6

Human Attention Modelization and Data Reduction

Matei Mancas, Dominique De Beul, Nicolas Riche and Xavier Siebert
IT Department, Faculty of Engineering (FPMs),
University of Mons (UMONS), Mons
Belgium

1. Introduction

Attention is so natural and so simple: every human, every animal and even every tiny insect is perfectly able to pay attention. In reality as William James, the father of psychology said: "Everybody knows what attention is". It is precisely because everybody "knows" what attention is that few people tried to analyze it before the 19th century. Even though the study of attention was initially developed in the field of psychology, it quickly spread into new domains such as neuroscience to understand its biological mechanisms and, most recently, computer science to model attention mechanisms. There is no common definition of attention, and one can find variations depending on the domain (psychology, neuroscience, engineering, ...) or the approach which is taken into account. But, to remain general, human attention can be defined as the natural capacity to selectively focus on part of the incoming stimuli, discarding less "interesting" signals. The main purpose of the attentional process is to make best use of the parallel processing resources of our brains to identify as quickly as possible those parts of our environment that are key to our survival.

This natural tendency in data selection shows that raw data is not even used by our multi-billion cells brain which prefers to focus on restricted regions of interest instead of processing the whole data. Human attention is thus the first natural compression algorithm. Several attempts towards the definition of attention state that it is very closely related to data compression and focus resources on the less redundant, thus less compressible data. Tsotosos suggested in Itti et al. (2005) that the one core issue which justifies attention regardless the discipline, methodology or intuition is "information reduction". Schmidhuber (2009) stated that ..."we pointed out that a surprisingly simple algorithmic principle based on the notions of data compression and data compression progress informally explains fundamental aspects of attention, novelty, surprise, interestingness ...". Attention modeling in engineering and computer science domain has very wide applications such as machine vision, audio processing, HCI (Human Computer Interfaces), advertising assessment, robotics and, of course, data reduction and compression.

In section 2, an introduction to the notions of saliency and attention will be given and the main computational models working on images, video and audio signals will be presented. In section 3 the ideas which either aims at replacing or complementing classical compression algorithms are reviewed. Saliency-based techniques to reduce the spatial and/or temporal resolution of non-interesting events are listed in section 4. Finally, in section 5, a discussion on the use of attention-based methods for data compression will conclude the chapter.

2. Attention modeling: what is saliency?

In this first part of the chapter, a global view of the methods used to model attention in computer science will be presented. The details provided here will be useful to understand the next parts of the chapter which are dedicated to attention-based image and video compression.

2.1 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" and implies a competition between "bottom-up" and "top-down" information. The idea of saliency maps is that the sight or gaze of people will direct to areas which, in some way, stand out from the background. The eye movements can be computed from the saliency map by using winner-take-all (Itti et al. (1998)) or more dynamical algorithms (Mancas, Pirri & Pizzoli (2011)). 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 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. The literature about attention modeling in computer science is not symmetric between those two approaches: the saliency-based methods are much more popular than the visibility models. For this reason, the following sections in this first part of the chapter will also mainly deal with saliency methods, but a review of visibility methods will be provided in the end.

2.2 Saliency approaches: bottom-up methods

Bottom-up approaches use features (most of the time low-level features but not always) extracted from the signal, 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 words 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, different methods are described for still images, videos and audio signals. All those modalities are of course interesting for multimedia compression which, by definition, contains both video and audio information.

2.2.1 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 the subject, 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 perfect 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 algorithmic 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. Therefore, we introduce here a new 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 or 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. Finally, the third class will take into account a context which is based on a model of what the normality should be. This model can be described as a priori probabilities, Fourier spectrum models ... In the following sections, the main methods from those three classes are described for still images.

2.2.1.1 Context: pixel's surroundings

This approach is based on a biological motivation and dates back to the work of Koch & Ullman (1985) on attention modeling. The main principle is to initially compute visual features at several scales in parallel, then 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, and 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.

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 computational models of bottom-up attention use the comparison of a central patch to its surroundings as a novelty indicator. An update is obtained by adding other features to the same architecture such as symmetry Privitera & Stark (2000) or curvedness Valenti et al. (2009). Le Meur et al. (2006) refined the model by using more biological cues like contrast sensitivity functions, perceptual decomposition, visual masking, and center-surround interactions. Another popular and efficient model is the Graph Based Visual Saliency model (GVBS, Harel et al. (2007)), which is very close to Itti et al. (1998) regarding feature extraction and center-surround, but differs from it in the fusion step where GBVS computes an activation map before normalization and combination. Other models like Gao et al. (2008) also used center-surround approaches even if the rest of the computation is made in a different mathematical framework based on a Bayesian approach.

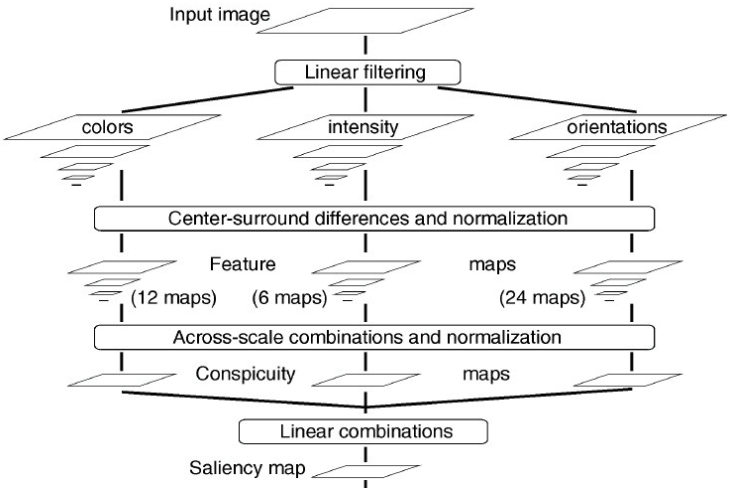


Fig. 1. Model of Itti et al. (1998). Three stages: center-surround differences, conspicuity maps, inter-feature fusion into saliency map.

2.2.1.2 Context: the whole image

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. 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 parts. 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).

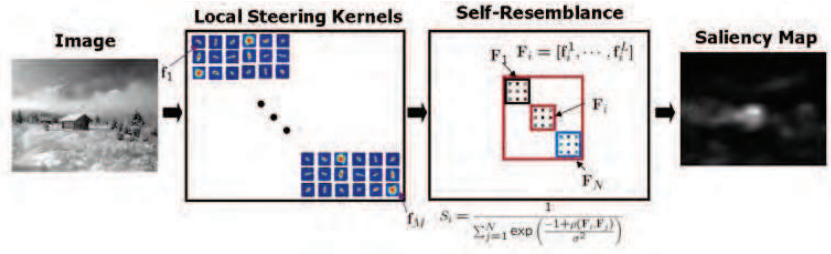


Fig. 2. Model of Seo & Milanfar (2009). Patches at different locations are compared.

A similar approach was developed in Mancas (2007) and Mancas (2009), that detects 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. The leftmost image is an apple with a defect shown 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 results from Mancas et al. (2007) which detected the apple edges, but also the defect. Finally the rightmost image is the mouse tracking result for more than 30 users.

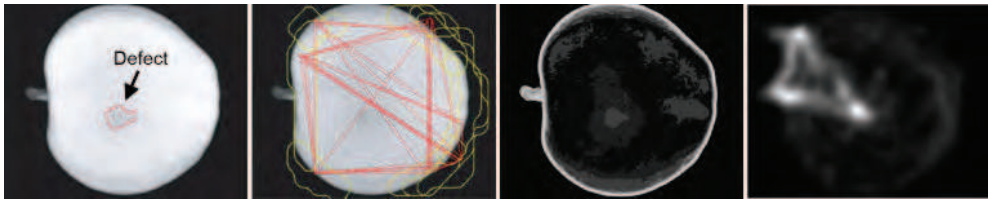


Fig. 3. Difference between locally contrasted and globally rare features. Left image: an apple with a defect in red, Second Image: Itti et al. (1998), Third image: Mancas et al. (2007), Right image: mouse tracking (ground truth).

A typical model using this context is the model of Stentiford (2001) which uses random neighborhoods and check if it is possible to find a lot of those neighborhoods or not in the rest of the image. If there are few possibilities, the patch was rare, thus salient. This model does not need feature extraction, as features remain included in the compared patches.

Oliva et al. (2003) also defined the saliency as the inverse likelihood of the features at each location. This likelihood is computed as a Gaussian probability all over the image on the features which are extracted by using a steerable pyramid. Boiman & Irani (2005) proposed a method where different patches were not only compared between them, but also their relative positions where taken into account.

A well-known model is Bruce & Tsotsos (2006). This model of bottom-up overt attention is proposed based on the principle of maximizing information sampled from a scene. The proposed operation is based on Shannon's self-information measure and is achieved in a neural circuit taking into account patches from the image projected on a new basis obtained by performing an ICA (Independent Component Analysis Hyvärinen et al. (2001)) on a large sample of 7x7 RGB patches drawn from natural images.

Recently, Goferman et al. (2010) has introduced context-aware saliency detection based on four principles. First, local low-level considerations, including factors such as contrast and color are used. Second, global considerations, which suppress frequently occurring features, while maintaining features that deviate from the norm are taken into account. Higher level information as visual organization rules, which state that visual forms may possess one or several centers of gravity about which the form is organized are then used. Finally, human faces detection are also integrated into the model. While the two first points are purely bottom-up, the two others may introduce some top-down information.

2.2.1.3 Context: a model of normality

This approach is probably less biologically-motivated in most of the 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 interesting. In Achanta et al. (2009) a very simple attention model was developed. His method, first, changes the color space from RGB to Lab and finds the Euclidean distance between the Lab pixel vectors in a Gaussian filtered image with the average Lab vector for the input image. This is illustrated in the Figure 4. The mean

image used is a kind of model of the image statistics and pixels which are far from those statistics are more salient.

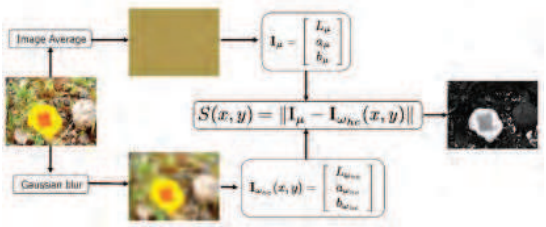


Fig. 4. Achanta et al. (2009) uses a model of the mean image.

In 2006, Itti & Baldi (2006) released the concept of surprise, central to attention. They described a formal Bayesian definition of surprise that is the only consistent formulation under minimal axiomatic assumptions. Surprise quantifies how data affects an observer, by measuring the difference between posterior and prior (model of normality) beliefs of the observer. In Hou & Zhang (2007), the authors proposed a model that is independent of any features. As it is known that natural images have a $\frac{1}{f}$ decreasing log-spectrum, the difference between this normality model obtained by low-pass filtering and the log-spectrum of the image is reconstructed into the image space and lead to the saliency map.

2.2.1.4 Attention models for still images: a comparison

It is not easy to classify attention models, for several reasons. First, there is a large variety of models. Second, some research groups (e.g., Itti's) have implemented different models, finding themselves in several categories. Also some approaches have several contexts and could be classified in more than one category, but based on the context notion, is seems possible to find this three main families of methods despite their diversity.

Figure 5 displays saliency maps computed with six models (available as Matlab codes), along with the eyestracking results to show where people really look at. For this purpose three images from the Bruce's database [1] were used. Along with the saliency maps of six models, one can find the most salient areas after automatic thresholding.

Figure 5 reveals that saliency maps can be quite different, from very fuzzy ones (Itti, Harrel or Seo) to high resolution ones (Mancas, Bruce or Achanta). It is not easy to compare those saliency maps (they should all be low-pass filtered to decrease their resolution). Nevertheless for the purpose of compression, one needs a model which is able to highlight the interesting areas but also the interesting edges as Mancas.

2.2.2 Videos

Part of the static models have been extended to video. Itti's model was generalized with the addition of motion features and flickering and in Itti & Baldi (2006) he applied another approach based on surprise to static but also dynamic images. Le Meur et al. (2007b) used motion in addition to spatial features. Gao et al. (2008) generalized his 2D square center-surround approach to 3D cubic shapes. Belardinelli et al. (2008) used an original approach of 3D Gabor filter banks to detect spatio-temporal saliency. Bruce & Tsotsos (2009)

[1] http://www-sop.inria.fr/members/Neil.Bruce/

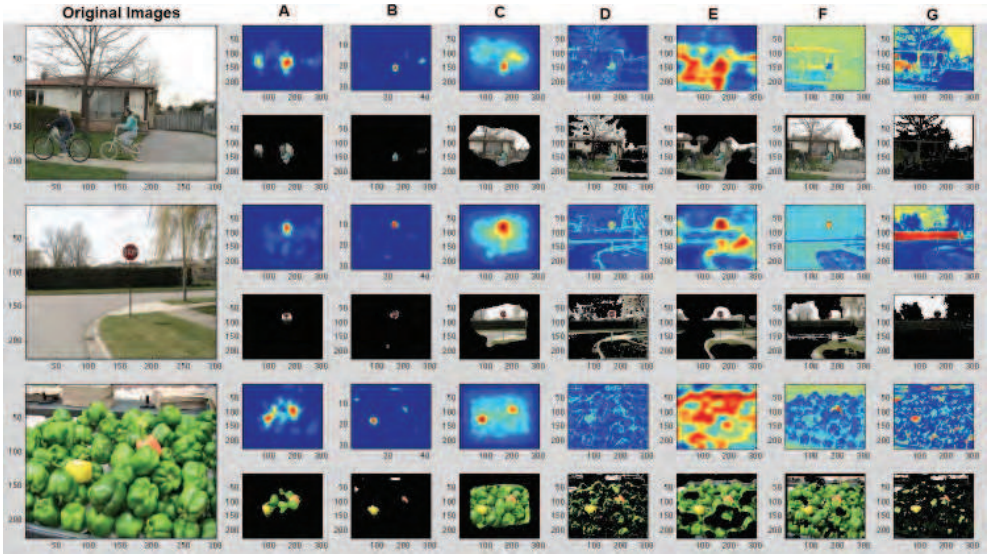


Fig. 5. For three original images (on the left), the eye-tracking results (column A) and six other saliency models maps. B = Itti and Koch (1998), C = Harrel et al. (2007); D = Mancas (2007), E = Seo and Milanfar (2009), F = Bruce and Tsotsos (2005), G = Achanta (2009). A thresholded applied on the saliency maps is shown on the images bellow.

extended his model by learning ICA not only on 2D patches but on spatio-temporal 3D patches. As shown in Figure 6, similarly to Gao, Seo & Milanfar (2009) introduced the time dimension in addition to his static model. Another model is SUN (Saliency Using Natural

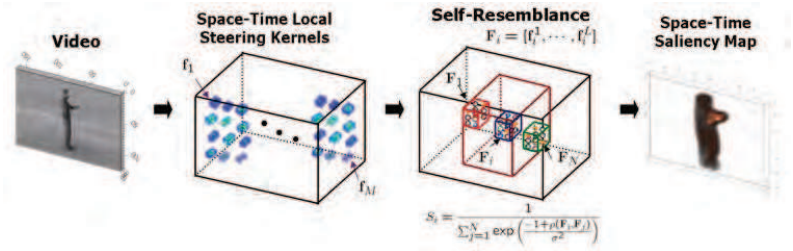


Fig. 6. Seo & Milanfar (2009) generalized to video in 2009.

statistics) from Butko et al. (2008) that propose a Bayesian framework for saliency. Two methods are implemented. First, the features are calculated as responses of biologically plausible linear filters, such as DoG (Differences of Gaussians) filters. Second, the features are calculated as the responses to filters learned from natural images using independent component analysis (ICA).

Frintrop (2006) introduces the biologically motivated computational attention system VOCUS (Visual Object detection with a Computational attention System) that detects regions of interest in images. It operates in two modes, in an exploration mode in which no task is provided, and in a search mode with a specified target. The bottom-up mode is based on an enhancement of the Itti model.

Finally, 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. It shows good results from four objects to dense crowds with increasing performance. The idea here is to show that motion is most of the time salient but within motion, there might be motion which is more or less salient. Mancas model is capable of extracting different motion behavior from complex videos or crowds (Figure 7).

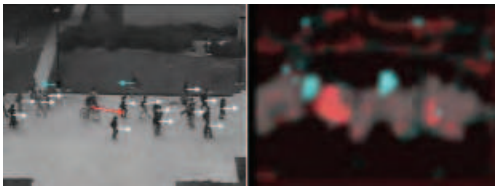


Fig. 7. Detection of salient motion compared to the rest of motion. Red motion is salient because of unexpected speed. Cyan motion is salient because of unexpected direction Mancas, Riche & J. Leroy (2011).

2.2.2.1 Extension to 3D

With the release of the Microsoft's Kinect sensor [2], in November 2010, 3D features have become easily accessible. 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. A 3D model-based motion detection in a scene has been implemented by Riche et al. (2011). The proposed algorithm has three main steps. First, 3D motion (speed and direction) features are extracted from the RGB video and the depth map of the Kinect sensor. The second step is a spatiotemporal filtering of the features at several scales to provide multi scale statistics. Finally, the third step is the rarity-based attention computation within the video frame.

2.2.3 Audio signals

There are very few auditory attention models compared to visual attention models. However, we can classify existing models into different categories.

As shown in Figure 8, 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.

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.

[2] http://www.xbox.com/kinect

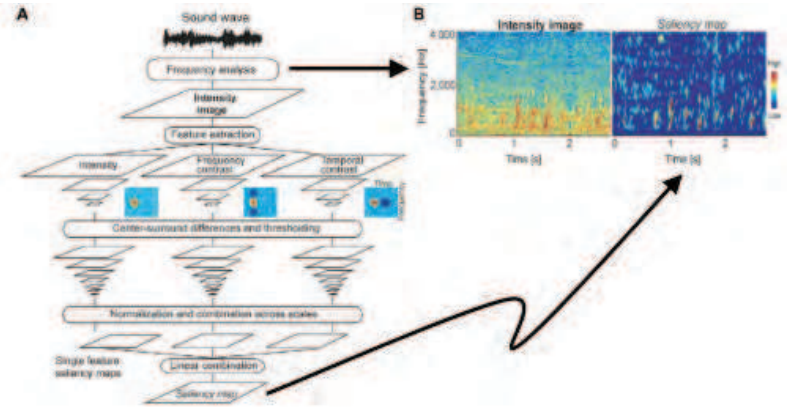


Fig. 8. Kayser et al. (2005) audio saliency model inspired from Itti.

Couvreur et al. (2007) define features that can be computed along audio signals in order to assess the level of auditory attention on a normalized scale, i.e. between 0 and 1. The proposed features are derived from a time-frequency representation of audio signals and highlight salient regions such as regions with high loudness, temporal and frequency contrasts. Normalized auditory attention levels can be used to detect sudden and unexpected changes of audio textures and to focus the attention of a surveillance operator to sound segments of interest in audio streams that are monitored.

2.3 Saliency models: including top-down information

There are two main families of top-down information which can be added to bottom-up attention. The first one mainly deals with learnt normality 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.

2.3.1 Top-down as learnt normality: attending unusual events

Concerning still images, the "normal" gaze behavior can be learnt from the "mean observer". Eye-tracking techniques can be used on several users, and the mean 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. (2009) 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 fact seems logical as natural images are acquired using cameras and the photographer will naturally tend to locate the objects of interest in the center of the picture. This observation might be interesting in the field of image compression as high quality compression seems to be required mainly in the center of the image while peripheral areas could be compressed with lower rates.

Of course, this observation for natural images is very different from more specific images which use a priori knowledge. Mancas (2009) showed using mouse tracking that gaze density

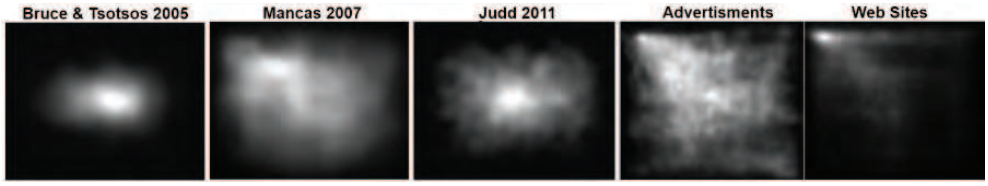


Fig. 9. Three models of the mean observer for natural images on the left. The two right images: model of the mean observer on a set of advertising and websites images.

is very different on a set of advertisements and on a set of websites as it is showed in 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 learnt 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 learnt 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 learnt through HMMs (Hidden Markov Models) Jouneau & Carincotte (2011).

2.3.2 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 the 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").

2.3.2.1 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 learnt object.

2.3.2.2 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.

Fig. 10. Bottom-up saliency model inhibited by top-down information to select only salient people.

2.4 Visibility models

Compared to other Bayesian frameworks (e.g. Oliva et al. (2003)), these models have a 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. 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 top-down centered preference for natural images we developed in section Top-down as learnt normality: attending unusual events.

3. Attention-based visual coding

Since the late 1990's techniques based on attention have been introduced in the field of image and video coding (e.g., Kortum & Geisler (1996); Maeder et al. (1996)). Attention can be used to compress videos or to transmit the most salient parts first during the data transfer from a server to a client. This section will first introduce general principles of video compression, then review some of the major achievements in saliency-based visual coding.

3.1 General video coding and compression

The goal of this section is to briefly introduce the concepts of video coding and compression, which tends to be used interchangeably since they are heavily related. What follows in this section is a short introduction to general video compression for which Lorente (2011) is an example of recent exhaustive review.

Video compression is the process of converting a video signal into a format that takes up less storage space or transmission bandwidth. It can be considered as a coding scheme that reduces bits of information representing the video. Nevertheless the overall visual quality has to be preserved, leading to a compromise between the level of artifacts and the bandwidth.

Two types of compression can be distinguished: lossy and lossless compression Richardson (2003). In a lossless compression system statistical redundancy is removed so that the original data can be perfectly reconstructed. Unfortunately, at the present time lossless methods only allows a modest amount of compression, insufficient for most video applications. On the other hand, lossy compression provides bigger compression ratios, at the expense of not being able to reconstruct perfectly the original signal. Lossy compression is the type of compression most commonly used for video, attention-based or not.

It is interesting to note that, even if generic compression algorithms do not explicitly use saliency, they implicitly exploit the mechanisms of human visual perception to remove redundant information Geisler & Perry (1998). For example retinal persistence of vision makes the human eye keep an instantaneous view of a scene for about one-tenth of a second at least. This allows video (theoretically a continuum of time) to be represented by a series of discrete frames (e.g., 24 frames per second) with no apparent motion distortion.

Coding standards define a representation of visual data as well as a method of decoding it to reconstruct visual information. Recent hybrid standards like H.264/AVC Wiegand et al. (2003) have led to significant progress in compression quality, allowing for instance the transmission of high definition (HD) television signals over a limited-capacity broadcast channel, and video streaming over the internet Richardson (2003).

Emerging video coding standard H.265 [3] aims at enhancing video coding efficiency using intensive spatiotemporal prediction and entropy coding. Nevertheless, this new standard only considers *objective* redundancy, as opposed to attention-based methods described below.

3.2 A review of the attention-based methods

The above mentioned 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 11). The aim of these methods is to achieve compression without significant degradation of perceived quality.

Saliency-based methods derive from biological properties of the human eye, that enable one to focus only on a limited region of an image at a time. It is thus a subjective notion, but a lot of research has been devoted to its modeling and quantification.

In the following there is an attempt to list the methods currently available in the literature, pointing to their strengths and weaknesses when possible. Although there is currently no

[3] http://www.h265.net

Fig. 11. Illustration of the distortions introduced by general compression methods (three first images on the left) compared to saliency-based compression (three last images on the right), at three different compression levels. Adapted from Yu & Lisin (2009))

unified taxonomy, we have divided the methods into interactive, indirect and direct, the latter being the most commonly studied.

3.2.1 Interactive approaches

As described above, earlier approaches for modeling the human visual system (HVS) relied on eye-tracking devices to monitor attention points (e.g., Kortum & Geisler (1996)).

With such devices, encoding continuously and efficiently follows the focus of the observer. Indeed, observers usually do not notice any degradation of the received frames. 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). A general coding scheme should be independent of the number of observers, the viewing distance, and any hardware device or user interaction.

Even in the absence of eye tracking, an interactive approach has demonstrated usefulness. Observers can for example explicitly point to priority regions with the mouse Geisler & Perry (1998). However, extending this approach to general-purpose non-interactive video compression presents severe limitations.

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. Despite progresses in attention modelling and even though human gaze is well modelled in the presence of salient objects, it is not possible to obtain a reliable model of human gaze in the absence of specific salient objects (as can be seen in Figure 12). Indeed, the highly dynamical process of eye movements is influenced a lot by previous gaze position if no salient objects pops out from the background.

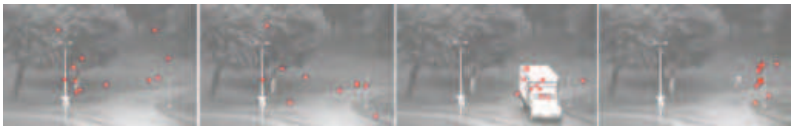


Fig. 12. The two left images show several users eye tracking results which are spread through the image and very different, while the two images on the right showing clear regions of interest will exhibit much more correlated fixations.

3.2.2 Indirect approaches

Indirect compression consists in 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.

The seminal model of Itti et al. (1998) was later applied to video compression in Itti (2004) by computing a saliency map for each frame of a video sequence and applying a smoothing filter 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.

The main advantages of this method are twofold. First, a high correlation with human eye movements on unconstrained video inputs is observed. Second a good compression rate is achieved, the average size of a compressed video being approximately half the size of the original one, for both MPEG-1 and MPEG-4 (DivX) encodings.

Another method combines both top-down and bottom-up information, using a wavelet decomposition for multiscale analysis Tsapatsoulis et al. (2007). Bit rate gain ranging from 15% to 64% for MPEG-1 videos and from 10.4% to 28.3% for MPEG-4 are reported.

Mancas et al. (2007) proposed an indirect approach based on their attention model. 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 13). Depending on the method parameters, images could be compressed twice as much for standard JPEG. Nevertheless even though the quality of the important areas remain unchanged, the quality of the less important regions can dramatically decrease. It is thus not easy to compare the compression rate as the quality of the images remains subjective.

Fig. 13. Two pairs of images (original and anisotropic filtered). Adapted from Mancas et al. (2007).

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 afterwards. This fact also led 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 Li et al. (2011).

3.2.3 Direct approaches

Recent work on modeling visual attention (Le Meur, Itti, Parkhurst, Chauvin ...) 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.

Li et al. (2011), a recent extension of Itti (2004), uses a similar neurobiological model of visual attention to generate a saliency map, whose 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 Li et al. (2011) 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 Hou & Zhang (2007) and the Phase spectrum of Quaternion Fourier Transform Guo & Zhang (2010). The goal here 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. Compression rates between 32.6% (from 8.88Mb for raw H-264 file to 5.98Mb for compressed file) and 38% (from 11.4Mb for raw MPEG-4 file to 7.07Mb for compressed file) are reported in Guo & Zhang (2010).

These Fourier-based approaches are computationally efficient, but they are less connected to the Human Visual System. They also 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 Li et al. (2011), a scheme for attention video compression has recently been suggested by Gupta & Chaudhury (2011). It proposes a learning-based 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.

3.2.4 Enhancing the objective quality

Many encoding techniques have sought to optimize perceptual rather than objective quality: these techniques allocate more bits to the image areas where human can easily see coding distortions, and allocate fewer bits to the areas where coding distortions are less noticeable. Experimental subjective quality assessment results show that visual artifacts can be reduced through this approach. However two problems arise: first, the mechanisms of human perceptual sensitivity are still not fully understood, especially as captured by computational models; second, perceptual sensitivity may not necessarily explain people's attention.

The use of top-down information is very efficient as it is very likely to be attended. Face detection is one of the crucial features, but also text detection, skin color, motion-related events for video-surveillance, . . . (see for example Tan & Davis (2004) and references therein).

4. Image retargeting based on saliency maps

In previous sections, the compression algorithms do not modify the original spatial (frame resolution) and temporal (video length) size of 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 will keep some context 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.

Images can be resized (zoomed) according to two families of methods, either taking into account the relevance of the content or not. In the first case, it not only requires to preserve the content but also the structure of the original image Shamir & Sorkine (2009) so the cropping should avoid to be too restrictive.

The second case comprises methods such as letter and pillar boxing, fixed windowing cropping and scaling. These methods are fast but often give poor results. Indeed, fixed windowing cropping can leave relevant contents outside the window, while scaling can engender losses of important information which could eventually lead to an unrecognizable image Liu & Gleicher (2005). Therefore, these classical methods are not mentioned further in this section.

4.1 Spatio-temporal visual data repurposing: perceptual zoom

Human beings are naturally able to perceive interesting areas of an image, as illustrated in the previous sections of this chapter. 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 presented above in this chapter. Interestingly, such techniques can also be used for real time spatio-temporal images broadcast Chamaret et al. (2010).

Figure 14 shows several perceptual zooms depending on a parameter which will threshold the smoothed Mancas (2009) saliency map.

In the next section, the main general retargeting methods based on content-aware cropping are presented. Then, attention-based retargeting methods are more specifically described.

4.1.1 General retargeting methods

An interactive cropping method is proposed by Santella et al. (2006), whose purpose is to create aesthetically pleasing crops without explicit interaction. The location of important content of the image is realized by segmentation and by eye-tracking. With a collection of gaze-based crops and an optimization function, the method identifies the best crops.

Another approach is to calculate automatically an "active window" with a predefined size, as presented by Tao et al. (2007). In this case the retargeted image has to satisfy two

Fig. 14. Example of images along with rectangles providing different attention-based automatic zooms. After a saliency map (Mancas (2009)) 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.

requirements: to preserve temporal and spatial distance, and to contain useful information such as objects shapes and motions. These methods are composed of the following three steps. First, extraction of the image foreground, for example by minimizing an energy function to automatically separate foreground from background pixels. Second, optimization of the active window to fit in the target size. Third, clustering to reduce the number of parameters to be estimated in the optimization process.

4.1.2 Saliency maps based methods

Figure 15 perfectly illustrates the process of the saliency-based retargeting. From the original image, on the left, a saliency map is computed (in the middle) from which an area with higher intensity is extracted using some algorithm and its bounding-box will represent the zoom.

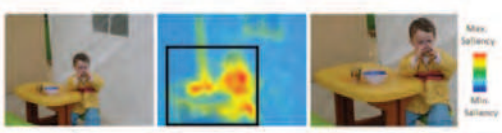


Fig. 15. Example of retargeting: left, the original picture; middle, the saliency map; right, the reframed picture. (adapted from Le Meur & Le Callet (2009))

A technique to determine automatically the "right" viewing area for spatio-temporal images is proposed in Deselaers et al. (2008). Images are first analyzed to determine relevant regions by using three strategies: the visual saliency of spatial images, optical flow for movements and the appearance of the image. A log-linear algorithm then computes the relevance for every position of the image to determine a sequence of cropping positions with a correct aspect ratio for the display device.

Suh et al. (2003) uses the Itti & Koch (2001) algorithm to compute the saliency map, 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 take into account most of the saliency while remaining sufficiently small.

The previous methods show that the perceptual zoom not only compresses the images, but it also allows better recognition during visual search!

The Self-Adaptive Image Cropping for Small Displays Ciocca et al. (2007) 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.

The RSVP (Rapid Serial Visual Presentation de Bruijn & Spence (2000)) method for images can also be adapted to allow in a sequential way and during a short time the visualization and browsing of the interest regions Fan et al. (2003). Here also, the bottom-up attention saliency is computed with Itti & Koch (2001) while top-down information is added: texts and faces detection. The most relevant interest regions are proposed to mobile phones as key images.

Liu et al. (2007) 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 Le Meur et al. (2007a). 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. A subset of the corners coordinates of the squares in which are predicted eye gaze centered on a local maximum of saliency is determined. The coordinates of the upper left and the lower right corners of the final zoom thumbnail are set to include a square area centered on the relevant local maximum.

4.2 Spatio-temporal resolution decrease for uninteresting regions: anisotropic resolution

Perceptual zoom does not always preserve the image structure. For example, Figure 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 main methods are described in this section: warping and seam carving. These methods may cause non-linear visual distortions on several regions of the image (Zhou et al. (2003)).

4.2.1 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 Liu & Gleicher (2005). Another approach of warping is to place a grid mesh onto the image and then compute a new geometry for this mesh (Figure 16), such that the boundaries fit the new desired image sizes, and the quad faces covering important image regions remain intact at the expense of larger distortion to the other quads Wang et al. (2008).

Automatic image retargeting with fisheye-view warping Liu & Gleicher (2005) uses an "importance map" that combines salience and object information to find automatically, with a greedy algorithm, a minimal rectangular region of interest. A non-linear function is then used for warping to ensure that the distortion in the region of interest is smaller than elsewhere in the image.

Non-homogeneous content-driven video-retargeting Wolf et al. (2007) 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.

Ren et al. (2009) introduces a retargeting method based on global energy optimization. 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 Ren et al. (2010). 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.

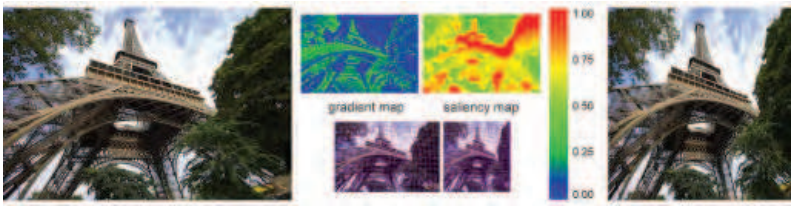


Fig. 16. The original image (left) is deformed by a grid mesh structure to be fit in the required size (right). The scaling and stretching depend on the gradient and saliency map. Source : `http://graphics.csie.ncku.edu.tw/Image_Resizing/`

Wang et al. (2008) present a warping method which uses the grid mesh of quads to retarget the images (figure 16). 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.

4.2.2 Seam carving

Seam carving Avidan & Shamir (2007) 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 oriented gradients, or saliency maps Vaquero et al. (2010). 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.

Figure 17 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 Wonjun et al. (2011) 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 Chamaret et al. (2010). Even if saliency maps

Fig. 17. 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

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, Rubinstein et al. (2008) 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.

Grundmann et al. (2010) proposed a saliency-based spatio-temporal seam-carving approach with much better spatio-temporal continuity than Rubinstein et al. (2008). 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 proposed by the author are temporally discontinuous providing only the appearance of a continuous seam which helps in keeping both spatial and temporal coherence.

5. Discussion and perspectives

5.1 Two main approaches

In this chapter we discussed the use of saliency-based methods on two main approaches to image and video compression. The first one uses the result of the saliency maps to compress the signal but it does not modify the original spatial (frame resolution) and temporal (video length) size of the signal. The second one uses saliency maps to crop or reduce the spatio-temporal resolution of the signal. In this latter case, the compression is not obtained through signal quality reduction, but through quantity reduction. Of course, both methods can be used together and they are more or less interesting depending on the application.

5.2 Automatic human attention prediction issues

As already shown in Figure 12, different viewers' gaze can be predictable or not depending on the situation, thus a compression system should take this fact into account. If there is not a real salient object standing out from the background, the compression scheme should not take saliency into account while, this one can help if salient objects are present.

Another point to take into account is the shape of the saliency maps. As stated in section Attention models for still images: a comparison, saliency maps with a high resolution and which also highlight edges might be more convenient for compression purposes than more

fuzzy approaches. Those maps preserve important details where artifacts would be clearly disturbing.

Attention-based visual coding seems to become less crucial as the bandwidth of Internet and TV continuously increase. Nevertheless, for precise applications like video-surveillance where the quality of uninteresting textures is not a problem and where the transmission bandwidth may be a problem, especially for massive HD multi-camera setups, the saliency-based approaches are very relevant. In the same way, storage of huge amount of live visual data is very resource-demanding and the best compression is needed while preserving the main events.

Concerning image and video retargeting and summarization, the perceptual zooming and smart resizing is of great importance in the context of smart mobile devices becoming common. Those devices have limited screen sizes and their bandwidth is much less easy to control in terms of quality of service and bandwidth. Intelligent and flexible methods of automatic thumbnailing, zoom, resizing and repurposing of audio-video data are crucial for a fast developing HD multimedia browsing market. Of course, in this case, a very good spatio-temporal continuity is required.

5.3 Quality evaluation and comparison issue

Coding artifacts in non-salient regions might attract attention of the viewer to these regions, thereby degrading visual quality. This problem is particularly noticeable at low bit rates as it can be seen in Figure 18: for example some repeating patterns like textures are not interesting but they become interesting (actually annoying) if they have compression artifacts or defects. Several methods have been proposed to detect and reduce such coding artifacts, to keep user's attention on the same regions that were salient before compression. It is however difficult to find appropriate criteria and quality metrics Farias (2010); Ninassi et al. (2007), and benchmark datasets (e.g.,Li et al. (2009)).

Fig. 18. First row: classical compression, Second row: attention-based compression. Adapted from `http://www.svcl.ucsd.edu/projects/ROI_coding/demo.htm`.

Another recurring problem encountered in writing this review is the lack of cross-comparison between the different methods. For example few authors report compression rates for an equivalent perceptual quality. The notion of "equivalent quality" itself seems difficult to define as even objective methods are not necessary perceptually relevant. This problem is particularly important for the methods in section Attention modeling: what is saliency? but it is also present in the retargeting and summarization methods from section Image retargeting based on saliency maps.

One way to fill in these data would be to provide datasets on the internet that would serve as benchmarks.

5.4 Saliency cross-modal integration: combining audio and visual attention

In a multimedia file a lot of information is included into the visual data. But also, supplemental or complementary information can be found within the audio track: audio data could confirm visual data information, help in being more selective or even bring new information that is not present in the camera field of view. Indeed, in some contexts sound might even be the only way to determine where to focus visual attention, for example if several persons are in a room but only one is talking. It seems thus that the use of both visual and audio saliency is a relevant idea.

Multimodal models of attention are unfortunately very few and they are mainly used in the field of robotics such in Ruesch et al. (2008). Another interesting idea is to localize the sound-emitting regions in a video. Recent work as Lee et al. (2010) has shown the ability to localize sounds in an image.

Given the computationally intensive nature and the real-time requirements of video compression methods and especially in the case of multimodal integration of saliency maps, some algorithms have exploited recent advances in Graphics Processing Unit (GPU) computing. In particular, a parallel implementation of a spatio-temporal visual saliency model has been proposed Rahman et al. (2011).

5.5 Saliency models and new trends in multimedia compression

Visual compression has been a very active field of research and development for over 20 years, leading to many different compression systems and to the definition of international standards. Even though video compression has become a mature field, a lot of research is still ongoing. Indeed, as the quality of the compression increases, so does users' level of expectations and their intolerance to artifacts. Exploiting saliency-based video compression is a challenging and exciting area of research and especially nowadays when saliency models include more and more top-down information and manage to better and better predict real human gaze.

Multimedia applications are a continuously evolving domain and compression algorithms must also evolve and adapt to new applications. The explosion of portable devices with less bandwidth and smaller screens, but also the future semantic TV/web and its object-based description will lead in the future to a higher importance of saliency-based algorithms for multimedia data repurposing and compression.

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)*.

Avidan, S. & Shamir, A. (2007). Seam carving for content-aware image resizing, *ACM Trans. Graph.* 26(3): 10.

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.

Belardinelli, A., Pirri, F. & Carbone, A. (2008). Motion saliency maps from spatiotemporal filtering, *In Proc. 5th International Workshop on Attention in Cognitive Systems - WAPCV 2008*, pp. 7–17.

Boiman, O. & Irani, M. (2005). Detecting irregularities in images and in video, *International Conference on Computer Vision (ICCV)*.

Bruce, N. D. B. & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach, *Journal of Vision* 9(3).

Bruce, N. & Tsotsos, J. (2006). Saliency based on information maximization, *in* Y. Weiss, B. Schölkopf & J. Platt (eds), *Advances in Neural Information Processing Systems 18*, MIT Press, Cambridge, MA, pp. 155–162.

Butko, N. J., Zhang, L., Cottrell, G. & Movellan, J. (2008). Visual saliency model for robot cameras, *IEEE Inter. Conf. on Robotics and Automation (ICRA)*, pp. 2398Ű-2403.

Chamaret, C., Le Meur, O., Guillotel, P. & Chevet, J.-C. (2010). How to measure the relevance of a retargeting approach?, *Workshop Media Retargeting ECCV 2010*, Crete, Grèce, pp. 1–14.

Ciocca, G., Cusano, C., Gasparini, F. & Schettini, R. (2007). Self-adaptive image cropping for small displays, *IEEE Transactions on Consumer Electronics* 53(4): 1622–1627.

Couvreur, L., Bettens, F., Hancq, J. & Mancas, M. (2007). Normalized auditory attention levels for automatic audio surveillance, *International Conference on Safety and Security Engineering (SAFE)*.

de Bruijn, O. & Spence, R. (2000). Rapid serial visual presentation: A space-timed trade-off in information presentation, *Advanced Visual Interfaces*, pp. 189–192.

Deselaers, T., Dreuw, P. & Ney, H. (2008). Pan, zoom, scan – time-coherent, trained automatic video cropping, *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Anchorage, AK, USA.

Fan, X., Xie, X., Ying Ma, W., Jiang Zhang, H. & qin Zhou, H. (2003). Visual attention based image browsing on mobile devices, *Proc. of ICME 2003*, IEEE Computer Society Press, pp. 53–56.

Farias, M. C. Q. (2010). *Video Quality Metrics (in: Digital Video)*, InTech.

Frintrop, S. (2006). Vocus: A visual attention system for object detection and goal-directed search, *Thesis print*, Vol. 3899 of *Lecture Notes in Artificial Intelligence*, Springer Berlin / Heidelberg.

Gao, D., Mahadevan, V. & Vasconcelos, N. (2008). On the plausibility of the discriminant center-surround hypothesis for visual saliency., *J Vis* 8(7): 13.1–1318.

Geisler, W. S. & Perry, J. S. (1998). A real-time foveated multiresolution system for low-bandwidth video communication, *in Proc. SPIE*, pp. 294–305.

Goferman, S., Zelnik-Manor, L. & Tal, A. (2010). Context-aware saliency detection, *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 2376–2383.

Grundmann, M., Kwatra, V., Han, M. & Essa, I. (2010). Discontinuous seam-carving for video retargeting, *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 569–576.

Guo, C. & Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression., *IEEE Trans Image Process* 19(1): 185–198.

Gupta, R. & Chaudhury, S. (2011). A scheme for attentional video compression, *Pattern Recognition and Machine Intelligence* 6744: 458–465.

Harel, J., Koch, C. & Perona, P. (2007). Graph-based visual saliency, *Advances in Neural Information Processing Systems 19*, MIT Press, pp. 545–552.

Hou, X. & Zhang, L. (2007). Saliency detection: A spectral residual approach, *Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR '07*, pp. 1–8.

Hyvärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*, New York: Wiley.

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention, *IEEE Transactions on Image Processing* 13(10): 1304–1318.

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. (2001). Computational modelling of visual attention, *Nature Reviews Neuroscience* 2(3): 194–203.

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.

Itti, L., Rees, G. & Tsotsos, J. (2005). *Neurobiology of Attention*, Elsevier Academic Press.

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 & Torralba, A. (2009). 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.

Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry., *Hum Neurobiol* 4(4): 219–227.

Kortum, P. & Geisler, W. (1996). Implementation of a foveated image coding system for image bandwidth reduction, *In Human Vision and Electronic Imaging, SPIE Proceedings*, pp. 350–360.

Le Meur, O. & Le Callet, P. (2009). What we see is most likely to be what matters: visual attention and applications, *Proceedings of the 16th IEEE international conference on Image processing*, ICIP'09, IEEE Press, Piscataway, NJ, USA, pp. 3049–3052.

Le Meur, O., Le Callet, P. & Barba, D. (2007a). Construction d'images miniatures avec recadrage automatique basé sur un modéle perceptuel bio-inspiré, *Traitement du signal*, Vol. 24(5), pp. 323–335.

Le Meur, O., Le Callet, P. & Barba, D. (2007b). Predicting visual fixations on video based on low-level visual features, *Vision Research* 47: 2483–2498.

Le Meur, O., Le Callet, P., Barba, D. & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28(5): 802 –817.

Lee, J., De Simone, F. & Ebrahimi, T. (2010). Efficient video coding based on audio-visual focus of attention, *Journal of Visual Communication and Image Representation* 22(8): 704–711.

Legge, Hooven, Klitz, Mansfield & Tjan (2002). Mr.chips 2002: new insights from an idealobserver model of reading, *Vision Research* pp. 2219–2234.

Li, J., Tian, Y., Huang, T. & Gao, W. (2009). A dataset and evaluation methodology for visual saliency in video, *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pp. 442 –445.

Li, Z., Qin, S. & Itti, L. (2011). Visual attention guided bit allocation in video compression, *Image and Vision Computing* 29(1): 1 – 14.

Liu, F. & Gleicher, M. (2005). Automatic image retargeting with fisheye-view warping, *Proceedings of User Interface Software Technologies (UIST)*.

Liu, H., Jiang, S., Huang, Q., Xu, C. & Gao, W. (2007). Region-based visual attention analysis with its application in image browsing on small displays, *ACM Multimedia*, pp. 305–308.

Lorente, J. D. S. (ed.) (2011). *Recent Advances on Video Coding*, InTech.

Maeder, A. J., Diederich, J. & Niebur, E. (1996). Limiting human perception for image sequences, *in* B. E. Rogowitz & J. P. Allebach (ed.), *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, Vol. 2657, pp. 330–337.

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. & Gosselin, B. (2010). Dense crowd analysis through bottom-up and top-down attention, *Proc. of the Brain Inspired Cognitive Sytems (BICS)*.

Mancas, M., Gosselin, B. & Macq, B. (2007). Perceptual image representation, *J. Image Video Process.* 2007: 3–3.

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*.

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.

Ninassi, A., Le Meur, O., Le Callet, P. & Barbba, D. (2007). Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric, *IEEE Inter. Conf. on Image Processing (ICIP)* , Vol. 2, pp. 169–172.

Oliva, A., Torralba, A., Castelhano, M. & Henderson, J. (2003). Top-down control of visual attention in object detection, *IEEE Inter. Conf. on Image Processing (ICIP)* , Vol. 1, pp. I – 253–6 vol.1.

Privitera, C. M. & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations, *IEEE Trans. Pattern Anal. Mach. Intell.* 22(9): 970–982.

Rahman, A., Houzet, D., Pellerin, D., Marat, S. & Guyader, N. (2011). Parallel implementation of a spatio-temporal visual saliency model, *Journal of Real-Time Image Processing* 6: 3–14.

Ren, T., Liu, Y. & Wu, G. (2009). Image retargeting using multi-map constrained region warping, *ACM Multimedia*, pp. 853–856.

Ren, T., Liu, Y. & Wu, G. (2010). Rapid image retargeting based on curve-edge grid representation, *IEEE Inter. Conf. on Image Processing (ICIP)* , pp. 869–872.

Richardson, I. E. (2003). *H.264 and MPEG-4 Video Compression: Video Coding for Next Generation Multimedia*, 1 edn, Wiley.

Riche, N., Mancas, M. & B. Gosselin, T. D. (2011). 3d saliency for abnormal motion selection: the role of the depth map, *Proceedings of the ICVS 2011*, Lecture Notes in Computer Science, Springer Berlin / Heidelberg.

Rubinstein, M., Shamir, A. & Avidan, S. (2008). Improved seam carving for video retargeting, *ACM Transactions on Graphics (SIGGRAPH)* 27(3): 1–9.

Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J. & Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub, *IEEE Int. Conf. on Robotics and Automation*, p. 6.

Santella, A., Agrawala, M., Decarlo, D., Salesin, D. & Cohen, M. (2006). Gaze-based interaction for semi-automatic photo cropping, *In CHI 2006*, ACM, pp. 771–780.

Schmidhuber, J. (2009). Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes, *in* G. Pezzulo, M. Butz, O. Sigaud & G. Baldassarre (eds), *Anticipatory Behavior in Adaptive Learning Systems*, Vol. 5499 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 48–76.

Seo, H. J. & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance, *Journal of Vision* 9(12).

Shamir, A. & Sorkine, O. (2009). Visual media retargeting, *ACM SIGGRAPH ASIA 2009 Courses*, SIGGRAPH ASIA '09, ACM, New York, NY, USA, pp. 11:1–11:13.

Stentiford, F. W. M. (2001). An estimator for visual attention through competitive novelty with application to image compression, *Proc. Picture Coding Symposium*, pp. 101–104.

Suh, B., Ling, H., Bederson, B. B. & Jacobs, D. W. (2003). Automatic thumbnail cropping and its effectiveness., *Proceedings of the 16th annual ACM symposium on User interface software and technology (UIST)*, pp. 95–104.

Tan, R. & Davis, J. W. (2004). Differential video coding of face and gesture events in presentation videos, *Computer Vision and Image Understanding* 96(2): 200 – 215. Special Issue on Event Detection in Video.

Tao, C., Jia, J. & Sun, H. (2007). Active window oriented dynamic video, *Workshop on Dynamical Vision at the Inter. Conf. on Comp. Vision (ICCV)*

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.

Tsapatsoulis, N., Rapantzikos, K. & Pattichis, C. (2007). An embedded saliency map estimator scheme: Application to video encoding, *International Journal of Neural Systems* 17(4): 1–16.

Valenti, R., Sebe, N. & Gevers, T. (2009). Image saliency by isocentric curvedness and color, *Inter. Conf. on Comp. Vision (ICCV)*.

Vaquero, D., Turk, M., Pulli, K., Tico, M. & Gelf, N. (2010). A survey of image retargeting techniques, *SPIE Applications of Digital Image Processing*.

Wang, Y.-S., Tai, C.-L., Sorkine, O. & Lee, T.-Y. (2008). Optimized scale-and-stretch for image resizing, *ACM Trans. Graph. (Proceedings of ACM SIGGRAPH ASIA)* 27(5).

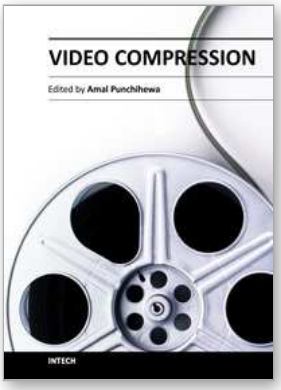
Wiegand, T., Sullivan, G. J., Bjntegaard, G. & Luthra, A. (2003). Overview of the h.264/avc video coding standard., *IEEE Trans. Circuits Syst. Video Techn.* pp. 560–576.

Wolf, L., Guttmann, M. & Cohen-Or, D. (2007). Non-homogeneous content-driven video-retargeting, *Proceedings of the Eleventh IEEE International Conference on Computer Vision (ICCV-07)*.

Wonjun, K., Chanho, J. & Changick, K. (2011). Spatiotemporal saliency detection and its applications in static and dynamic scenes, *IEEE Trans. Circuits and Systems for Video Tech.* 21(4): 10.

Yu, S. X. & Lisin, D. A. (2009). Image compression based on visual saliency at individual scales., *International Symposium on Visual Computing*, pp. 157–166.

Zhou, Lu, L. & Bovik., A. (2003). Foveation scalable video coding with automatic fixation selection, *IEEE Transactions on Image Processing* 12(2): 243–254.



Video Compression

Edited by Dr. Amal Punchihewa

ISBN 978-953-51-0422-3
Hard cover, 154 pages
Publisher InTech
Published online 23, March, 2012
Published in print edition March, 2012

Even though video compression has become a mature field, a lot of research is still ongoing. Indeed, as the quality of the compressed video for a given size or bit rate increases, so does users' level of expectations and their intolerance to artefacts. The development of compression technology has enabled number of applications; key applications in television broadcast field. Compression technology is the basis for digital television. The "Video Compression" book was written for scientists and development engineers. The aim of the book is to showcase the state of the art in the wider field of compression beyond encoder centric approach and to appreciate the need for video quality assurance. It covers compressive video coding, distributed video coding, motion estimation and video quality.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Matei Mancas, Dominique De Beul, Nicolas Riche and Xavier Siebert (2012). Human Attention Modelization and Data Reduction, Video Compression, Dr. Amal Punchihewa (Ed.), ISBN: 978-953-51-0422-3, InTech, Available from: http://www.intechopen.com/books/video-compression/human-attention-modelization-and-data-reduction

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821