channel participates in triggering a saccade, if it is 'very' salient *anywhere* on the given stimuli (frame) by its 'optimal' receptive field. The first bar in every quintuplets shows the results for the case when only one channel creates the final saliency map, the second bar if two channels, etc. For the sake of simplicity we only depict the first 5 bar instead of the 10. (This already shows tendencies.) The horizontal bars indicate the accidental chance for making a 'hit'.
Once the saliency maps were ready, we made our predictions as the locations with the highest saliency values in the master map, with at least 4.2° distance between them. We defined 'hit', if the distance between the predicted and the measured location was less then 5 degree. We defined two predictions 'different', if their distance was at least 2.9 degree (agrees with the receptive field size index: 5)
a) shows the results if we yield the channel- weights from the un-normalized channel-dependent saliency map, and **b)** depicts the case, when these are from the normalized map.

For a comparison of the three strategies see figure 14.

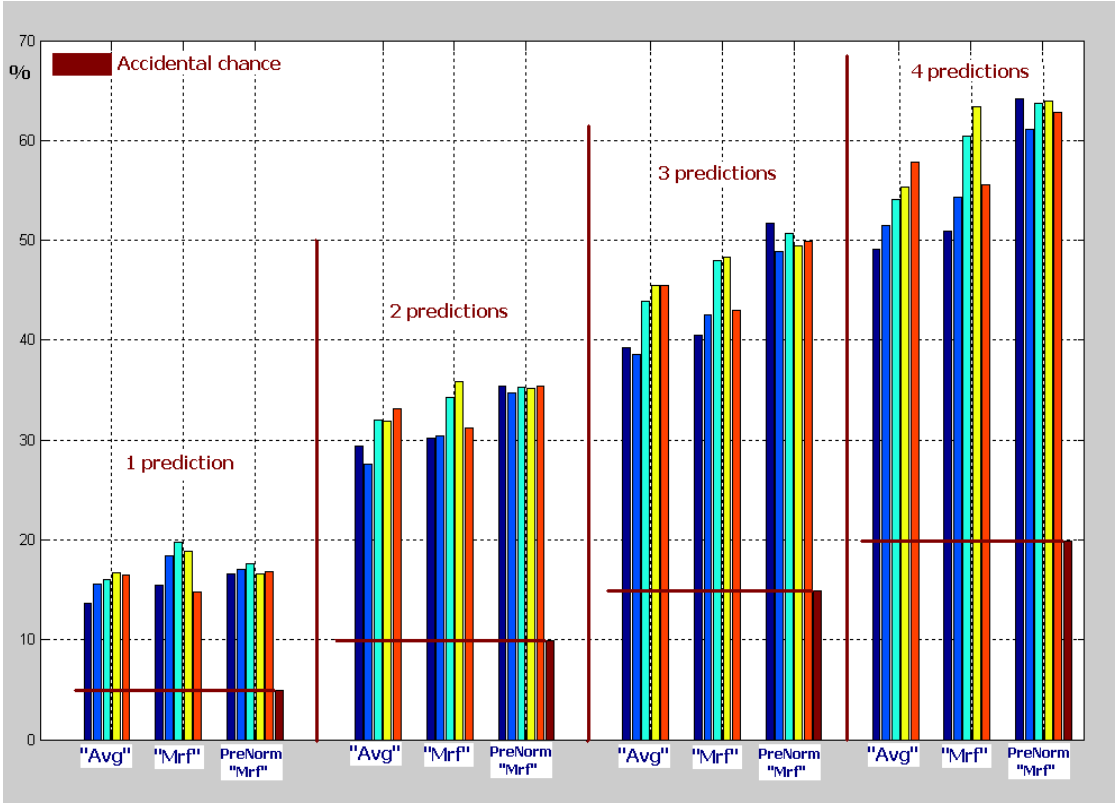

Figure 14: Comparison of the efficiency of the three continually updated strategies. For details of these approaches see text.

5) Conclusions

The goal of the present study has been the development of a bio-inspired model of stimulus-driven, bottom up attentional selection, whose performance will match as closely as possible human attentional selection - as reflected in their eye movements –under free-viewing conditions in case of dynamic natural scenes. To this end, as an input we have used short movies of dynamic natural scenes, which – according to recent studies - under free-viewing conditions evoke primarily bottom-up attentional selection mechanisms [20]. The important novel properties of our model are: firstly, that it is built on real retina channels instead of a few heuristic ones, and secondly, that the parameters (*receptive field sizes* for the individual channels and the channel *weights* during the creation of the master saliency map) are set after human measurements.

We have used two different channel weighting strategies. In the case of constant weights, during the verification experiment, we have used the same channel weights, whereas, in the case of the dynamic, continuously updated channel weights were updated on each frame, according to the specific properties of the visual input. The constant weights approach proved to be the more efficient. The probability that the first 4 predicted locations will include the location of the first human fixations measured on the same input reached 74% in the case of constant weights approach. For comparison, the probability of a match between human fixations and arbitrary predicted locations is less then 20% under the very same conditions. However, in the case of continuously updated channel weights, the hit ratio was around 65%

on 4 predictions (see figures 11-14). These results – although unexpected – from an engineering point of view turn out to be advantageous, since constant weights approach, as compared to the continuously updated approach, requires less processing resources, it is easier to implement and faster.

The developed model might have a broad range of application by adapting the channel weights and receptive field shapes to the specific task-requirements, as well as, by using only the task-relevant channels. With these modifications, there are already some practical applications, primarily, in the so called "Bionic Eyeglass Project", an on-going project meaning to help the everyday-life of blind or visually impaired people. In this, we have three main "scenes": home, workplace and the way between them, with sub-tasks to be solved. By modifying the *shape* of the receptive fields we get an effective algorithm to define the direction of an escalator – a sub-task on the street. If the receptive field is horizontal bar–shaped (in the brain hierarchy, from V1 such RFs can also be found), then the edges of the escalator will be salient. From here, the task is to define whether their vertical coordinates lessen (the steps move away) or grow (the steps draw near). Or, as another modification, by combining different channel information, primary beaming (lamps indoor, sky outdoor) can be detected. For performing this task in the same project, the data from two channels has proved to be enough. Thus, the proposed algorithm is general, in the sense that, it is capable of receiving any kind of input (not only natural scenes) and, with task-dependent simplifications, localize different objects or features.

6) Appendix

6.1) Channel order

The order of the channels:

1) Intensity
2) Transient
3) LED (Local Edge Detector)
4) Red-green opposition.
5) Blue-yellow opposition
6) Alpha
7) Beta
8) Delta
9) Bistratified
10) Polar

6.2) Constant channel weights

The constant channel-based saliency map weights for the three strategies investigated in section 4.2:

TABLE I

THE STIMULUS-INDPENDENT CHANNEL WEIGHTS DURING THE CREATION OF THE FINAL SALIENCY MAP. THE DIFFERENT ROWS BELONG TO DIFFERENT STRATEGIES.

	Intensity	Transient	LED	Red-Green opp.	Blue-Yellow opp.	Alpha	Beta	Delta	Bistratified	Polar
W1 "aRf"	9.75	36.55	9.52	6.18	9.86	4.6	2.91	4.87	7.33	8.4
W2 "OptRf"	7.83	26.33	10.32	7.83	10.36	5.52	4.95	3.99	6.79	16.05
W3 "Avg"	7.25	40.07	5.87	9.21	9.52	6.67	4.33	4.10	6.71	6.21

That is, for all w1, w2 and w3 weight-vectors, the final saliency map is weighted sum of the channel-dependent saliency matrices:

$$FinalSM_k = \sum_{c=1}^{10} w_c \overline{SM_{k,c}}$$

E.g. for the k^{th} frame, according to the w2 weights-vector, the master saliency map is:

$$FinalSM_k = 7.83 * \overline{SM_{k,Intenzi}} + 26.33 * \overline{SM_{k,Tranzi}} + 10.32 * \overline{SM_{k,Edge}} + 7.83 * \overline{SM_{k,Red-Green}} + 10.36 * \overline{SM_{k,Blue-Yellow}} + 5.52 * \overline{SM_{k,Alpha}} + 4.95 * \overline{SM_{k,Beta}} + 3.99 * \overline{SM_{k,Delta}} + 6.79 * \overline{SM_{k,Bistrati}} + 16.05 * \overline{SM_{k,Polar}}$$

6.3) Optimal receptive field sizes

TABLE II

THE 'OPTIMAL' RECEPTIVE FIELD SIZES USED IN THE DIFFERENT STIMULUS-INDPENDENT CHANNEL WEIGHTING STRATEGIES

	Intensity	Transient	LED	Red-Green opp.	Blue-Yellow opp.	Alpha	Beta	Delta	Bistratified	Polar
"aRf"	3	9	2	12	3	12	4	4	4	15
"Avg"	20	9	21	20	31	22	19	4	15	15

The second row depicts the receptive field indexes courted in section 4.1, figure 8 and 9, and used in section 4.2. The third row contains the receptive field indexes used to approximate the average saliency values in section 4.3.

The index i means α view-angle as next:

$$tg\frac{\alpha}{2} = \frac{4i - 3}{100} 0.147$$

where the $4i$-3 defines the diameter of the receptive field in pixels.

7) References

[1] 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

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

[3] X. Vilasís-Cardona, S. Luegno, J. Solsona, A. Maraschini, G. Apicella and M. Balsi, "Guiding a mobile robot with cellular neural networks", *Int. J. Circ Theor Appl*., Vol 30, pp. 611-624, 2002

[4] B. E. Shi, "An eight layer cellular neural network for spatio-temporal image filtering", *Int. J. Circ Theor Appl*., Vol 34, pp. 141-164, 2006

[5] A. M. Treisman and G. Galade, "A feature-integration theory of attention", *Cogn. Psychol*. 12, 97-136, 1980

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

[7] J. Khan and O. Komogortsev "A hybrid scheme for perceptual object window design with joint scene analysis and eye-gaze tracking for media encoding based on perceptual attention", *SPIE Journal of Electronic Imaging* 15 (2), April, 2006.

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

[9] 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.

[10] L. Itti and Christof Koch, "Computational modeling of visual attention," *Nature Neuroscience,* Vol 2, 2001

[11] 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

[12] 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

[13] W. Osberger and A. M. Rohaly "Automatic detection of regions of interest in complex video sequences", *Human Vision and Electroning Imaging VI, Proceedings of SPIE*, Vol. 4299, 2001

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

[15] E. R. Kandel and J. H. Schwartz, "Principles of Neuroscience" Elsevier, New York, 3rd edition, 1991

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

[17] 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

[18] 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

[19] L. Itti, "Modeling Primate Visual Attention," **In:** *Computational Neuroscience: A Comprehensive*

Approach, (J. Feng **Ed.**), pp. 635-655, Boca Raton: CRC Press, 2003

[20] 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