

# Modeling attention in engineering

Matei Mancias

*Numediart Institute, Faculty of Engineering (FPMs), University of Mons (UMONS)  
31 Bd. Dolez, 7000 Mons  
Belgium*

## 1. Attention in computer science: idea and approaches

There are two main approaches to attention modeling in computer science. The first one is based on the notion of "saliency", while the second one is based on the idea of "visibility". The number of papers and the amount of work is dramatically different between these two approaches and the models based on saliency are by far more spread than the visibility models in computer science.

The notion of "saliency" implies a competition between "bottom-up" or exogenous and "top-down" or endogenous information. The idea of bottom-up saliency maps is that the sight or gaze of people will direct to areas which, in some way, stand out from the background based on novel or rare features. This bottom-up saliency can be modulated by top-down information based on memory, emotions or goals. The eye movements can be computed from the saliency map by using winner-take-all (Itti et al. (1998)) or more dynamical algorithms (Mancias, Pirri & Pizzoli (2011) Le Meur & Liu (2015)).

The second approach to attention modeling is based on the notion of "visibility" which assumes that people look to locations that will lead to successful task performance. Those models are dynamic and intend to maximize the information acquired by the eye (the visibility) of eccentric regions compared to the current eye fixation to solve a given task (which can also be simply free viewing). In this case top-down information is naturally included in the notion of task along with the dynamic bottom-up information maximization. The eye movements are in this approach directly an output from the model and do not have to be inferred from a "saliency map" which is considered as a surface giving the posterior probability (following each fixation) that the target is at each scene location Geisler & Cormack (2011).

## 2. Visibility models

Compared to other Bayesian frameworks, like the one of Oliva et al. (2003), visibility models have one main difference. The saliency map is dynamic even for static images, as it will change depending on the eye fixations and not only the signal features: of course, given the resolution drop-off from the fixation point to the periphery, it is clear that some features are well identified in some eye fixation, while less or even not visible during other eye fixations. At the contrary of saliency models, visibility models make explicit the resolution variability of the retina: in that way an attention map is "re-computed" at each new fixation, as the feature visibility changes at each of these fixations.

Najemnik & Geisler (2005) found that an ideal observer based on a Bayesian framework can predict eye search patterns including the number of saccades needed to find a target, the amount of time needed as well as the saccades spatial distribution.

Other authors like Legge et al. (2002) proposed a visibility model capable to predict the eye fixations during the task of reading. In the same way, Reninger used similar approaches for the task of shape recognition. Tatler (2007) introduces a tendency of the eye gaze to stay in the middle of the scene to maximize the visibility over the image (which reminds the centered preference for natural images or centred Gaussian bias illustrated in Figure 9).

The visibility models are much more used in the case of strong tasks and few of them are applied to free viewing which is considered as a weak task Geisler & Cormack (2011).

### **3. Saliency approaches: bottom-up methods**

While visibility models are more used in cognitive sciences and with strong tasks, in computer science, bottom-up approaches use features extracted only once from the signal independently from the eye fixations, such as luminance, color, orientation, texture, objects relative position or even simply neighborhoods or patches from the signal. Once those features are extracted, all the existing methods are essentially based on the same principle: looking for contrasted, rare, surprising, novel, worthy to learn, less compressible, maximizing the information areas. All those definitions are actually synonyms and they all amount to searching for some unusual features in a given context which can be spatial or temporal. In the following, we provide examples of contexts used for different kind of signals.

#### **3.1 Still images**

The literature is very active concerning still images saliency models. While some years ago only some labs in the world were working on this topic, nowadays hundreds of different models are available. Those models have various implementations and technical approaches even if initially they all derive from the same idea.

It is thus very hard to find a simple taxonomy which classifies all the methods. Some attempts of taxonomies proposed an opposition between "biologically-driven" and "mathematically-based" methods with a third class including "top-down information". This approach implies that only some methods can handle top-down information while all bottom-up methods could use top-down information more or less naturally. Another difficult point is to judge the biological plausibility which can be obvious for some methods but much less for the others. Another criterion is the computational time or the algorithm complexity, but it is very difficult to make this comparison as all the existing models do not provide cues about their complexity. Finally a classification of methods based on center-surround contrast compared to information theory based methods do not take into account different approaches as the spectral residual one for example. Other taxonomies will also be introduced in the next chapters as for example the dependence on image features. Here, we show a taxonomy of the saliency methods which is based on the context that those methods take into account to exhibit signal novelty. In this framework, there are three classes of methods.

The first one is pixel's surroundings: here a pixel, a group of pixels or a patch is compared with its surroundings at one or several scales.

A second class of methods will use as a context the entire image and compare pixels or patches of pixels with other pixels or patches from other locations in the image but not necessarily in the surroundings of the initial patch. Some models even use more than one image as a context: an entire dataset can be used here.

Finally, the third class will take into account a context which is based on a model of what the normality should be.

In the following sections, these three classes of models are illustrated.

### 3.1.0.1 Context: pixel's surroundings

This approach is initially based on a biological motivation. Its origins come from the work of Koch & Ullman (1985) on attention modeling. The main idea is to compute visual features at several scales in parallel, to apply center-surround inhibition, combination into conspicuity maps (one per feature) and finally to fuse them into a single saliency map. There are a lot of models derived from this approach which mainly use local center-surround contrast as a local measure of novelty. A good example of this family of approaches is the Itti's model (Figure 1) Itti et al. (1998) which is the first implementation of the Koch and Ullman model. It is composed of three main steps. First, three types of static visual features are selected (colors, intensity and orientations) at several scales. The second step is the center-surround inhibition which will provide high response in case of high contrast, while it will have low response in case of low contrast. This step results in a set of feature maps for each scale. The third step consists in an across-scale combination, followed by normalization to form "conspicuity" maps which are single multiscale contrast maps for each feature. Finally, a linear combination is made to achieve inter-features fusion. Itti proposed several combination strategies: a simple and efficient one is to provide higher weights to conspicuity maps which have global peaks much bigger than their mean. This is an interesting step which integrates global information in addition to the local multi-scale contrast information.

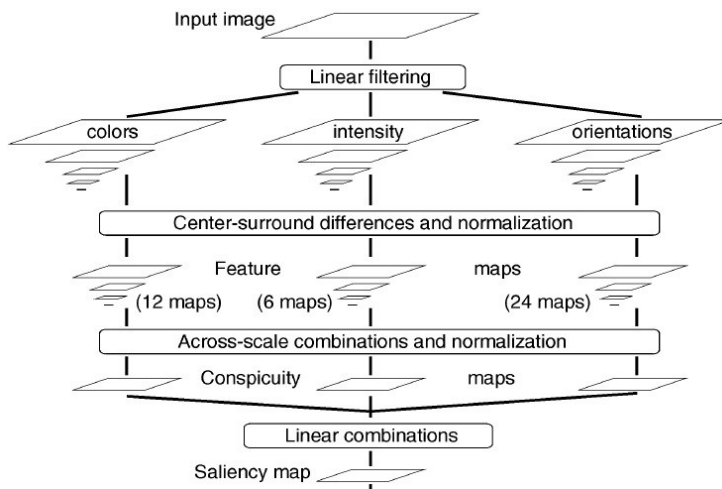


Figure 1: Model of Itti et al. (1998). Three stages: center-surround differences, conspicuity maps, inter-feature fusion into saliency map. Adapted from Itti et al. (1998)

This implementation proved to be the first successful approach of attention computation by providing better predictions of the human gaze than chance or simple descriptors like

entropy. Following this success, most of the computational models of bottom-up attention use the comparison of a central patch to its surroundings as a novelty indicator.

### 3.1.0.2 Context: the whole image or a dataset of images

In this approach, the context which is used to provide a degree of novelty or rarity to image patches is not necessarily the surroundings of the patch, but can be other patches in its neighborhood or even anywhere in the image or an image database. The idea can be divided in two steps. First, local features are computed in parallel from a given image. The second step measures the likeness of a pixel or a neighborhood of pixels to other pixels or neighborhoods within the image. This kind of visual saliency is called "self-resemblance". A good example is shown in Figure 2. The model has two steps. First it proposes to use local regression kernels as features. Second it proposes to use a nonparametric kernel density estimation for such features, which results in a saliency map consisting of local "self-resemblance" measure, indicating likelihood of saliency Seo & Milanfar (2009).

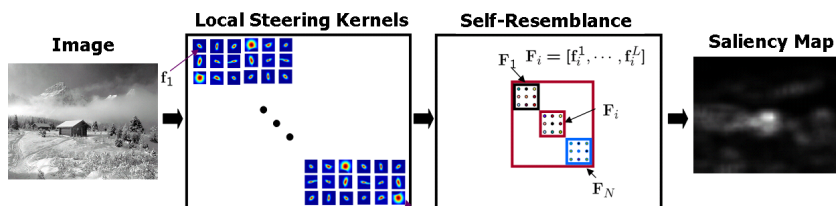


Figure 2: Model of Seo & Milanfar (2009). Patches at different locations in the image are compared. Adapted from Seo & Milanfar (2009).

Mancas (2009) and Riche et al. (2013) focus on the entire image. These models are designed to detect saliency in the areas which are globally rare and locally contrasted. After a feature extraction step, both local contrast and global rarity of pixels are taken into account to compute a saliency map. An example of the difference between locally contrasted features and globally rare is given in Figure 3. On the left there is the initial image of an apple with a defect in red, the second image shows the fixations predicted by Itti et al. (1998) where the locally contrasted apple edges are well detected while its less contrasted but rare defect is not. The third image shows Mancas, Gosselin & Macq (2007) which detected the apple edges, but also the defect. Finally the rightmost is the mouse tracking result for more than 30 users. Boiman & Irani (2007) look for similar patches and relative positions of these patches in an image database which provide more cues about what should be normal. The use of a database might be view as an introduction of top-down information.

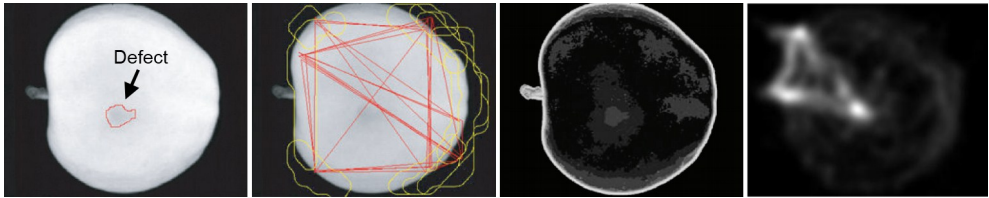


Figure 3: Difference between locally contrasted and globally rare features. Left image: an apple with a defect in red, Second Image: Itti et al. (1998), Third image: Mancas, Gosselin & Macq (2007), Right image: mouse tracking (ground truth).

### 3.1.0.3 Context: a model of normality

This approach is probably less biologically-motivated in most of the other implementations. The context which is used here is a model of what the image should be: if things are not like they should be, this can be surprising, thus attract people attention. Achanta et al. (2009) proposed a very simple attention model (Figure 4): first, the color space is converted from RGB to Lab, second the Euclidean distance is computed between a Gaussian filtered version of the input image and the average Lab vector of the input image. The mean image used is a kind of model of the image statistics: pixels which are far from those statistics are more salient. This model is mainly useful in salient objects detection.

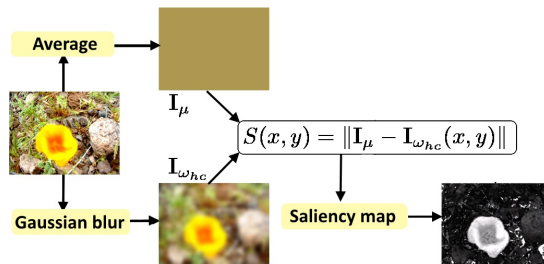


Figure 4: Achanta et al. (2009) uses a model of the mean image. Adapted from Achanta et al. (2009).

Another approach to "normality" can be found in Hou & Zhang (2007), where the authors proposed a spectral model that is independent of any features. As it is known that natural images have a  $\frac{1}{f}$  decreasing Fourier log-spectrum, the difference between the log-spectrum of the image and its smoothed log-spectrum (spectral residual) is reconstructed into a saliency map. Indeed, a smoothed version of the log-spectrum is closer to an  $\frac{1}{f}$  decreasing log-spectrum as small variations are removed. This approach is almost as simple as Achanta et al. (2009) but much more efficient in predicting eye fixations.

More details on still images saliency modeling can be found in the chapters "Bottom-Up Visual Attention for Still Images: A Global View" and "Bottom-up models for still images: a practical review".

### 3.2 Videos

Part of the static models have been extended to video. As shown in Figure 5, it is the case of Seo & Milanfar (2009) where the time dimension is introduced by replacing square spatial patches by 3D spatio-temporal cube patches where the third dimension is the time. Also, Itti's model was generalized with the addition of motion features and flickering to the initial spatial set of features containing luminance, color and orientations.

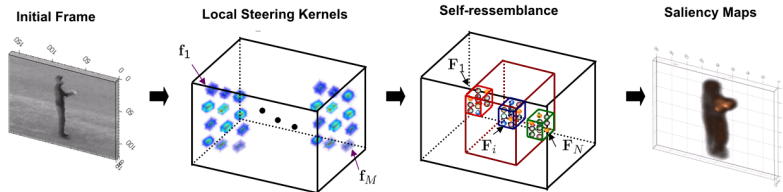


Figure 5: Seo & Milanfar (2009) generalized to video by introducing the spatio-temporal cubes. Adapted from Seo & Milanfar (2009).

Those models mainly show that important motion is well detected. Other models like Mancas, Riche & J. Leroy (2011) has developed a bottom-up saliency map to detect abnormal motion. The proposed method is based on a multi-scale approach using features extracted from optical flow and global rarity quantification to compute bottom-up saliency maps. The model exhibits promising results from a few moving objects to dense crowds with increasing performance (Figure 6). The idea here is to show that motion is most of the time salient but within motion, some motion areas are more interesting than others.

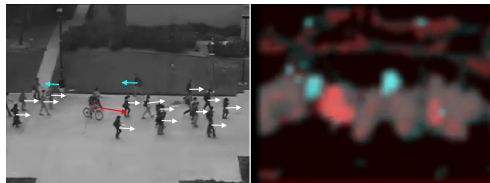


Figure 6: Detection of salient motion compared to the rest of motion. Red motion is salient because of unexpected speed. Cyan motion is salient because of unexpected direction Mancas, Riche & J. Leroy (2011).

More details on video saliency modeling can be found in chapter "Bottom-up models for videos: a practical review".

### 3.3 Extension to 3D

3D saliency modeling is an emerging area of research which was boosted by two main evolutions.

First, the arrival of affordable RGB-D cameras which provide both classical RGB images and a depth map describing pixels distance from the camera. In terms of computational attention this depth information is very important. For example, in all models released

up to now, movement perpendicular to the plane of the camera could not be taken into account while now it is directly available in the depth map. Those cameras (as MS Kinect for example) provide a whole set of new features to the community through the depth map but also through the available point cloud and its 3D geometric features (surface normals, curvature, compactness, convexity, ...).

The second event is the arrival of 3D printers which democratized the 3D models used to print objects. 3D models are more easily available and libraries like PCL ? can handle 3D point clouds, convert formats and compute features from those point clouds.

Most of the 3D saliency models are extensions of still images models. Some use the 3D meshes based on Itti's approach, others just add the depth as an additional feature while recent models are based on the use of point clouds. More details can be found in the chapter "Towards 3D visual saliency modelling".

As 3D saliency models are mainly extensions of 2D models, depending on the extended model, the different features can be taken into account locally and/or globally on the 3D objects.

### **3.4 Audio Signals**

There are very few auditory attention models compared to visual attention models. One of the main issues is that it is not easy to find easy ground truth in the audio domain (contrary to eye-tracking for visual attention). Also, the audio modality taken alone is much less informative on the scene than the visual modality. However, we can classify existing models into different categories.

The first one represents the local context for audio signals. As shown in Figure 7, Kayser et al. (2005) computes auditory saliency maps based on Itti's visual model (1998). First, the sound wave is converted to a time-frequency representation ("intensity image"). Then three auditory features are extracted on different scales and in parallel (intensity, frequency contrast, and temporal contrast). For each feature, the maps obtained at different scales are compared using a center-surround mechanism and normalized. The center-surround maps are fused across scales achieving saliency maps for individual features. Finally, a linear combination builds the saliency map which is then reduced to one dimension to be able to fit on the one-dimensional audio signal.

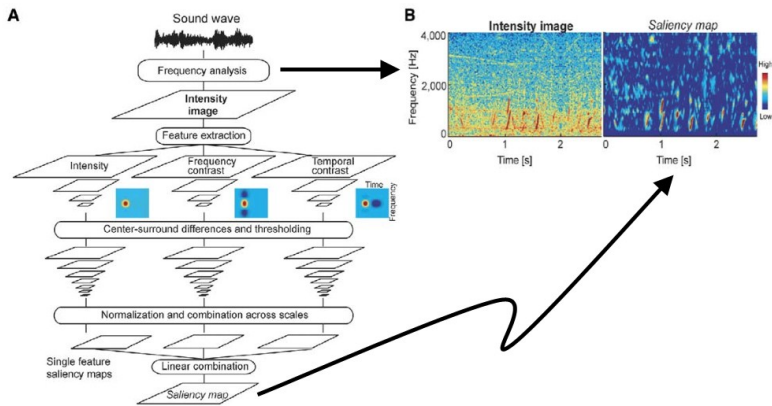


Figure 7: Kayser et al. (2005) audio saliency model inspired from Itti. Adapted from Kayser et al. (2005).

Another approach to compute auditory saliency map is based on following the well-established approach of Bayesian Surprise in computer vision (Itti & Baldi (2006)). An auditory surprise is introduced to detect acoustically salient events. First, a Short-Time Fourier transform (STFT) is used to calculate the spectrogram. The surprise is computed in the Bayesian framework. This surprise approach represent the "normality" context for audio signals. In the case of audio signal there is no real "global" context as the time dimension has no real boundaries as the spatial dimensions have. A global context will be a long period of time in the past.

Mancas, Couvreur, Gosselin, Macq et al. (2007) directly use as features the amplitude of the STFT and quantifies their rarity compared to a long audio history. The model detects sudden and unexpected changes of audio textures and focus the attention of a surveillance operator to sound segments of interest in audio streams that are monitored.

### 3.5 Mixing video and audio signals

The superior colliculus (SC) is the brain structure which directly communicates with the eye motor command in charge of eyes orientation. Its originality is to integrate information coming from different sensory areas but mainly visual and auditory. The information within the SC (Fig. 8) has a retinotopic representation. Visual information is displayed on the superficial layers and the auditory information on the deeper layers King (2004). Once in the same coordinate system, multi-sensory information will be fused in order to take a decision on the eye movement. The main task of the SC is thus to direct the eyes onto the "important" areas of the surrounding space in terms of both vision and sounds and mix those two modalities.

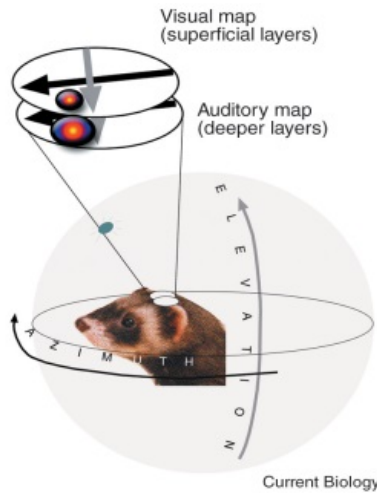


Figure 8: Data fusion within the superior colliculus. Adapted from King (2004)

Some attempts in mixing visual (still, video and 3D) to audio signals saliency showed that the result is much more complex than expected. The final result is NOT the simple addition of visual and audio saliency taken together and it also depends on the scene (natural, social, action, ...) Coutrot & Guyader (2014) King (2009).

Basically, the visual modality seems to take the lead of attention unless the audio event is congruent spatially AND temporally with an image object/action. In this case, the audio has a great impact on the global attention. Given the retinotopic representation in the superior colliculus, a correspondence between the audio and visual location in the same time range is necessary for the fusion to be effective. This task should also be easier in the future as arrays of microphones which also provides the direction of a sound are available together with the RGB and depth map on low-cost sensors as the MS. Kinect.

More details on mixing audio and visual saliency can be found in the chapter "Multimodal saliency models for videos".

#### 4. Saliency models: including top-down information

Top-down is endogenous information and comes from the inner world (information from memory, their related emotional level and also the task-related information). The separation between bottom-up and top-down information is far from being clear. Depending on the viewpoint and the definitions, some notions can be considered as bottom-up or top-down.

One can say that top-down is not involved if the memory/learning is not involved. In this case all the hard-wired features which might be low-level (luminance, color, orientation, motion direction), mid-level (object basic properties as the size, centred-Gaussian as a default context) or high-level (face detection, people detection) which involve specific brain areas but do not need memory and learning are bottom-up.

Top-down involves learning and memory and will deal with specific contexts (e.g. websites, adds, ...), object recognition (face recognition, people recognition, specific animal or object) or a given task coming from inner needs (looking for the keys, ...).

It is thus interesting that face detection can be considered as bottom-up (face feature detection does not necessary need memory and might be located in a specific brain area, the fusiform gyrus McCarthy et al. (1997)) while face recognition is clearly top-down as it directly uses memory to remember a specific person.

In practice, three main families of top-down information can be added to bottom-up attention models.

The first one mainly deals with learned normality in a given context which can come from the experience from the current signal if it is time varying, or from previous experience (tests, databases) for still images.

The second approach is about task modeling which can either use object recognition-related techniques or which can model the usual location of those objects of interest.

The third one use learning to extract both bottom-up and top-down information from eye-tracking results on a dataset of images.

#### 4.1 Top-down as learned normality: attending unusual events

Concerning still images, the "normal" gaze behavior can be learned from the "mean observer". Eye-tracking techniques can be used on several users, and the average of their gaze on a set of natural images can be computed. This was achieved by several authors as it can be seen on Figure 9. Bruce and Judd et al. (2009b) used eye-trackers while Mancas (2007) used mouse-tracking techniques to compute this mean observer. In all cases, it seems clear that, for natural images, the eye gaze is attracted by the center of the images. This information is not top-down as it is generic enough not to be learned.

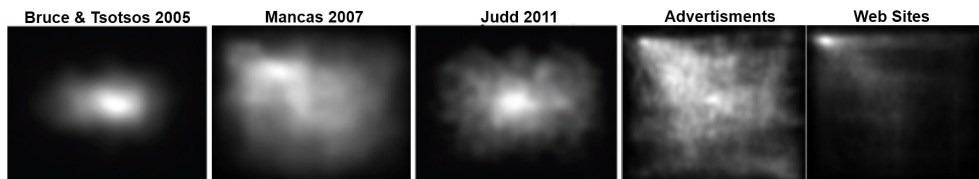


Figure 9: Three models of the mean observer for natural images on the left. The two right images: model of the mean observer on a set of advertising and websites images.

This observation seems logical as natural images are taken using cameras and the photographer will naturally tend to locate the objects of interest in the center of the picture. Another point is that the objects in the center of the visual field are the ones one might interact with, they are then more important than the others.

This observation for natural images is very different from more specific images which use a priori knowledge and which are top-down. Mancas (2009) showed using mouse tracking that gaze density is very different on a set of advertisements and on a set of websites as displayed on Figure 9 on the two right images. This is partly due to a priori knowledge that people have about those images. For example, when viewing a website, the upper part has high chance to contain the logo and title, while the left part should contain the menu. During images or video viewing, the default template is the one of natural images with a high weight on the center of the image. If supplemental knowledge is known about the image, the top-down information will modify the mean behavior towards the optimized gaze density. Those top-down maps can highly influence the bottom-up saliency map but

this influence is variable. In Mancas (2009) it appears that top-down information seems more important in the case of websites, than advertisements and natural images. Other kinds of models can be learned from videos, especially if the camera is still. It is possible to accumulate motion patterns for each extracted feature which provides a model of normality. As an example, after a given period of observation, one can say: here moving objects are generally fast (first feature: speed) and going from left to right (second feature: direction). If an object, at the same location is slow and/or going from right to left, this is surprising given what was previously learned from the scene, thus attention will be directed to this object. This kind of considerations can be found in Mancas & Gosselin (2010). It is possible to go further and to have different cyclic models in time. In a metro station, for example, normal people behavior when a train arrives in the station is different from the one during the waiting period in terms of people direction, speed, density . . . In the literature (mainly in video surveillance) the variations in time of the normality models is learned through HMMs (Hidden Markov Models) Jouneau & Carincotte (2011).

For 3D signals, another information is the proximity of objects. For natural images, centered objects also attract our attention because they might be the ones we will interact with as they are in the center of the visual field. In the same way, a close object is more likely to attract attention as it is more likely to be the first that we will have to interact with. In real world the default context is a mix between a centered Gaussian and proximity value: centered close objects are the most important while far objects on the sides the less.

## **4.2 Top-down as a task: attending to objects or their usual position**

While the previous section dealt with attention attracted by events which lead to situations which are not consistent with the knowledge acquired about the scene, here we focus on a second main top-down cue which is a visual task ("Find the keys!"). This task will also have a huge influence on the way the image is attended and it will imply object recognition ("Recognize the keys") and object usual location ("they could be on the floor, but never on the ceiling").

### **4.2.0.1 Object recognition**

Object recognition can be achieved through classical methods or using points of interest (like SIFT, SURF . . . Bay et al. (2008)) which are somehow related to saliency. Some authors integrated the notion of object recognition into the architecture of their model like Navalpakkam & Itti (2005). They extract the same features as for the bottom-up model, from the object and learn them. This learning step will provide weight modification for the fusion of the conspicuity maps which will lead to the detection of the areas which contain the same feature combination as the learned object.

### **4.2.0.2 Object location**

Another approach is in providing with a higher weight the areas from the image which have a higher probability to contain the searched object. Several authors as Oliva et al. (2003) developed methods to learn objects' location. Vectors of features are extracted from the images and their dimension is reduced by using PCA (Principal Component Analysis). Those vectors are then compared to the ones from a database of images containing the given object. Figure 10 shows the potential people location that has been extracted from the image. This information, combined with bottom-up saliency lead to the selection of a person sitting down on the left part of the image.

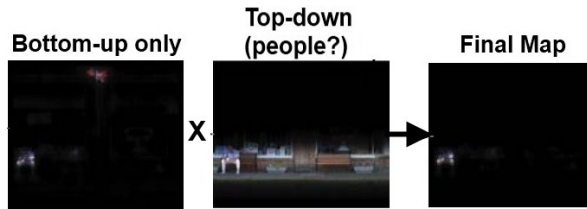


Figure 10: Bottom-up saliency model inhibited by top-down information to select only salient people. Adapted from Oliva et al. (2003).

#### 4.2.0.3 Task, context and learning

Recently, learning the salient features becomes more and more popular: the idea here is not to find the rare regions, but to find an optimal description of those rare regions which are already known from eye-tracking or mouse-tracking ground truth. The learning is based on deep neural networks, sparse coding and pooling based on large images datasets where the regions of interest are known. The most attended regions based on eye-tracking results are used to train classifiers which will extract the main features of these areas.

The use of deep neural networks greatly improved those techniques which are now able to extract meaningful middle and high level features which can describe the best the salient regions Shen & Zhao (2014). Figure 11 shows examples of interesting feature extraction in the context of the training set which was here the MIT dataset Judd et al. (2009a). This dataset contains general purpose images and free viewing, thus specific top-down information is not included. The top row of the figure shows the features after the second layer. One can see mid-level features like corners or textures which naturally pop out from learning. More interestingly higher-level features such as text-like texture, faces, circular objects, man-made structures are learned in the third layer. Those features might be considered top-down even if generic face detection for example can also be considered as bottom-up. These features are than mixed with weights which are again learned from the ground truth into saliency maps.

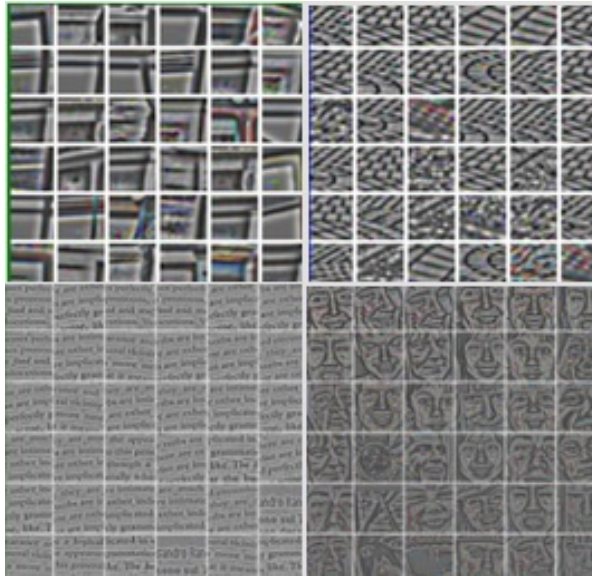


Figure 11: Deep learning of salient features: at the second layer (mid-level features, top row) and at the third layer (high-level features, bottom row). Adapted from Shen & Zhao (2014).

An interesting thing with this kind of approach is that it can be tailored to datasets where specific contexts (like outdoor pictures) or specific tasks (looking for wild animals) are taken into account. In that case the initial feature learning phase could exhibit features which are more related to this context and task and which integrate both bottom-up and top-down information.

## 5. Modeling attention in computer science

In computer science there are two families of models: some are based on feature visibility and others on the concept of saliency maps, the latter approach being the most prolific.

For saliency-based bottom-up attention the idea is the same for all the models: find areas in the image which are the most surprising in a given context. Three main type of contexts can be found: a local one mainly focusing on contrast, a global one quantifying the feature rarity and a normality-based which use normal forms in image or Fourrier space.

Saliency models can be also applied to video, audio and even 3D signals. When mixing audio and visual signals, the influence of the audio seem to be taken into account only if it is congruent with a visual event.

Finally a set of top-down features which can influence the saliency-based models are reviewed. While some of them are in fact bottom-up (centered Gaussian, face detection, ...), others are real top-down features (context-related, object and face recognition, object location).

In the next chapters the saliency-based models will be described for still images, for videos but also for 3D and multimedia models. A strong validation of still and video models is also done to see how effective the models are.

## 6. References

- Achanta, R., Hemami, S., Estrada, F. & Susstrunk, S. (2009). Frequency-tuned Salient Region Detection, *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.  
**URL:** <http://www.cvpr2009.org/>
- Aldoma, A., Marton, Z.-C., Tombari, F., Wohlkinger, W., Potthast, C., Zeisl, B., Rusu, R. B., Gedikli, S. & Vincze, M. (2012). Point cloud library, *IEEE Robotics & Automation Magazine* **1070**(9932/12).
- Bay, H., Ess, A., Tuytelaars, T. & Gool, L. V. (2008). Surf: Speeded up robust features, *Computer Vision and Image Understanding (CVIU)* **110**(3): 346–359.
- Boiman, O. & Irani, M. (2007). Detecting irregularities in images and in video, *International Journal of Computer Vision* **74**(1): 17–31.
- Coutrot, A. & Guyader, N. (2014). How saliency, faces, and sound influence gaze in dynamic social scenes, *Journal of Vision* **14**(8): 5.
- Geisler, W. S. & Cormack, L. (2011). *Chapter 24: Models of Overt Attention, in The Oxford handbook of eye movements*, Oxford University Press.
- Hou, X. & Zhang, L. (2007). Saliency detection: A spectral residual approach, *Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR '07*, pp. 1–8.
- Itti, L. & Baldi, P. F. (2006). Modeling what attracts human gaze over dynamic natural scenes, in L. Harris & M. Jenkin (eds), *Computational Vision in Neural and Machine Systems*, Cambridge University Press, Cambridge, MA.
- Itti, L., Koch, C. & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(11): 1254–1259.
- Jouneau, E. & Carincotte, C. (2011). Particle-based tracking model for automatic anomaly detection, *IEEE Int. Conference on Image Processing (ICIP)*.
- Judd, T., Ehinger, K., Durand, F. & Torralba, A. (2009a). Learning to predict where humans look, *Computer Vision, 2009 IEEE 12th international conference on*, IEEE, pp. 2106–2113.
- Judd, T., Ehinger, K., Durand & Torralba, A. (2009b). Learning to predict where humans look, *IEEE Inter. Conf. on Computer Vision (ICCV)*, pp. 2376–2383.
- Kayser, C., Petkov, C., Lippert, M. & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map, *Curr. Biol.* **15**: 1943–1947.
- King, A. J. (2004). The superior colliculus, *Current Biology* **14**(9): R335–R338.
- King, A. J. (2009). Visual influences on auditory spatial learning, *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**(1515): 331–339.
- Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry., *Hum Neurobiol* **4**(4): 219–227.
- Le Meur, O. & Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition, *Vision research* .
- Legge, Hooven, Klitz, Mansfield & Tjan (2002). Mr.chips 2002: new insights from an idealobserver model of reading, *Vision Research* pp. 2219–2234.
- Mancas, M. (2007). *Computational Attention Towards Attentive Computers*, Presses universitaires de Louvain.
- Mancas, M. (2009). "relative influence of bottom-up and top-down attention, *Attention in Cognitive Systems*, Vol. 5395 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg.

- Mancas, M., Couvreur, L., Gosselin, B., Macq, B. et al. (2007). Computational attention for event detection, *Proc. Fifth Int'l Conf. Computer Vision Systems*.
- Mancas, M. & Gosselin, B. (2010). Dense crowd analysis through bottom-up and top-down attention, *Proc. of the Brain Inspired Cognitive Systems (BICS 2019)*.
- Mancas, M., Gosselin, B. & Macq, B. (2007). Perceptual image representation, *J. Image Video Process.* **2007**: 3–3.  
**URL:** <http://dx.doi.org/10.1155/2007/98181>
- Mancas, M., Pirri, F. & Pizzoli, M. (2011). From saliency to eye gaze: embodied visual selection for a pan-tilt-based robotic head, *Proc. of the 7th Inter. Symp. on Visual Computing (ISVC)*, Las Vegas, USA.
- Mancas, M., Riche, N. & J. Leroy, B. G. (2011). Abnormal motion selection in crowds using bottom-up saliency, *IEEE ICIP*.
- McCarthy, G., Puce, A., Gore, J. C. & Allison, T. (1997). Face-specific processing in the human fusiform gyrus, *Journal of cognitive neuroscience* **9**(5): 605–610.
- Najemnik, J. & Geisler, W. (2005). Optimal eye movement strategies in visual search, *Nature* pp. 387–391.
- Navalpakkam, V. & Itti, L. (2005). Modeling the influence of task on attention, *Vision Research* **45**(2): 205–231.
- Oliva, A., Torralba, A., Castelano, M. & Henderson, J. (2003). Top-down control of visual attention in object detection, *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, Vol. 1, pp. I – 253–6 vol.1.
- Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B. & Dutoit, T. (2013). Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis, *Signal Processing: Image Communication* **28**(6): 642–658.
- Seo, H. J. & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance, *Journal of Vision* **9**(12).  
**URL:** <http://www.journalofvision.org/content/9/12/15.abstract>
- Shen, C. & Zhao, Q. (2014). Learning to predict eye fixations for semantic contents using multi-layer sparse network, *Neurocomputing* **138**: 61–68.
- Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions, *Journal of Vision* **7**.