1 Applications of saliency models

*M. Mancas and O. Le Meur*

1.1 Attention modeling: a very wide spectrum of applications

In engineering, the automatically computed output of a saliency model is called a saliency map. Saliency maps can be computed on still images (chapter “Bottom-up models for still images: a practical review”), videos (chapter “Bottom-up models for videos: a practical review”), audio signal (chapter “Multimodal saliency models for videos”) and even on 3D data (chapter “Towards 3D visual saliency modelling”). Those maps provide for each pixel in an image or video frame, each voxel on a 3D model or at a given time position in an audio file the probability to be attended by human gaze. They include bottom-up information using low-level features directly extracted from the signal or they can also include top-down information related to memory or emotions.

The applications of saliency maps are numerous and they can occur in many domains. For some applications, like in advertising or interfaces optimization, the saliency maps and their analysis are the final goal, while for others (compression, object recognition, ...) saliency maps are not a goal per se, but they act like informational filters to improve the efficiency of other techniques.

While an exhaustive list of saliency maps applications would be difficult to provide and to structure, we propose in this chapter a taxonomy composed of three categories. Different application domains can be split within those three categories (1.1).

Basically, attention maps provide cues about the surprising parts of a signal. A first category of applications directly takes advantage of the detection of those surprising, thus abnormal areas in the signal. We will call this class of applications “Abnormality detection”. Surveillance or events detection are examples of applications domains in this category.

A second category will focus more on the opposite of the first one: as the attention maps provide us with an idea about the surprising parts of the signal, one can deduce where is the normal (homogenous, repetitive, usual, etc...) signal. We will call this category: “Normality modeling”. The main application domains are in signal compression or re-targeting.

Finally, the third application category is related to the surprising parts of the signal
but will go further than a simple detection. This application family will be called “Abnormality processing” and it will need to compare and further process the most salient regions. Domains such as robotics, object retrieval or interfaces optimization can be found in this category.

In the rest of the chapter we will follow this taxonomy which has the advantage to group dozens of applications into only 3 categories (Figure 1.1). The review of the applications of attention modeling has the ambition to be as exhaustive as possible by listing all the known applications. Nevertheless, if examples and references are provided for each application, those references are not necessarily exhaustive.

<table>
<thead>
<tr>
<th>Category</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormality detection</td>
<td>Video/Audio Surveillance (Event detection, Crowd monitoring), Machine vision (defect detection), Medical imaging (pathology detection), ...</td>
</tr>
<tr>
<td>Normality detection</td>
<td>Texture finding, Compression (2D, video, 3D), Re-targetting, Summarization, Computer Graphics (image mosaicking, adaptive rendering), ...</td>
</tr>
<tr>
<td>Abnormality processing</td>
<td>Robotics and CBIR (Image registration and scene reconstruction, Object retrieval, Extraction of object-of-interest), Communication optimization (human-machine interfaces, advertisement, web sites, 3D views optimization, Memorability), ...</td>
</tr>
</tbody>
</table>

**Abbildung 1.1** A 3-classes taxonomy for saliency applications.

In each of the three categories some applications will be listed and others further detailed. Moreover, some of other chapters such as “Multimodal saliency models for videos”, “Towards 3D visual saliency modelling”, “Attentive Content-Based Image Retrieval ”, “Saliency and Attention for Video Quality Assessment ” or “Attentive Robots” are dedicated to specific saliency maps applications.

### 1.2 Applications based on abnormality detection

In this section the main focus is on the first category of applications within our taxonomy: the applications which use the detection of the areas having the higher saliency scores. Those areas correspond with events, defects, pathologies or social-related interactions in real-life applications.

#### 1.2.1 Video surveillance

Video surveillance encompasses a sheer number of applications which can benefit from attention modeling. Here we provide some non exhaustive examples of the use
An interesting European research project, called “SeaRise”, focused on saliency and video surveillance in real-life situation. The main purpose of the project is to develop a trinocular active cognitive vision system called “Smart-Eyes” which first detects abnormal motion using saliency models, and then focuses on the detected area for tracking and categorization of salient events \[1\]. Saliency models were developed to take into account spatial but also motion information \[2\]. Additional long-term information was used to distinguish usual paths or motion from abnormal motion (Figure 1.2).

Abbildung 1.2 Example of the use of long-term information. While the regions A and B on top contain a high amount of motion and are very salient when only short-term information is taken into account, when “usual” paths are taken into account, the region C becomes much more interesting. Adapted from \[2\].

Other authors took into account the concept of “usual motion” either by using accumulation of motion features in given regions which provide a “normality” of the motion in those regions \[3\] or using more complex systems as Hidden Markov Models (HMMs) to predict future normal motion \[4\]. Crowd monitoring is an important topic in surveillance as it is very difficult to detect quickly and in an automatic way a suspect behavior. Some approaches rely on motion rarity \[5\] or motion textures \[6\] while \[7\] worked on human gaze modeling on crowd videos. A benchmark of several models on a dataset which also includes crowd videos is available in \[8\].

While abnormal motion has been mostly used for crowd scenes, some authors like in \[9\] provide models which work on any general scene containing motion (Figure 1.3). An issue with this model is that it is very computationally expensive as it takes into account not only video patches, but also the relative position of those patches. In addition video datasets are needed to learn normal motion before being able to detect...
any suspect behavior.

Abbildung 1.3 Top-left: a frame from a normal motion dataset. Bottom-left: a test frame containing similar motion. Top-right: a frame with motion different from the dataset. Bottom-right: suspicious motion detection and localization. Adapted from [9].

1.2.1.1 A discussion
Attention in video surveillance is a prolific domain with lots of recent references. The search for abnormal events is one of the most important “quests” in the domain. While saliency is currently not a mainstream idea in video surveillance, there are good chances for it to become an important axis of research in the domain in the next years.

1.2.2 Audio surveillance
Audio surveillance is a domain which is much less investigated compared to video surveillance. Nevertheless, microphones could also be added to surveillance cameras. While a human operator still can watch several screens in the same time to monitor data from the cameras, it is almost impossible for him to listen to several audio sources simultaneously. In this latter case, automatic methods for audio event detection are crucial.

For instance, some saliency models were used [10][11] to spot unusual sounds in classical contextual sounds like a gunshot in the middle of a metro station audio ambiance (see Figure 1.4). The idea is to automatically select the camera corresponding with the microphone where the unusual audio event is detected. A normalized en-
vironment adaptive audio attention model based on space and audio clues was also proposed in [12].

Abbildung 1.4 Top: audio signal with reference segmentation. Middle: a space-time representation of the signal (cochleogram). Bottom: Attention peak detection. Adapted from [10].

1.2.2.1 A discussion
Compared to video surveillance, audio surveillance is a smaller investigation field by itself. Moreover, there are very few audio models existing. The use of both audio and video saliency is only achieved in the fields of robotics or social interactions (see the chapters “Multimodal saliency models for videos” and “Attentive Robots”). Nevertheless, attention models have a real interest in the domain. While this application should stay rather limited in a short-term perspective, there is a lot of potential at a long-term perspective.

1.2.3 Machine vision: defect detection

Machine vision is the application of computer vision to industry and manufacturing. One of the applications of machine vision is the automatic inspection of manufactured goods. Machine vision systems perform precise tasks such as counting objects on a conveyor, reading serial numbers, and searching for surface defects. These systems
are preferred for repetitive high-speed tasks and they are sometimes used to complement human’s work which provides a finer perception over a short period of time and which is much more flexible in classification and adaptation to new defects.

In [13], machine vision was applied first to automatic fruit grading. Automatic quality inspection of fresh fruits by machine vision is a challenge not only due to their largely varying physical appearances, but also because of the need to decrease the cost, time and error of inspection introduced by human experts. Figure 1.5 shows the results using a global rarity-based model. As the apple is the main object of the scene, local contrast is not needed and the use of global rarity alone is the best approach once a pre-processing step which eliminates the apple contours and background is achieved. The results in the middle are promising but some regions that are neither contours nor defective have also high attention scores. These “false positives” are mainly due to illumination artifacts or to the presence of stem or calyx regions which are quite similar to defects.

An “atlas” is used to provide the algorithm with images containing healthy apples. If uneven illumination and shadows often occur on healthy apples images, they will be also found in the atlas, thus even if these artifacts are rare within the initial image they will be less rare if the entire atlas is taken into account. On the contrary, the defective skin will be even more rare as it never occurs within the atlas, but only on the test image. The results of the use of healthy apples examples are visible in the right column of Figure 1.5. The defects become in this case more visible and noise due to illumination is eliminated.

Abbildung 1.5 Left: initial apple images. Middle: global rarity-based saliency maps after pre-processing. Right: global rarity saliency maps using an “atlas” or dataset of healthy apples. Adapted from [13].

In [9], in addition to video surveillance, their model can also apply to static images using or not an additional atlas. Figure 1.6 shows the results of defect detection using small atlases (sometimes only one image with no defect is enough to be able to find the defect afterwards).
Abbildung 1.6 Left column: initial images containing defects. Right column: defected located using examples of images with no defects. Adapted from [9].

Saliency models are applied for defect detection on a wide variety of applications such as the semiconductor manufacturing and electronic production [14], metallic surfaces [15] or wafer defects [16] ...

1.2.3.1 A discussion
Defect detection and saliency modeling is a niche application field which is rapidly developing with recent references. As the image-based saliency models become reliable, these field of application should grow in the next years.
1.2.4

Medical imaging: pathology detection

Pathologies, such as tumors, might be considered also as defects with respect the healthy tissues which are considered as normal. In [11], head and neck tumors are taken into account. The two main features characterizing those tumors are that 1) they are located close to the throat and 2) they induce an asymmetry in the neck tissues relative to the throat.

The first feature is taken into account by computing the log-polar image of the Computed Tomography scanner (CT scan) slices (Figure 1.7 - b). The logarithmic approach gives much more importance to the areas which are located around the center point, which is here the throat.

The second feature is taken into account by computing for each gray-level (reduced to only 16 in the paper) the ratio between the pixels on the right-side and left-side of the image. This provides, for each image, 16 coefficients of symmetry which are close to 0 if the gray level is symmetric or close to 1 or -1 if there is an asymmetry towards one or the other side. In Figure 1.7 - c) each column represent the symmetry coefficients for one slice: there are 16 values on the Y axis while the X axis represents the number of slices in the CT scan volume.

A rarity-based attention approach is applied on each line to provide the result in Figure 1.7 - d). Indeed if a gray-level symmetry is rare (unusual) in the context of the other slices of the CT scan volume, this means that the gray level is abnormally asymmetric for the given slice. Abnormal asymmetries are thus detected in Figure 1.7 - d). This result, once projected on the X axis, shows pics for the slices having a high probability of containing tumors (Figure 1.7 - e).

As in previous sections, the use of a set of additional healthy slices called “atlas” can increase the efficiency of the algorithm and it can provide a level of normality. An atlas is used in Figure 1.7 - f) at the left of the vertical line. This atlas provide a reference for a threshold of the final result (Figure 1.7 - h). The approach in [11] is able to detect the slices in a CT scan which might contain tumors.

Many other papers introduce saliency models as promising approaches in improving existing medical imaging techniques. In [17] a rarity-based approach is also used on Magnetic Resonance Imaging (MRI images). In [18], they use saliency to improve medical images registration. In [19] bright lesion detection and classification in color retinal images are based on saliency models. Several saliency models are tested on different image modalities in [20].

1.2.4.1 A discussion

While some years ago the interaction between attention models and medical imaging was sparse, more and more publications make use of saliency models. With the improvement of saliency models for still images and, with the arrival of 3D models of attention, there is a real development potential in the medical domain at middle term.
Abbildung 1.7 a) Initial CT SCAN slice example with annotated tumor and body parts. b) Log-polar representation of image a) centered on the throat. c) 16 gray levels symmetry coefficients (Y axis) for all the slices into the 3D CT Scan volume (X axis). d) rarity-based attention computed on each line (a gray level symmetry coefficient is rare if it is different from the coefficients of the same gray level in the other slices of the 3D CT scan volume). Clear coefficients are the ones which are unusually asymmetric. e) vertical projection of d) which shows pics of abnormality for some slices which might contain tumors. f) same as c) but using also an atlas with healthy CT scan slices (left of the vertical line). g) rarity-based attention computed on f). h) slices possibly containing tumors after thresholding based on the maximum attention level in the atlas where no tumors are present. Adapted from [11].

1.2.5 Expressive gestures and social abilities based on saliency

Gestures are an important part of non verbal communication. They are extensively used in robotics but also in the study of the expressiveness and communication or in Human-Computer Interactions (HCI). Although there are few references in the domain, some papers used saliency models to investigate the role of gestures. In [21] and [22] close gestures are analyzed using dynamic saliency models to show how changes in gestures are more interesting than repetitive gestures. In [23] it is shown that the important moments in gestures which are detected by an attention model are close to what several users provided as manual annotation.

The computation of saliency which is common to several points of view, several images or videos can be used to exhibit the common interesting objects for several people/agents, robots [24]. Human gaze is also a very important social cue which
will instinctively push others to gaze in the same direction. This joint attention and saliency modeling can be used together for robot to robot communication like in [25]. Moreover, saliency models can help in refining the estimation of a viewer gaze (Figure 1.8) by proposing a set of salient areas close to the estimated gaze point [26], [27].

Abbildung 1.8 Left image: experimental setup. Middle image: several possibilities from the face direction system and the important objects detected using a saliency models. Right image: closest salient object to the face direction gaze is selected. Adapted from [27].

In robotics, several references also use saliency models to introduce the notion of gestures like in [28], [29]. The use of saliency models in human-robot interaction is evaluated in [30] and pointing gestures are related to saliency measures in [31]. Research on how to manage the point of focus of a robot using important objects and habituation is described in [32]. A set of interesting projects using attention models in robotics are also described in [33].

For avatar synthesis, a more natural behavior can be inferred by using attention models [34] which will direct the avatar attention on events which are of interest for humans. Other references can be found in this domain like [35] or [36].

1.2.5.1 A discussion
Social signal processing in human-machine or human-human interaction is a growing research field. For the moment the efforts of integrating saliency models are sparse and quite rare due to the fact that the community is not yet aware of what some models can bring and also to the fact that audio-visual and video models are not yet mature enough to act in very complex and dynamic scenes containing a lot of top-down information such as the social scenes. The fact that attention is a filter which brings real-world signals towards conscience and awareness, social interaction extensively uses signals which aim to attract others’ attention. Gesture saliency modeling, multimodal saliency, co-saliency or joint attention are all very important points which should bring a lot to the field in the next years.
1.2.6

**Attention-based computer graphics**

In computer graphics, saliency maps can help in rendering with less details the areas with lower saliency and with more details the areas which are more salient (Figure 1.9). The idea is close to the one of compression, but applied to rendering in computer graphics [37]. Other attention-based rendering techniques can be found in [38]. In addition to rendering, other computer graphics techniques as the meshes for 3D models can also be taken into account by saliency models like in [39].

![Abbildung 1.9 Left: Initial full rendering (see shadows and lights in the back). Middle: Saliency map. Right: Attention-based rendering. Adapted from: [37].](image)

Tone mapping can also take advantage from saliency models. Tone mapping is a technique for mapping a set of colors to another to approximate the appearance of High Dynamic Range (HDR) images in image which has a more limited dynamic range. In [40] they use visual attention for tone mapping on HDR images. In [41], the authors showed that tone mapping has a real influence on human perception which can be a disadvantage (in the case of compression for example) or an advantage (in the case of artistic or computer graphic applications).

In [42] several artistic effects are applied based on a saliency model. The authors show that the use of saliency maps help the main objects to remain less altered and more visible (Figure 1.10). Other references as [43] can be found in this domain which provides automatically interesting perceptually-aware artistic effects.

Other papers related to attentive art deal with saliency-based aesthetics: [44], [45]. The saliency models take as input the segmented image and an order of importance of each segment. This input helps the algorithm to adapt the parameters of color, orientation and sharpness to change the image in order to stick to the proposed regions order. Those new images are considered as more aesthetic. In [46] the authors compute the image aesthetics based on saliency models.
1.2.6.1 **A discussion**
The use of saliency models for artistic purposes or for computer graphics is a new axis of research and one can find a lot of new references in this domain. This direction of research will probably grow especially in 3D rendering.

1.2.7 **Attention-based quality metric**

A great deal of interest and research has been devoted to the design and development of visual quality metrics, leading to the definition of three types of quality metrics: no-reference, reduced-reference and full-reference video quality metrics. A full-reference video quality metrics requires to have the original and the impaired video
sequences. This is obviously a strong limitation in practice. To overcome this limitation, a reduced-reference quality metrics can be used. It requires to get a reduced description of the reference video. This reduced description is compared to a similar description extracted from the impaired video in order to infer a quality score. The more the descriptions are close, the higher the quality is. For some application such as monitoring the quality in a transmission chain, this kind of approach is much more convenient than a full-reference video quality. However, one drawback is that the description extracted from the original video sequence must be encoded without loss. The last solution is to use a no-reference quality metrics for which only the impaired video sequence is available. No-reference quality metrics are less complex but less powerful.

The most relevant quality metrics (IQM (Image Quality Metric) or VQM (Video Quality Metric)) use human visual system properties to predict accurately the quality score that an observer would have given. Hierarchical perceptual decomposition, contrast sensitivity functions, visual masking, etc are the common components of a perceptual metric. These operations simulate different levels of human perception and are now well mastered. In this section, we present quality metrics using visual attention. Assessing the quality of an image or video sequence is a complex process, involving the visual perception as well as the visual attention. It is actually wrong to think that all areas of the picture or video sequence are accurately inspected during a quality assessment task. People preferentially and unconsciously focus on regions of interest. For these types of regions, our sensitivity to distortions might be significantly increased compared to non-salient regions. Even though we are aware of this, very few IQM or VQM approaches take this property into account. Therefore, it seems natural to use saliency maps to give more importance to distortion occurring on salient part. Before describing saliency-based quality metrics, we need to understand more accurately the visual strategy deployed by observers while assessing the quality of an image or video sequences.

1.2.7.1 Eye-movement during a quality task

The use of saliency map in a quality metric raises two main issues.

The first issue deals with the way we compute the saliency map. Indeed, a bottom-up saliency model classically outputs a saliency map indicating where people look. This kind of model makes the assumption that observers watch the scene without performing any task. We then have a paradoxical situation; indeed we want to know where observers look within the scene while they perform a quality task and not when they freely viewed the scene. So the question is whether a bottom-up saliency map could be used to weight distortions or not. To make clear this point, Le Meur et al. [47] investigated the influence of quality assessment task on the visual deployment. Two eye tracking experiments were carried out, one in free-viewing task and the second during a quality-task. The comparison between gaze allocations indicates the quality task has a moderate impact on the visual attention deployment with no significant difference between the two conditions (free-task vs quality-task).

The second issue is related to the presence of strong visual coding impairments
which could disturb the deployment of visual attention in a free-viewing task. In other words, should we compute the saliency map from the original unimpaired image or from the impaired image. In 48, this point was investigated by performing eye-tracking experiments on video sequences with and without video coding artifacts. Observers were asked to watch the video clips without specific instruction. Results indicate that visual deployment is almost the same in both cases. This conclusion is interesting knowing that the distortions of the video clips were estimated as being as visually annoying by a panel of observers.

These two experiments indicate that the use of a bottom-up visual attention model can be used in the context of quality assessment.

1.2.7.2 Saliency-based quality metrics
For most of saliency-based metric 49,50,51,52,53, the use of saliency map consists in modifying the pooling strategy. Quality metrics are composed of several stages. The last one is called the pooling which aims at computing the final quality score from a 2D distortion (or error) map. The degree of saliency of a given pixel can be used as a weight, giving more or less importance to the error occurring on this pixel location.

The difference between these methods concerns the way the weights are defined. As presented in Ninassi et al. 50, different methods to compute the weights can be used:

\[
\begin{align*}
    w_0(x, y, t) &= 1 \\
    w_1(x, y, t) &= SM_n(x, y, t) \\
    w_2(x, y, t) &= 1 + SM_n(x, y, t) \\
    w_3(x, y, t) &= SM(x, y, t) \\
    w_4(x, y, t) &= 1 + SM(x, y, t) \\
    w_5(x, y, t) &= SM_b(x, y, t) \\
    w_6(x, y, t) &= 1 + SM_b(x, y, t)
\end{align*}
\]  

(1.1)

where \( SM(x, y, t) \) is the unnormalized human saliency map, \( SM_n(x, y, t) \) is the human saliency map normalized in the range \([0, 1]\) and \( SM_b(x, y, t) \) is a binarized human saliency map. The weighting function \( w_0 \) is the baseline quality metrics in which the pooling is not modified. The functions \( w_1, w_3 \) and \( w_5 \) give more importance to the salient areas than the others. Indeed, the offset value of 1 in the weighting functions \( w_2, w_4 \) and \( w_6 \) allows us to take into account distortions appearing also on the non salient areas.

The use of saliency map in the pooling stage provides contrasted results. In 50, the use of saliency map does not improve the performance of the quality metric. In the other hand, Akamine and Farias 53 showed that the performance of very simple metrics (PSNR and MSE) has been improved by the use of saliency information. However, for the SSIM metric 54, the saliency does not allow to improve the metric performance. In addition, they showed that the performance improvement depends both on the saliency model used to generate the saliency map and on the distortion type (white noise, JPEG distortions).
1.2.7.3 Quality in stereoscopic 3D images
The conflicting vergence and accommodation cues is widely accepted to be a main cause of visual discomfort in stereoscopic viewing [55]. In addition, fast salient object motion has also been proposed as a cause for viewing discomfort [56]. In both cases, the use of efficient saliency maps is very useful in improving the viewing comfort.

Attention models have been first used to assess the viewing discomfort as in [57], [58], [59]. In this case, saliency maps are compared with the disparity (depth) maps to provide objective metrics for discomfort. A second use of saliency models is in the enhancement of the viewer comfort as in [60], [61] where the blurring of the image is done according to the saliency of the areas.

1.2.7.4 A discussion
Many authors working in this field consider that the visual attention is important for assessing the visual quality of images. However, there are still a number of open issues as demonstrated by [50], [53]. New strategies to incorporate visual attention into quality metrics as well as a better understanding of the interactions between saliency and distortion need to be addressed.

The development of the research in 3D stereoscopic viewing comfort opens a new research avenue to saliency models in the near future.

1.3 Applications based on normality detection
In this section we focus on the second category of applications from our taxonomy: applications using the prior knowledge of the locations having the lowest saliency scores. Those areas correspond with repeating and less informative regions, thus they can be easily compressed or cropped for example.

1.3.1 Attention-based texture
In [62], the authors show that saliency models, which are based on the global information of the image, can be related directly to the homogeneity of the texture. The more the image is complex and unique, the less the saliency is high. The natural tendency of saliency models to only focus on important signal while discarding the usual and repetitive one is very useful in the case of image texture. Indeed, low saliency is synonym of highly homogenous textures or colors. More recently, other authors proposed to use an attention model as a regularity metric for textures [63], [64].

1.3.1.1 A discussion
Very few papers deal with saliency models and texture regularity, even if this is a promising research field. Recent publications should make the texture segmentation
community more aware about the potential of saliency models in texture segmentation and feature extraction. One of the applications of texture regularity detection is in image compression which will be further discussed in the following section.

1.3.2
Attention-based compression

Video compression is the process of converting a signal into a format that takes up less storage space or transmission bandwidth. It can thus be considered as a coding scheme that reduces bits of information representing the original signal (audio, images, videos).

Since the late 1990’s techniques based on attention have been introduced in the field of image and video coding [65] [66]. Attention can be used to select the less interesting areas in images or videos and compress them or to transmit the most salient parts first during the data transfer from a server to a client.

The classical compression methods tend to distribute the coding resources evenly in an image. On the contrary, attention-based methods encode visually salient regions with high priority, while treating less interesting regions with low priority (Figure 1.12). The aim of these methods is to achieve compression without significant degradation of perceived quality.

Although there is currently no unified taxonomy, we have divided the methods into indirect and direct methods, the latter being the most commonly studied.

A first family of compression methods can be called “interactive”. Early approaches
relied on eye-tracking devices to monitor human attention focus \cite{66}.

With such devices which are able to follow the focus of the observer, encoding continuously and efficiently the images is natural. Indeed, observers usually do not even notice any degradation of the frames they watch. However, these techniques are neither practical (because of the use of the eye-tracking device) nor general (because they are restricted to a single viewer).

Attempts to automatize this approach by using attention-based methods are very complex as top-down information is very important and if clear salient objects are not present in a frame, people gaze can be very different. Even in the case where progresses in attention modeling are achieved, it is not possible to have a reliable model of human gaze in case where there is no specific salient object in the frame and where the viewers’ gaze has naturally a very high dispersion.

1.3.2.1 **Indirect approaches**

Indirect compression consists of modifying the source image to be coded, while keeping the same coding scheme. Such methods are thus generally driven by a saliency map based methods.

In \cite{68}, a saliency map for each frame of a video sequence is computed and a smoothing filter is applied to all non-salient regions. Smoothing leads to higher spatial correlation, a better prediction efficiency of the encoder, and therefore a reduced bit-rate of the encoded video.

Another method combines both top-down and bottom-up information, using a wavelet decomposition for multiscale analysis \cite{69}. Bit rate gain ranging from 15\% to 64\% for MPEG-1 videos and from 10.4\% to 28.3\% for MPEG-4 are reported.

An indirect approach based on their attention model is proposed by \cite{70}. An anisotropic pre-filtering of the images or frames is achieved keeping highly salient regions with a good resolution, while low-pass filtering the regions with less important details (Figure 1.13).

In \cite{71}, the depth based on the level of blur of the regions in an image is also taken into account: closer areas should be thus less compressed than objects which might be located far from the camera.

The main advantage of indirect approaches is that they are easy to set up because the coding scheme remains the same. The intelligence of the algorithm is applied as a pre-processing step while standard coding algorithms are used afterward. This fact also let people to easily quantify the gain in terms of compression as the main
compression algorithm can be used directly on the image or on the saliency pre-processed image. However, one possible problem is that the degradation of less salient zones can become strong. Selective blurring can lead to artifacts and distortions in low-saliency regions \[72\].

### 1.3.2.2 Direct approaches

Recent works on modeling visual attention (Le Meur, Itti, Parkhurst, Chauvin \ldots) paved the way to efficient compression applications that modify the heart of the coding scheme to enhance the perceived quality. In this case some modifications to the saliency map are generally necessary to dedicate it directly to coding. The saliency maps will not only be used in the pre-processing step, but also in the entire compression algorithm.

An extension of \[68\], uses a similar neurobiological model of visual attention to generate a saliency map \[72\]. The most salient locations are used to generate a so-called guidance map. The latter is used to guide the bit allocation through quantization parameter (QP) tuning by constrained global optimization. Considering its efficiency at achieving compression while preserving visual quality and the general nature of the algorithm, the authors suggest that it might be integrated in general-purpose video codecs. Future work in this direction should include a study of possible artifacts in the low-bit rate regions of the compressed video, which may themselves become salient and attract human attention. Another possible issue pointed out in \[72\] is that the attention model does not always predict accurately where people look at. For example high speed motion increases saliency, but regions with lower motion can attract more attention (e.g., a person running on the sidewalk, while cars are going faster).

Other approaches with lower computational complexity have been investigated, and in particular two methods using the spectrum of the images: the Spectral Residual \[73\] and the Phase spectrum of Quaternion Fourier Transform \[74\]. The goal of both approaches is to suppress spectral elements corresponding to frequently occurring features.

The Phase spectrum of Quaternion Fourier Transform (PQFT) is an extension of the phase spectrum of Fourier transform (PFT) to quaternions incorporating inter-frame motion. The latter method derives from the property of the Fourier transform, that the phase information specifies the location each of the sinusoidal components resides within the image. Thus the locations with less periodicity or less homogeneity in an image create the so-called proto objects in the reconstruction of the image’s phase spectrum, which indicates where the object candidates are located. A multi-resolution
wavelet foveation filter suppressing coefficients corresponding to background is then applied.

These Fourier-based approaches have two main drawbacks linked to the properties of the Fourier transform. First, if an object occupied most of the image, only its boundaries will be detected, unless resampling is used (at the expense of a blurring of the boundaries). Second, an image with a smooth object in front of a textured background will lead to the background being detected (saliency reversal).

Using the bit allocation model of [72], a scheme for attention video compression has been suggested by [75]. This method is based on learning feature integration algorithm, with a Relevance Vector Machine architecture, incorporating visual saliency propagation (using motion vectors), to save computational time. This architecture is based on thresholding of mutual information between successive frames for flagging frames requiring recomputation of saliency.

Recently, attention-based image compression patents like [76] has been accepted, which also show that compression algorithms are more and more efficient in real-life applications and become close to reach the market.

1.3.2.3  A discussion
For images and video, the expectations from the saliency models were very high. In a first step not all these expectations were met. As the saliency models are not perfect and the classical compression already includes some cognitive elements, the compression factor given the information quality decrease is not optimal. Current saliency-based compression algorithms are mainly suitable for applications where the too high compression of some areas (which creates artifacts catching human attention) are not an issue like in video surveillance. Indeed, in video surveillance the perceived quality of background regions is not important if the foreground is not degraded. However, current work shows an enhancement of the techniques from which some become close to market as recent patents like [76] demonstrate.

Future developments in the direction of 3D compression seem very interesting and new research avenues should be shortly opened in that direction. Indeed, a simple MS. Kinect One device records RGB, depth and infra-red images at almost two Gigabytes per second. Devices able to provide 3D images or point clouds all need efficient ways of compression to cope with the huge amount of data they deliver.

1.3.3  Attention-based retargeting
Compression aims in reducing the amount of data in a signal. A usual approach consist in modifying the coding rate, but other approaches can also reduce the amount of data in the signal by cropping or resizing the signal. An obvious idea which drastically compresses an image is of course to decrease its size. This size decrease can be brutal (zoom on a region and the rest of the image is discarded) or softer (the resolution of the context of the region of interest is decreased but not fully discarded). The first approach will of course be more efficient from a compression point of view,
but it will fully discard the context of the regions of interest which can be disturbing. The direct image cropping will be called here “perceptual zoom” while the second approach which still keeps some context information around the region of interest will be called “anisotropic resolution”. Both approaches provide image retargeting. Retargeting is the process of resizing images while minimizing visual distortion and keeping at best the salient content.

1.3.3.1 Spatio-temporal visual data repurposing: perceptual zoom

Human beings are naturally able to perceive interesting areas of an image. Zooming in images should therefore focus on such regions of interest.

Images manipulation programs provide tools to manually draw these rectangles of interest, but the process can be automated with the help of attention algorithms. Interestingly, such techniques can also be used for real time spatio-temporal images broadcast [77].

Figure 1.14 shows several perceptual zooms depending on a parameter which will threshold the smoothed saliency map from [78].

Abbildung 1.14 Example of images along with rectangles providing different attention-based automatic zooms. After a saliency map [78] is computed and low-pass filtered, several threshold values are used to extract the bounding boxes of the more interesting areas. Depending on this threshold, the zoom is more or less precise/important.

The authors in [79] use Itti algorithm to compute the saliency map [80], that serves as a basis to automatically delineate a rectangular cropping window. A fast greedy algorithm was developed to optimize the window, that has to encompass most of the saliency while remaining sufficiently small.

The Self-Adaptive Image Cropping for Small Displays [81] is based on an Itti and Koch bottom-up attention algorithm but also on top-down considerations as face detection, skin color ... According to a given threshold, the region is either kept or eliminated.

In [82], the authors start by segmenting the image into several regions, for which saliency is calculated to provide a global saliency map. The regions are classified according to their attractiveness, which allows to present image regions on small size screens and to browse in big size images.

A completely automatic solution to create thumbnails according to the saliency distribution or the cover rate is presented by [83]. The size of the thumbnail can be fixed and centered on the saliency map global maximum or adapted to certain
parameters such as the saliency distribution. The gaze fixation predicted by a Winner-Take-All algorithm can thus be used and the search for the thumbnail location ends when a given percentage of the total image saliency is reached.

An algorithm proposed in [84] starts by adaptively partition the input image into number of strips according to the combined saliency map, which contains both gradient information and visual saliency to measure significant regions and is also used to guide the sampling process when scaling image strips.

A video retargeting method based on a spatio-temporal saliency model is described in [85]. Based on a spatio-temporal saliency map, a salient object detection method is used to locate salient object regions in the video. Finally, cropping and uniform scaling operations are performed on the basis of salient object regions to generate the retargeted video.

A hybrid framework of video retargeting with a domain enhanced spatial–temporal grid optimization can be found in [86]. First, they combine visual attention with higher level features. Second, they build a semantic importance map representing the spatial importance and temporal continuity, which is incorporated with a 3D rectilinear grid scaleplate to map frames to a target display, thereby keeping the aspect ratio of semantically salient objects as well as the perceptual coherency.

The methods of intelligent perceptual zooming based on saliency algorithms become more and more interesting with the advances in saliency maps computation in terms of both real-time and spatio-temporal cues integration. Even big companies as Google [87] become more and more involved in developing applications based on perceptual zooms. The idea is to generalize the perceptual zoom for images and videos and keep the temporal coherence of the zoomed image even in case of objects of interest which might brutally appear in the image far from the previous zoom area.

#### 1.3.3.2 Spatio-temporal resolution decrease for uninteresting regions: anisotropic resolution

Perceptual zoom does not always preserve the image structure. For example, Figure 1.14 shows that the smallest zoom on the left image only comprises part of the castle, which is likely to attract attention. In this case the zoom loses the structure and context of the original image. To keep the image structure when retargeting two methods are described in this section: warping and seam carving. These methods may cause non-linear visual distortions on several regions of the image ([88]), but they provide enough contextual information to let the viewer understand the main structures. When adapted to video, those techniques are also easier to stabilize as the context is more present than for the perceptual zoom.

**Warping**  Warping is an operation that maps a position in a source image to a position in a target image by a spatial transformation. This transformation could be a simple scaling transformation [89].

Non-homogeneous content-driven video-retargeting [90] proposes a real-time retargeting algorithm for video. Spatial saliency, face detection and motion detection are computed to provide a saliency matrix. An optimized mapping is computed with
a sparse linear system of equations which takes into account some constraints such as importance modeling, boundary substitutions, spatial and time continuity.

A retargeting method based on global energy optimization is detailed in [91]. Some content-aware methods only preserve high energy pixels, which only achieve local optimization. They calculate an energy map which depends on the static saliency and face detection. The optimal new size of each pixel is computed by linear programming.

The same group proposes a retargeting approach that combines an uniform sampling and a structure-aware image representation [92]. The image is decomposed with a curve-edge grid, which is determined by using a carving graph such that each image pixel corresponds to a vertex in the graph. A weight is assigned to each vertex connection (only vertical direction) which depends on an energy map using saliency region and face detection. The paths with high connection weight sums in the graph are selected and the target image is generated by uniformly sampling the pixels within the grids.

Abbildung 1.15 The original image (left) is deformed by a grid mesh structure to be fit in the required size (right). The scaling and stretching depend on the gradient and saliency map. (adapted from: http://graphics.csie.ncku.edu.tw/Image_Resizing/)

A warping method which uses the grid mesh of quads to retarget the images (figure 1.15) is defined in [93]. The method determines an optimal scaling factor for regions with high content importance as well as for regions with homogeneous content which will be distorted. A significance map is computed based on the product of the gradient and the saliency measure which characterizes the visual attractiveness of each pixel. The regions are deformed according to the significance map. A global optimizing process is used repetitively to minimize the quad deformation and grid bending.

Another approach is a patch-based retargeting scheme [94] with an extended significance measurement to preserve shapes of both visually salient objects and structure lines while minimizing visual distortions. In the proposed scheme, a similarity transformation constraint is used to force visually salient content to undergo as-rigid-as-possible deformation, while an optimization process is performed to smoothly propagate distortions. These processes enable to yield more pleasing content-aware warping and retargeting.

Seam Carving Seam carving [95] allows to retarget the image thanks to an energy function which defines the pixels importance. The most classical energy function is the gradient map, but other functions can be used such as entropy, histograms of ori-
ented gradients, or saliency maps [96]. Low-energy pixels are connected together to make a seam path. The seam paths cross vertically and horizontally the image and are removed. Dynamic programming is used to calculate the optimal seams. The image is readjusted by shifting pixels to compensate the disappeared seams. The process is repeated as often as required to reach the expected sizes.

Abbildung 1.16 The original images (A and B) and for each one seams removal (vertical seams for A and horizontal seams for B) using gradient (top-row) and using a saliency map (bottom row). Adapted from: http://cilabs.kaist.ac.kr

Figure 1.16 shows an example of seam carving: the original images (A and B) are reduced either by discarding vertical or horizontal seams. On the top row, the classical gradient is used as the energy map, while saliency maps of [97] are used for the bottom row. Depending on the energy map which is used distances, shapes as well as aspect ratio distortions can cause anisotropic stretching [77]. Even if saliency maps most of the time work better than simple gradient, they are not perfect and the results can be very different depending on the method used.

For spatio-temporal images, [98] propose to remove 2D seam manifolds from 3D space-time volumes by replacing dynamic programming method with graph cuts optimization to find the optimal seams. A forward energy criterion is presented which improves the visual quality of the retargeted images. Indeed, the seam carving method removes the seams with the least amount of energy, and might introduce energy into the images due to previously non-adjacent neighbors becoming neighbors. The optimal seam is the one which introduces a minimum amount of energy.

A saliency-based spatio-temporal seam-carving approach with much better spatio-temporal continuity than [98] is proposed by [99]. The spatial saliency maps are computed on each frame but they are averaged over and history of frames in order to smooth the maps from a temporal point of view. Moreover, the seams are temporally discontinuous providing only the appearance of a continuous seam which helps in keeping both spatial and temporal coherence.

In [100], the authors describe a saliency map which takes more into account the context and proposes to apply it to seam carving. The idea leads to good results as shown in Figure 1.17.

In [101] the authors used attention algorithms for videos retargeting based on seam carving. An efficient spatio-temporal grouping is done to determine the temporal rate of reduction depending on the content, to suppress groups of isolated seams, to identify spatio-temporal groups of seams and to approximate by constant segments the number of seams for each group, while keeping the total sum of seams constant.
Abbildung 1.17 Left: original images, Middle: saliency maps, Right: retargeted images. Adapted from: [100].

Problems of geometric distortion, anachronism and length of summary have been also addressed.

Interestingly, recent papers as [102] propose to mix seam carving and warping techniques. Firstly, based on the importance partition with the saliency map, they apply a weighted seam carving approach to make the seams distributed dispersedly in the important regions. Then they propose Content Aware Image Distance (CAID) to assess the deformation caused by removing seams. The weighted seam carving will stop with a fixed threshold to guarantee little visual image quality degradation. Finally, the grid based warping is utilized to achieve the final size with a global optimization model, since warping tends to avoid discontinuity artifacts of important region and typically make the distortion distribution of unimportant region more coherently.

1.3.3.3 Attention-based summarization

Summarization of images or videos is a term which is similar to retargeting. It might be based on cropping (closer to the first retargeting family) [103]. It might also be closer to the second family based on carving as in [104]. The main purpose is to provide a relevant summary of a video or an image.

In [105] the authors used video summarization to provide a mashup of several videos into a unique pleasant video containing the important sequences of all the concatenated videos. This approach shows the possible extension of the notion of summarization from a single image or video document to a whole archive of documents. This application has common points with section [1.2.6] and image mosaics. In [106] the authors proposed to make an intelligent collage based on saliency maps (Figure 1.18). This approach also led to a patent [107] on this topic.
1.3.3.4 A discussion

While the use of saliency maps for classical compression does not bring the expected improvements when using the nowadays state of the art models, the retargeting methods (perceptual zooms, warping or seam carving) can benefit a lot from saliency methods. Automatic attention computation is based on the use of context (contrast, rarity, surprise in a given spatial and/or temporal context). These models can highly improve retargeting methods and preserve objects of interest while also keeping the minimum of context information. Industrial applications begin to rise with the enhancement of the saliency models both in terms of accuracy and computational efficiency.

1.3.4 Watermarking and security

Watermarking consists of hiding data in an image with a minimal visual altering of this image.

An idea is to hide data in the most interesting areas of the image which are computed based on a saliency model [108] to get more robust watermarks. Indeed, the high-frequencies of the watermark are less easy to notice if hidden within other high-frequency areas which is generally the case for salient regions.

Another idea is, on the opposite, to hide data in the less salient regions as those regions have a lower probability to be noticed [109]. This assumption is true if the background is cluttered (grass, . . . ) as watermarks are easier to hide in high frequency areas. The saliency-based watermarking is capable of hiding lower injected-watermark energy onto more sensitive regions and higher energy onto the less perceptually significant regions in the image [110]. The use of saliency models helps to get better visual quality of the watermarked image and an improved robustness of the watermark.
1.3.4.1 A discussion
Watermarking only uses the saliency model as a filter to select the areas where information should be inserted. They could be inserted depending on the image entropy in the most salient or less salient areas.

1.3.5 Attention-based advertising insertion

In [111] one can find an interesting description of attention-based advertising insertion. The first approach is called linear advertising, while the second is called non-linear.

Linear advertising will insert content-related ad clips into less intrusive temporal positions. In [112], the authors propose a two-step approach. The first step aims in selecting an ad which is related to the current content. In the paper this was done using text mining. The second step goal is to find the moment which has low spatio-temporal saliency to insert the ad in a less intrusive way. Figure 1.19 shows the main scheme of the system.

Abbildung 1.19 Linear ads insertion at the less salient moments. Adapted from: [112].

Another approach is the non-linear one [113]. In this case, there are also two steps. The first consists in finding the right location in the frame where the ad should be inserted. This step uses the saliency map to locate an area close the most interesting one. The second step will produce color harmonization of the ad to be less intrusive when projected onto the frame. Finally, the harmonized ad is projected close to the most interesting area. This will lead to a very noticeable ad (close to salient regions) while the color harmonization reduces its intrusiveness.
1.3.5.1 **A discussion**

The way an ad is shown to viewers is of utmost importance in the way they perceive and remember the message. The linear and non linear approaches might find their way in the audiovisual production and broadcasting if the attention models become more efficient and the systems can really be real-time.

1.4 **Applications based on abnormality processing**

The third category of our attention-based applications taxonomy concerns abnormality processing. Some applications go further than the use of the simple detection of the areas of interest. They use comparisons between the areas, relative positioning and other operations on the saliency maps. Application domain such as robotics or advertisement highly benefit from this category of applications.

1.4.1 **Attention-based robotics, object recognition and registration**

Robotics is a very large domain of application with various needs. As robotics aim at mimicking human reactions, the field aggregates several techniques of which some can be used in other domains as robotics. We describe here rapidly three research axis where robots can take advantage from saliency models: 1) image registration and landmarks extraction, 2) object recognition, and 3) robots action guidance. We only provide a rapid view about those research actions here as they will be explained more in detail in the chapter on “Attentive Robots”.

1.4.1.1 **Image registration and landmarks**

An important need of a robot is to know where it is located. For this aim, the robot can use the data from its sensors to find landmarks (salient features extraction) and register images taken at different times (salient features comparison) to build a model of the scene. The general process of real-time building of a view of the scene is called Simultaneous Localization and Mapping (SLAM). The use of RGB cameras using or not the depth information is called visual SLAM. Saliency models can help a lot in the extraction of more stable landmarks from images which can be more robustly compared [114]. A detailed review of attentive SLAM can be found in the “Attentive Robots” chapter.

Saliency maps are also used in other domains as for medical image registration [18], lunar images and crater impacts detection [115] or on 3D object registration [116]. All those techniques imply first the computation of saliency maps, but the results are not used directly: they need to be further processed (like extraction of regions of interest and their comparison).
1.4.1.2 Object recognition

Another important need of robots after they establish the scene, is to recognize the objects which are present in this scene and which might be interesting to interact with. To recognize objects two steps are needed. First of all, the robot needs to detect the object in a scene. For this goal saliency models can help a lot as they can provide information about proto-objects \cite{117} or areas objectness \cite{118}. For more details on proto-objects and objects detection, see the chapter “Object-based attention: biological and computational perspectives”.

Once objects are detected, they need to be recognized. In this area the main approach is to 1) extract features (SIFT, SURF or any others) from the object 2) filter the features based on a saliency map 3) perform the recognition based on a classifier (such as a SVM or others). Papers like \cite{119} or \cite{120} apply this technique which let a computer drastically decrease the number of needed keypoints to perform the object recognition. Further details can be found in the chapter “Attentive Content Based Image Retrieval” which is focusing on this approach.

Another approach was used in \cite{121} or \cite{122}. Here the features which are mostly present in the searched object and not present in the surroundings are learned and this learning phase provides a new set of weights for bottom-up attention models. In this way, the features which are the most discriminant in the searched object will get the higher response in the final saliency map. The bottom-up model is in that way tuned by top-down information on the discriminant features learned from the searched object.

A third approach can be found in \cite{123} where relative position of salient points (called cliques) are used for image recognition. More details on this approach can be found in the chapter “Bottom-Up Visual Attention for Still Images: A Global View”.

1.4.1.3 Action guidance

Once robots know where they are (attentive visual SLAM) and they also recognize objects around them (attentive object recognition), they need to decide what to do next. One of the decisions they need to make is to know where to look next and this decision is obviously taken based on visual attention. Several robots implement multi-modal attention like the iCub robot in Figure 1.20. They combine visual and audio saliency in an ego-sphere and this is used to point the gaze on the next location. More details about robots and gaze can be found in the “Attentive robots” chapter and an interesting survey on attention for interactive robots can be found in \cite{124}.

The social interactions also include gestures as the pointing gestures which are important top-down factors. The use of gesture direction is used in \cite{126}, \cite{29} to detect the object of utmost interest and to learn to the system where to look.

Robots are embodied agents, but other agents like virtual agents can implement attention models \cite{34}. In \cite{127} an attentive system based on high-level features (people skeleton extracted using a RGB-D camera) is described. More details on this approach can be found in the chapter “3D Saliency”.

1.4.1.4  A discussion
Robotics, especially humanoid robotics is a very complete field of research. Even if we restrain the domain to electrical engineering and computer science, there are still an impressive list of topics necessary to build a convincing robot. Here we focused on the use of attention models in robots and related fields which give us three main axis of research. The advances in those topics are huge, but still it is difficult to have a realistic social robot capable of naturally interact and adapt to novel unusual situations. This is one of the big challenges to which attention modeling might bring a solution in the future years.

1.4.2  Attention-based marketing and communication optimization
Marketing optimization can be applied to a large amount of practical cases such as: web sites, advertisement, product placement in supermarkets, signage, 2D and 3D objects placement in galleries, ldots All this application cases can benefit from attention maps themselves, but also from regions of interest comparison and further analysis of the attention maps. Moreover, attention only tells if people will notice the important message in an ad, but not if they remember it. Thus the “memorability” of an image is an important topic where attention models can help.

This section is structured in three subsections: we investigate the use of saliency maps in 1) web sites and ads optimization, 2) images memorability and 3) 3D objects best viewpoint calculation.

Abbildung 1.20 iCUB robot head. The robot implements a multi-modal saliency system. Adapted from: [125].
1.4.2.1 Attention-based web sites and advertisement optimization

Among the different applications of automatic saliency computation, the marketing and communication optimization is probably one of the closest to market. As it is possible to predict an image attention map, which is a map of the probability that people attend each pixel of the image, it is possible to predict where people are likely to look on a marketing material like an advertisement or a website. Attracting customer attention is the first step of the process of attracting people interest, induce desire and need for the product and finally push the client to action as described in the AIDA pyramid [128].

It is important to stress the fact that attention alone is not enough to push a potential client to action, but at least it is a key step towards this goal.

There are already two main techniques which are able to provide information about people attention on marketing material. The first one uses eye-tracking studies on marketing material like in [129]. This approach is very accurate as the precise gaze location of the users on a website/advertisement are measured. The drawbacks of this approach are in the time needed to conduct the study, the price and the fact that only finished or almost finished documents can be tested. Another drawback is that long fixations do not mean that this area is necessary very salient: it might only mean that it is difficult to understand and people spend a lot of time in trying to figure out what this area is about.

Figure 1.21 shows an example of heatmap computed from the average eye-tracking data from several users on a web site page.

Another technique uses mouse tracking and is also used on marketing material like in [130] or [131]. There are two ways of using mouse tracking: either with no special indication like in [130] or by telling people to locate their mouse where they look [132]. The second version is precise but only used for research purposes and
Abbildung 1.22 Mouse-tracking on a website: gaze heatmap overlay on the initial image from PicNet. Extracted from [130].

provides a quite good approximation of the user gaze [133] (more than 80% of the eye-tracking). In the case of [130], the accuracy is much lower. The advantage of mouse-tracking techniques is that they are cheaper than eye-tracking and more users are available via websites while eye-tracking needs the user to be physically present in front of an eye-tracking device. The study time varies but could be a little shorter than by using eye-tracking.

Figure 1.22 shows an example of heatmap computed from the average mouse-tracking data from several users on a web site page.

If no special indication is provided like in [130], the result is less accurate than eye-tracking which is due to the fact that the mouse pointer never exactly focus on the object of interest at least for visibility reasons. Also the mouse motion does not always follow the eye motion. However, the mass effect of the number of users which can be much higher than in the case of eye-tracking can partly compensate this issue.

Finally the predictive method which is the main focus here uses automatic saliency maps. This approach is much faster than eye-tracking tests (seconds versus days) and also much cheaper (around 10 times cheaper). The prices are in the same range as the ones of mouse tracking. The results are less good than eye-tracking and they are equivalent to mouse tracking following proprietary studies such as for example [134]. The predictive methods can be achieved in real time and used at any time in the creative process: while humans (both using eye or mouse tracking) will be disturbed by unfinished documents, an automatic algorithm will not. This means that eye-tracking or mouse-tracking can mainly be achieved once the document is almost finalized (which might be too late for important changes) while the use of automatic algorithms can be used in real time and provide several feedback loops during the creative process. Another advantage of the predictive method is that it is image-based only and it is possible to screen any kind of website or advertisement including the ones of concurrent companies.

Figure 1.23 shows an example of heatmap automatically computed on a web site page using the automatic saliency maps from Eyequant [135].
As the approach of saliency and marketing is one of the closest to the market, several companies based on the use of saliency in marketing were set up.

Feng-GUI [136] is an Israeli company mainly focusing on web pages and advertising optimization even if the algorithm is also capable to analyze video sequences. Among the bottom-up features they use one can find color, orientation, density and contrast, intensity, size and weight and intersections. The top down features are face detection, text detection and skin detection. They also use context information about the type of the document (Natural image, Website, Billboard, Advertisement) which probably correspond to different probability densities depending on the kind of document as in [132]. The main targeted applications are web pages and advertising optimization even if the algorithm is also capable to analyze video sequences.

AttentionWizzard [137] is a US company mainly focusing on web pages. There are few hints on the used algorithm, but it uses bottom-up features like: color differences, contrast, density, brightness and intensity, edges and intersections, length and width, curves and line orientations. Top-down features include face detection, skin color and text (especially big text) detection.

3M VAS [138] is the only big international player in this field. Very few details are given on the used algorithm, but it is also capable to provide video saliency. The main difference with the others competitors is in the customer segments with a much wider range of possible applications. They provide attention maps for web pages optimization, but also advertisement with static images or videos, packaging or in-store merchandising.

Eyequant [135] is a German company specialized in website optimization. Their algorithm use extensive eye-tracking tests to train the algorithm and make it closer to real eye-tracking for a given task. They also can modify the saliency map if the viewer is involved or not in a task or if he simply goes through the page by modifying their algorithm weights. Finally they provide a cue on “visual clarity” which seems to be related to a study on the image entropy.

All those companies claim around 90% accuracy for the first 3/5 viewing seconds [134]. They base their claim on different comparison between their algorithm and several existing databases using several ROC metrics. They always compare the re-
sults with the maximum ROC score obtained by the human users. Nevertheless, for real-life images and for given tasks and emotion-based communication, this accuracy dramatically drops but still remains usable.

In addition to those four companies, another approach of using saliency models is proposed by a US company called Eye Predict [139]. The main idea is to test a maximum of combination of product catalog and propose the configuration which best optimizes a given product visibility.

1.4.2.2 Predicting Memorability of Pictures
The study of images memorability in computer science is a recent topic [140][141][142][143]. From those first attempts it appears that it is possible to predict the degree of image’s memorability quite well. In this section, we present the concept of memorability of pictures, the relationship between memorability and eye movement and finally the computational models predicting the extent to which a picture is memorable.

Humans have an amazing visual memory. Only a few seconds is enough to memorize an image [144]. However, not all images are equally memorable. Some are very easy to memorize and to recall whereas the memorization task appears to be much more difficult for other pictures. [140] was the first paper to build a large dataset of pictures associated to their own memorability score. The score varies between 0 and 1. 0 indicates that the picture is not memorable at all while 1 indicates the highest score of memorability. The memorability has been quantified by performing a visual memory game. 665 participants were involved in the test to score the memorability of 2222 images. This dataset is freely available on author’s website.

From this large amount of data, authors in [140] investigated the contributions of different factors and envisioned the first computational model for predicting the memorability scores.

An interesting step which followed was the use of the temporal context in memorability: indeed when seeing a lot of desert images, if a single image of forest appears, that one will be very memorable [143].

Memorability and Eye-Movement
An eye tracking experiment was performed in order to investigate whether the memorability of a picture has an influence on our visual deployment [142]. For that, 135 pictures were extracted from the dataset proposed by [140]. They are organized into three classes of memorability (statistically significantly different), each composed of 45 pictures. The first class consists of the most memorable pictures (C1, score 0.82 ± 0.05), the second of typical memorability (C2, score 0.68 ± 0.04) and the third of the least memorable images (C3, score 0.51 ± 0.08).

As visual attention might be a step towards memory and therefore, the image memorability should influence the intrinsic parameters of eye movements such as the duration of visual fixations, the congruency between observers and the saccade lengths. From the collected eye tracking data, the visual behavior of participants is analyzed according to the picture’s memorability. Figure[1.24] illustrates this point. Four pictures are depicted; the first two pictures have a low memorability score whereas this score is
high for the last two pictures. The first one has a memorability score of 0.81 whereas
the second has a memorability score of 0.4. The average fixation durations for these
two pictures are 391 and 278 ms, respectively. The average lengths of saccades are
2.39 and 2.99 degree of visual angle, respectively. In addition, if there is something in
the picture that stands out from the background, the inter-observer congruency should
be higher for memorable pictures.

From the proposed experiment in [142], several conclusions have been drawn. First,
the fixation durations decrease with the degree of memorability of pictures. This trend
is especially noticeable just after the stimuli onset. Fixations are the longest one when
observers watch memorable pictures. A statistically significant difference is found
between fixation durations when the top 20 most memorable and the bottom 20 less
memorable are considered. This difference is confirmed for different viewing times.
These results are important since the duration of fixations reflects the deepness of the
visual processing in the brain [145].

The congruency between observers watching the same stimulus is the second indica-
tor that has been analyzed. It indicates the degree of similarity between observers’ fi-
xations. A high congruency would mean that observers look at the same regions of the
stimuli. Otherwise, the congruency is low. Generally the consistency between visual
fixations of different participants is high just after the stimulus onset but progressive-
ly decreases over time [146]. To quantify inter-observer congruency, two metrics can
be used: ROC [147] or a bounding box approach [148]. The former is a parametric
approach contrary to the latter. The main drawback of the bounding box approach is
its high sensitivity of outliers. A value of 1 indicates a perfect similarity between ob-
servers whereas the value 0 corresponds to the minimal congruency. Results of [142]
indicate that the congruency is highest on the class C1 (especially after the stimuli
onset (first two fixations)). The difference between congruency of class C1 and C2
is not statistically significant. However, there is a significant difference between con-
gruency of pictures belonging to C1 and C3. This indicates that pictures of classes
C1 and C2 are composed of more salient areas which would attract more observer’s
attention.

These results show that memorability and attention are linked. It would then be rea-
sonable to use attention-based visual features to predict the memorability of pictures.

**Memorability and saliency models** As mentioned earlier, Isola et al. [140] we-
re the first to propose a computation model for predicting the memorability score
of an image. Authors used a mixture of several low-level features which have been
automatically extracted. A support vector regression classifier were used to infer the
relationship between those features and the memorability. The best result was achie-
ved by mixing together GIST [149], SIFT [150], HOG [151], SSIM [152] and pixel
histograms (PH).

In [142], the authors proposed to go one step further by considering saliency-based
features, namely the saliency coverage and the visibility of structure. The saliency co-
verage which describes the spatial computational saliency density distribution could
be approximated by the mean of the normalized saliency maps (computed by the
RARE model [153]). A low coverage would indicate that there is at least one salient
Abbildung 1.24 (a) original pictures; (b) fixation map (a green circle represents the first fixation of observers); (c) Saliency map and (d) heat map. From top to bottom, the memorability score is 0.346, 0.346, 0.897 and 0.903, respectively (from a low to high memorability).
region in the image. A high coverage may indicate that there is nothing in the scene visually important as most of the pixels are attended. The second feature related to the visibility of structure is obtained by applying a low-pass filter several times on images with kernels of increasing sizes like in Gaussian pyramids (see [142] for more details). By using saliency-based features, the performance in term of linear correlation increases by 2% while reducing the number of features required to perform the learning (86% less features).

The same year, the work in [140] was extended by [154] who proposed an attention-driven spatial pooling strategy. Instead of considering all the features (SIFT, HOG...) with an equal contribution, the idea is to emphasize features of salient areas. This saliency-based pooling strategy improves the memorability prediction. Two levels of saliency were used: a bottom-up saliency and an object-level saliency. A linear correlation coefficient of 0.47 was obtained.

In [143], the context of the displayed images is studied and the influence of the viewing context is shown. Figure 1.25 shows the difference of memorability score function of the scene categories when images of those categories are shown surrounded of other images from the same category (in blue) or of different categories (in red). The memorability score dramatically increases when an image is shown surrounded of images from other scene categories.

![Abbildung 1.25 Memorability vs. scene categories. In blue, the images are shown in the middle of images of the same category. In Red, the images are shown in the context of other scene categories. Adapted from: [143].](image-url)
1.4.2.3 **Best viewpoint**
With more and more 3D objects which are created, manipulated, sold or even printed, 3D saliency is a very promising future research direction. The main idea is to compute the saliency score of each view of a 3D model: the best viewpoint is the one where the total object saliency is maximized [155]. Mesh saliency was introduced based on adapting to the mesh structure concepts for 2D saliency [39]. The notion of viewpoint and mesh simplification are also related through the use of mesh saliency [156].

As the 3D approach for best viewpoint is quite novel, it is not obvious to validate if the best viewpoint proposed by the algorithm is the one which people would select. The authors in [157] proposed a web-based solution for viewpoint validation based on votes. Those votes are projected to a 3D heatmap which is used to choose the best viewpoint (Figure 1.26).

While the best viewpoint application can be used for computer graphics or even 3D mesh compression, marketing is one of the targets of this research topic: more and more 3D objects are shown even on internet and the question of how to display them in an optimal way is very interesting in marketing.

![Image](image-url)

**Abbildung 1.26** Viewpoint evaluation: a web-based evaluation can be done on a lot of viewers who vote for their best view. A 3D heatmap of their votes can be projected on a sphere around the object and the maximum of this heatmap represents the best view. Adapted from: [157].

1.4.2.4 **A discussion**
The marketing optimization application of the automatic saliency algorithms has a promising future with already existing companies making money from the idea alone. However, even if the results are very promising, there is room for a lot of improvement. More and more top-down information must be added to classical bottom-up attention to make the result fit with more precise user categories.

An issue which might be stressed is the banner blindness [158] which consists in ignoring the areas where the presence of an advertising is detected. Even if those advertisements are ignored, they are actually seen [159] and if one of them is of interest for the user, he will for sure attend to it.

In addition, features linked to images memorability (see Section 1.4.3) will also be taken into account. Indeed, in [160] features related to memorability seems to provide better visibility to advertisements like higher gray level contrast or a smaller number of salient components, with all components close to the center of the creative and the
major component consistent with the rule of third.

Finally, as more and more objects can be represented in 3D to be better visualized or even directly sold as 3D models for 3D printing for example, the use of saliency on those models is a new research avenue. The automatic computation of the best viewpoint can provide interesting insights for 3D objects visualization.

1.4.3
Attention-based focus or symmetry

Saliency maps provide areas of interests or key points. By comparing those key points and their relative position, higher level characteristics of the image can be found. For example in [161] the author shows how symmetry axis can be found using a saliency model (Figure 1.27). Using again the comparison between patches in the image, it is also possible to find the vanishing points [162]. Using again the same approach, the auto-focus of camera can be controlled [163].

Abbildung 1.27 Comparison between different regions of the image and object symmetry detection. Adapted from: [161].

1.4.3.1 A discussion

This work led to numerous applications. Several patents like [164] or [165] which show that the technology becomes mature enough to be integrated in consumer electronics.
1.5 Conclusion

During the last two decades, significant progresses have been made in the area of visual attention. Although that the picture is much clearer, there are still a number of hurdles to overcome. For instance, the eye-tracking data sets used for evaluating the performance of computational models are more or less corrupted by biases. Among them, the central bias, which is the tendency of observers to look near the screen center, is probably the most important [166, 147]. The central bias, which is extremely difficult to cancel or to remove, is a fundamental flaw which can significantly undermine conclusions of some studies and model’s performance. Also other evaluation frameworks like the ones using segmented objects and even application-driven validation [167] will improve validation of the saliency models for real-life applications.

Regarding the applications, we decided to build a taxonomy made of three big categories:

- Abnormality detection: use the most salient areas detection.
- Normality detection: use the less salient areas detection.
- Abnormality processing: compare and further process the most salient areas.

This categories let us simplify and classify a very long list of applications which can benefit from attention models. We are just at the early stages of the use of saliency maps into computer vision applications. Nevertheless, the number of already existing applications shows a promising avenue for saliency models in improving existing applications, and for the creation of new applications. Indeed, several factors are nowadays turning saliency computation from labs to industry:

- The models accuracy drastically increased in two decades both concerning bottom-up saliency and top-down information and learning. The results of the recent models are way better than the first results in 1998.
- The models working both on videos and images are more and more numerous and provide more and more realistic results. New models including audio signals and 3D data are released and are expected to provide convincing results in the near future.
- The combined enhancement of computing hardware and algorithms optimization led to real-time or almost real-time good quality saliency computation.

While some industry already began to use attention maps (marketing), others (TV, Multimedia) come now to the use of such algorithms. Video surveillance and video summarization will also come into the game of using saliency maps shortly. This move from labs to industry will further encourage research on the topic towards understanding human attention, memory and human motivation. New models both using bottom-up and more and more top-down information will appear. Moreover, more validation techniques, mainly application-driven should be available in the next years to convince industry to use more attention modeling in their applications.
Literaturnachweis


18 Shiwei, Y., Tingzhu, H., Xiaoyun, L., and Wufan, C. (2013) Partial mutual information...


images using saliency-adaptive nonlinear disparity mapping. Displays.


video summary application, in *Digital Signal Processing (DSP)*, 2013 18th International Conference on, pp. 1–8, doi:10.1109/ICDSP.2013.6622744.


129 Miratech website proposes eye-tacking experiments for marketing material. URL [http://miratech.fr/](http://miratech.fr/)

130 Crazy egg website proposes mouse-tacking experiments for marketing material. URL [http://www.crazyegg.com/](http://www.crazyegg.com/)


135 Eyequant website proposes automatic saliency maps for marketing material. URL [http://www.eyequant.com/](http://www.eyequant.com/)


137 Attention wizzard website proposes automatic saliency maps for marketing material. URL [https://www.attentionwizard.com/](https://www.attentionwizard.com/)


166 Awad, D., Mancas, M., Riche, N., Courboulay, V., and Revel, A. A cbir-based evaluation framework for visual attention models.