Visual Attention and Applications in Multimedia Technologies

Patrick Le Callet, *Member, IEEE,* and Ernst Niebur

***Abstract*—Making technological advances in the field of human-machine interactions requires that the capabilities and limitations of the human perceptual system are taken into account. The focus of this report is an important mechanism of perception, visual selective attention, which is becoming more and more important for multimedia applications. We introduce the concept of visual attention and describe its underlying mechanisms. In particular, we introduce the concepts of overt and covert visual attention, and of bottom-up and top-down processing. Challenges related to modeling visual attention and their validation using *ad hoc* ground truth are also discussed. Examples of the usage of visual attention models in image and video processing are presented. We emphasize multimedia delivery, retargeting and quality assessment of image and video, medical imaging, and the field of stereoscopic 3D images applications.**

***Index Terms*—Visual system, video signal processing, multimedia systems, image analysis, image processing, image communication, image coding, stereo vision**

I. Introduction

SELECTIVE attention is nature's answer to a problem that is present in all but the simplest organisms and increasingly also in machines: information overload. To work efficiently in a variety of complex environments, animals and machines are equipped with an array of sensors, all of which are needed in one situation or another to assure survival of the animal or proper function of the machine. *In any given situation,*, however, only a subset of the sensory input is needed and it would be wasteful (and in many cases practically impossible) to process all sensory input at all times. Therefore, selection has to be made which sensors are relevant at a given time, and only information provided by those is allowed access to central processing resources, Frequently, even the input stream from one sensor may be overwhelmingly rich. For instance, all visual input to the human brain[1] is provided by about 10^6 retinal ganglion cells per eye. Assuming a maximal firing rate of these neurons of about 100 Hz results in a channel capacity of 100 Mbits per second per eye. Indeed, analyses of spike train statistics of visual input to the brain in primates [1], carnivores [2] and insects [3] confirm that the rate of the transmitted information is within an order of magnitude of the channel capacity. This torrent of information cannot be, and does not have to be, processed in detail. Instead, only a fraction of the instantaneously available information is selected for detailed processing while the remainder is discarded.

The filtering process is called *selective attention* and its mechanisms have been studied systematically for well over a century [4]–[6]. The first parallel stages of sensory processing are followed by a bottleneck that restricts the amount of information allowed to proceed to more central processing stages [7], [8]. Information processing in these later stages occurs sequentially rather than in parallel. This allows the application of powerful algorithms to the selected parts of the input that would be too costly to implement for all of sensory input.

For instance, search for a "singleton" target (that is distinguished from distractors by one feature, *e.g.,* by its color) is usually a parallel process (with search times nearly independent of the number of distractors) while search times for "conjunctive" targets (that can be distinguished from distractors only be considering more than one feature, *e.g.,* color *and* orientation) increase linearly with the number of distractors, suggesting a serial search. Treisman and colleagues argue in their Feature Integration Theory [9] that identification of conjunctive targets requires to bind its various features to a coherent object, a task that cannot be performed by elementary feature maps but requires the resources of a more powerful attentional mechanism. This mechanism is not available in parallel for the whole visual field but needs to be applied sequentially. A more differentiated view of visual search has emerged since Treisman's original theory, *e.g.,* refs [10]–[13], but it is generally accepted that visual processing consists of a parallel stage that is fast but relatively simple, followed (if the task requires it) by application of a more powerful mechanism that needs to be applied sequentially to one (or possibly a few) parts of visual input. Exploitation of this limitation of the human visual system is the basis for applications in multimedia which is the topic of this paper.

In section II, we discuss mechanisms of selective selective attention in primate vision and existing computational models. In section III, we focus on some multi-media applications without seeking for exhaustiveness, and we conclude in section IV.

II. Visual attention mechanisms and computational models

In this Section, we introduce detailed computational models of selective attention and some of their limitations. We define two dichotomies, overt *vs.* covert attention in Section II-A and bottom-up *vs.* top-down attention in Section II-B. In section II-C, we briefly discuss difficulties in obtaining ground truth for model predictions.

Patrick Le Callet is with LUNAM Université, Université de Nantes, Institut de Recherche en Communications et Cybernétique de Nantes, Polytech Nantes, UMR CNRS 6597, France email: patrick.lecallet@univ-nantes.fr

Ernst Niebur is with the Solomon Snyder Department of Neuroscience and the Zanvyl Krieger Mind Brain Institute, Johns Hopkins University, Baltimore MD 21218 USA e-mail: niebur@jhu.edu

Manuscript received April 30, 2013.

[1] Although attention controls input from all senses, we focus on vision throughout this article.

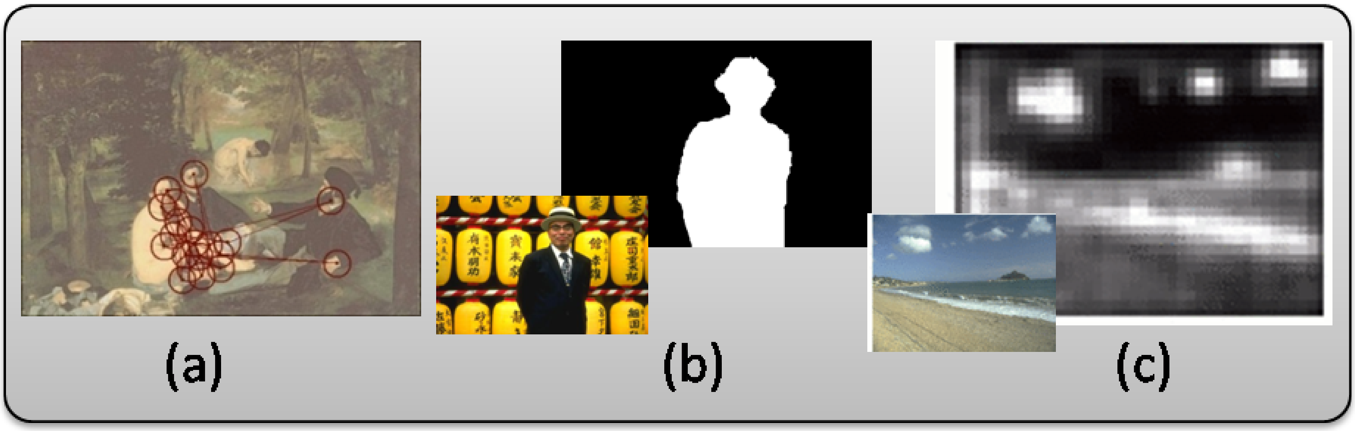

Fig. 1. Examples of scanpath (a), Region of Interest Map and corresponding content (b), Fixation density Map and corresponding content (c).

A. Overt versus covert visual attention

Due to the much higher resolution in the center of the retina compared to its more peripheral regions, humans and other primates usually direct their center of gaze towards the most relevant areas of the visual scene. This generates a series of fixations (and smooth eye movements although the latter are not often discussed in the context of selective attention) called "overt attention," since allocation of the high-resolution resources in the fovea can be easily observed by following the person's eyes, most conveniently and quantifiably with an eye tracker. It has been proposed that far-reaching conclusions can be drawn about the state of the human mind by analyzing the details of this so-called "scan path" [14], [15].

Primates, however, do not have to attend compulsively to objects in their center of gaze. As discovered early on both experimentally [4] and through introspection [6], humans are able to focus their attention to peripheral locations, away from their center of gaze. An illustration of this process is a car driver who fixates the road while simultaneously and covertly monitoring road signs and lights that appear in the retinal periphery. Since this redirection of attention is not visible immediately, it is referred to as covert attention.

There are many experimental paradigms that can determine the movements of the covert focus of attention but none is as convenient, fast, and easy to understand as tracking the eyes of an observer; in other words, measuring his or her overt attentional state. Fortunately, although the locations of overt and covert attention can be dissociated, as discussed, psychophysical evidence shows that an eye movement to a new location is necessarily preceded by focal attention to this locationq [16]–[21]. This makes it possible to easily obtain a close correlate of overt attentional selections by recording eye movements which thus serve as a proxy for shifts of covert attention. Of course, prediction of eye movements is also of immense interest by itself and of great practical interest, including for multimedia applications, Section III. Frequently, models for covert attention are, explicitly or implicitly, used to predict eye movements.

B. Bottom-up versus top-down attention

Attentional selection is a central part of perception and cognition. As such, it is influenced by many factors, both internal and external to the observer. What is attended depends, for instance, on the observer's motivation and the specific task he or she is performing. In a set of classic experiments, Yarbus [22] showed that eye movements (overt attention) of the same observer viewing the same visual scene differ dramatically depending on what information the observer is looking for in the scene. Attentional selection that depends on the internal state of the observer is referred to as "top-down attention." It is very difficult to develop biologically realistic detailed models of such mechanisms which may include influences such as the personal history of the observer.

On the other hand, "bottom-up" selection only depends on the visual input provided instantaneously or in the very recent past (as in immediately preceding frames of a movie). As such, it is not only much easier to control but it is also easier to quantify the correlation between input and resulting behavior. For this reason, Koch and Ullman [23] proposed that bottom-up attention is a suitable candidate for detailed computational models of selective attention. Specifically, they proposed that bottom-up attention is directed to salient parts of the visual scene and they proposed the concept of a saliency map. This is a topographic map of the visual field whose scalar value is the saliency at the respective location. Saliency is computed at multiple scales from the local differences in visual submodalities (color, orientation, . . .). If both the basic premise that bottom-up attention is attracted by salience as well as their concept how salience is computed are correct, attentional control is then reduced to finding the local maxima in the saliency map and assigning the successively visited foci of attention to those maxima in order of decreasing peak value. This results in a "covert attentional scan path,", see Figure 1 for an illustrative example, in analogy to the sequence of eye movements in overt attention.

This conceptual idea of attentional control by a saliency map was subsequently implemented in biologically realistic computational models [24]–[26]. Over the last decade and a half, these models have been refined, tested and applied by a large number of groups. Borji and Itti [27] provide an excellent overview of the current state-of-the-art of visual attention modeling including a taxonomy of models (information theoretical, cognitive, graphical, spectral, pattern classification, Bayesian, . . .).

The simplicity of the original saliency map model makes it attractive both conceptually as well as for applications but it also engenders limitations. For instance, it has been found that eye movements are typically oriented towards the centers of objects, rather than their borders which is where bottom-up saliency peaks [28]. Such deviations can be explained, at the cost of slightly higher complexity, by directing attention to proto-objects, rather than purely spatially defined regions of the visual scene [28], [29].

While bottom-up influences are thus important, it is clear that in many situations top-down attention plays a role, too. One consequence of the saliency map model is that its first selections in a new scene should agree better with observed eye movements than later ones, since less top-down guidance is expected to exist for input never seen before; this was confirmed experimentally [30]. It is also important to distinguish the term “salience” and “importance” (as in, *e.g.,* Region of Interest/Importance, RoI) which are frequently considered synonyms in the signal processing literature. While both visual salience and visual importance denote the most visually “relevant” parts of the scene, it is useful to reserve the term “salience” for strictly bottom-up influences, while “important” areas can be selected based on both bottom-up and top-down criteria. The two mechanisms are thus driven by a different combination of sources. The interplay between these mechanisms has been studied showing that their relationship might vary along viewing time [31].

Even though top-down influences play an important role in attentional selection, we have already discussed that developing computational models of top-down attention in as much detail as for bottom-up attention is virtually impossible. Some progress has been made for parts of the general problem, for instance for finding objects [32]. This field must be considered, however, as being in its infancy. For instance, it is known that not only the properties of objects and their immediate surrounds but the interaction between objects on the image scale as well as the “gist” of the scene [33], [34] strongly influence search patterns and response times [35]. On the other hand, it was shown that low-level saliency is significantly predictive not only of eye movements but, surprisingly, even for conscious decisions of what observers consider interesting [36]. The fact that the very simple quantities computed in the original saliency map [24], [25] significantly influence human behavior after conscious deliberation and after many seconds of response time engenders hope that these easily and cheaply computed models and their derivatives can be useful for technical applications, even when humans are “in the loop,” as in multi-media applications.

C. *Visual attention models and ground truth*

Developing and testing computational models of visual attention depends on the availability of ground truth. Many studies rely on fixation density maps (FDM) generated from eye-tracking experiments (see Figure 2 for an illustration of the process leading to the generation of FDM). Consequently, most of the models are supposed to address mainly overt visual attention. Nevertheless, recommendations to properly generate FDM are still missing. Several eye-tracking FDM databases have been made publicly available corresponding to experiments conducted independently in different conditions. The question of corresponding viewing time is particularly critical regarding top-down and bottom-up competition, while rarely considered. How the difference between various experimental set up to obtain FDM may impact image processing applications has been recently investigated [37].

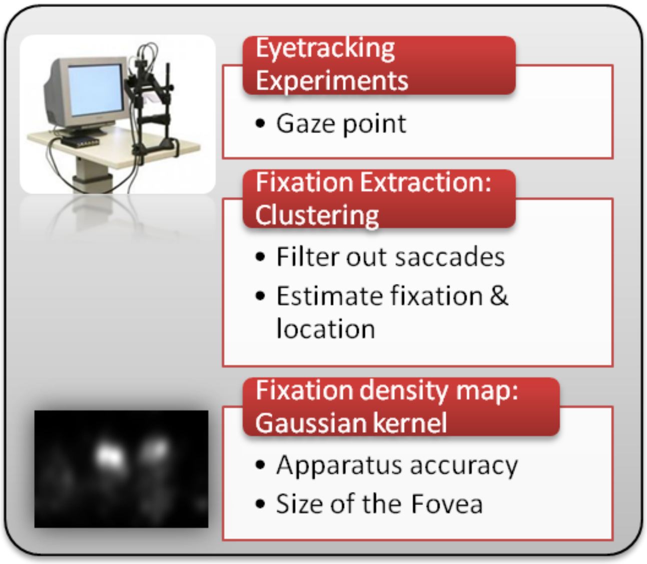
Eyetracking Experiments
• Gaze point
Fixation Extraction: Clustering
• Filter out saccades
• Estimate fixation & location
Fixation density map: Gaussian kernel
• Apparatus accuracy
• Size of the Fovea

Fig. 2. Steps for transforming eye-tracking data into a Fixation Density Map. After gathering the raw data (top), saccades are identified and fixation locations are determined (center). The fixation map is then obtained by convolving fixation locations with a Gaussian whose size is determined by a combination of mean eye tracking error and the size of the human fovea (bottom).

Computational models of attention produce very different predictions for FDM (see examples in Figure 3). How to quantitatively compare the performance of different models given the ground truth is another topic of research, see ref. [38] for a recent study proposing several metrics to assess model performance. It should also be noted that using FDM as ground truth may not be warranted for all models since some are designed to explain aspects of visual attention mechanisms that are not reflected in FDM.

Given these caveats, the usage of a given visual attention models should be achieved cautiously in image processing applications, considering the model type (e.g: top-down vs bottom-up) but also its performance regarding a given application context. Better characterization of a model should lead to comprehensive recommendation of proper usage.

III. Applications of Visual attention models in image and video processing

In the following, we give an overview of applications of models of visual selective attention. In Sections III-A

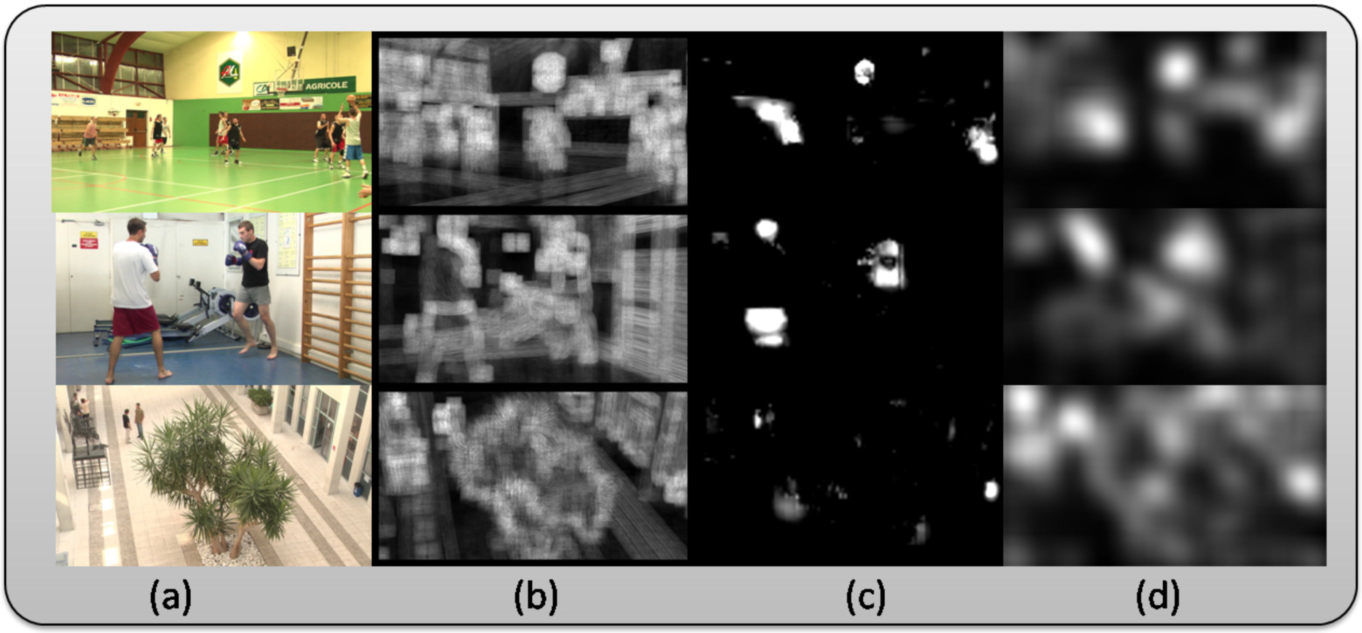

Fig. 3. Examples of FDM generated by visual attention models: original content (a), AIM model [?] (b), STB model [28] (c) , SR model [39] (d)

and III-B, we discuss multimedia delivery. Section III-C is devoted to re-targeting, Section III-D to quality assessment, and Section III-E to applications in medical imaging. Finally in Section III-F we discuss stereoscopic 3D images.

A. *Multimedia Delivery: improving source coding*

Several stages of the media delivery chain can benefit from insights into visual attention mechanisms. The first attempts were applied to selective compression of image and video contents. A survey on this topic can be found in [40]. Sensitivity and resolution reduction of the human visual system as a function of eccentricity is one the that could benefits to improve compression performance once salient location identified [43]. Selective compression is based on two priors: a prior of selection, that defines the most informative areas of an image, and a prior of compression that defines the coding nature and bit rate allocation strategy. Compression rate (prior of coding), and consequently the visual quality, can be differentially adapted to different image areas depending on the level of attention devoted to them by the human observers (prior of selection). The importance of a given image region can be computed based on the contribution of different features (contrast in color, orientation, intensity, . . .) [24], [41], [42] or in a simplified version under the assumption that human faces attract attention [43]. There are two principle approaches to prioritize coding of different image areas using saliency information. The first is the indirect approach [44] in which the graphical contents is pre-processed. Image agreas are selectively encoded according to their saliency, *e.g.,* by low-pass filtering less important regions. The choice of preprocessing methods needs to be compatible with the coding scheme, especially with the quantization operator.

The direct approach is applied in block based coding methods. Bit rates are allocated to each macro block separately according to a visual saliency criterion (see Figure 4). Most of the time, this is achieved by changing the quantization parameters. This can be done using conventional RDO (Rate Distortion Optimization) techniques [45]–[48] or by providing a map based on a preceding analysis of the contents [49].

With the recent availability of low-cost, consumer-grade eye trackers, visual attention-based bit allocation techniques for network video streaming have been introduced [50]. To improve the efficacy of such gaze-based networked systems, gaze prediction strategies can be used to predict future gaze locations to lower the end-to-end reaction delay due to the finite round trip time (RTT) of transmission networks. Feng *et al.* [50] demonstrated that the bit rate can be reduced by slightly more than 20% without noticeable visual quality degradation even when end-to-end network delays ares as high as 200ms.

In another approach [51], the audio component is also taken into account to improve RoI encoding based on the observation that sound-emitting regions in an audio-visual sequence typically draw a viewer’s attention.

B. *Multimedia Delivery: Improving Resilience to Transmission Errors*

Packets in a video bitstream contain data with different levels of importance from the visual information point of view. This results in unequal amounts of perceived image quality degradation when these packages are lost. Quality assessment experiments with observers have demonstrated that the effect of a lost packet depends on the spatio-temporal location of the visual information coded in the packet. Perceived quality degradation is lowest when the loss affects regions of “non interest” [52]–[54]. Visual attention based error resilience or RoI based channel coding methods are consequently good candidates to attenuate the perceptual quality loss resulting

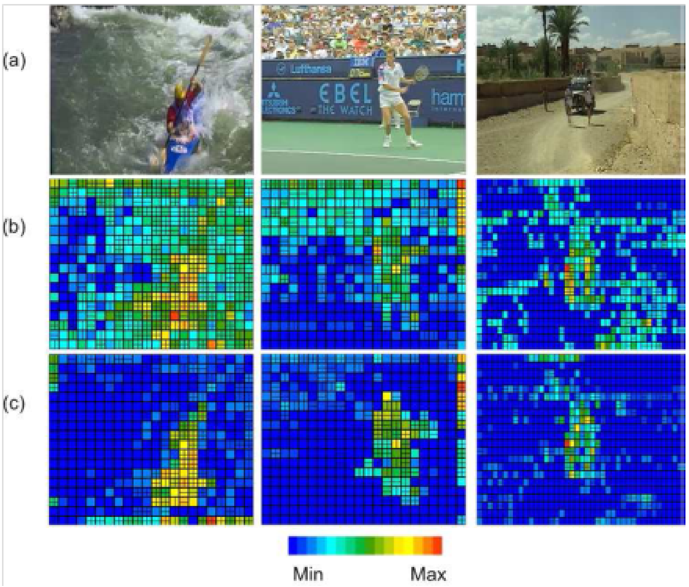

Fig. 4. Distribution of encoding cost of natural scenes (shown in a) for a conventional H.264 coding (b) and a saliency based approach (c) (from O. Le Meur, P. Le Callet, D. Barba Selective H.264 video coding based on a saliency map, http://people.irisa.fr/Olivier.Le_Meur). Color coded pixels show the cost in the respective areas. The color scale at the bottom is common for all panels in rows b and c.

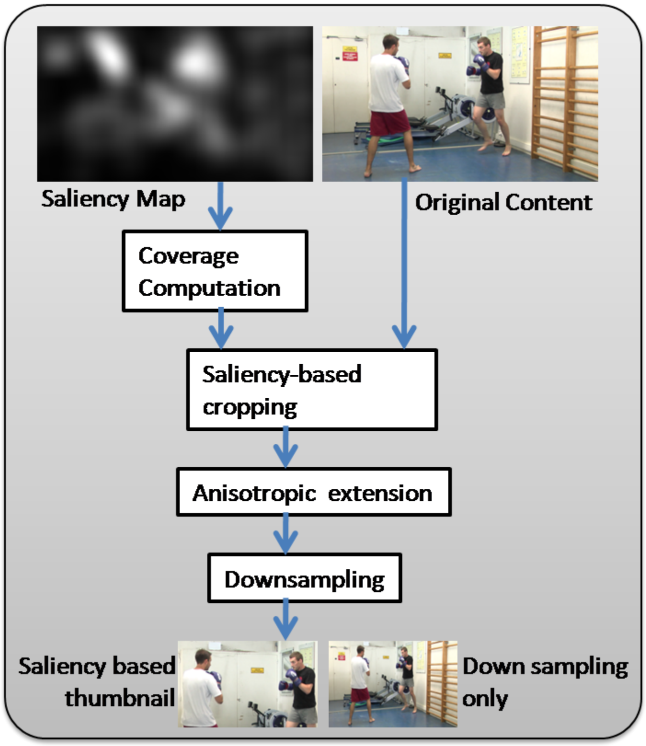

Fig. 5. Process of saliency-based reframing [58]. The saliency-based thumbnail focuses on the most relevant image parts.

from packet loss. In the context of highly prediction based coding technologies such as H.264/AVC, for good compression performance there is a high dependency between many parts of the coded video sequence. However, this dependency comes with the drawback of allowing a spatio-temporal propagation of the error resulting from a packet loss. RoI based coding should also consider attenuating the effect of this spatio-temporal dependency when important parts of the bitstream are lost. As part of the H.264/AVC video coding standard, error resilience features such as Flexible Macroblock Ordering (FMO) and Data Partitioning (DP) can be exploited to improve resilience of salient regions of video content. DP partitions code slice into three separate NAL (Network Abstract Layer) units, containing each different part of the slice. FMO allows the ordering of macroblocks in slices according to a predefined map rather than using the usual raster scan order. Coupled with RoI-based coding, FMO is can be used to gather RoI macroblocks into a single slice [55]. An alternative approach [56] consist in confining the RoI in separate slices to prevent error propagation within a picture and then constraining the coding prediction process in the RoIs to avoid that the resulting loss distortion reaching RoIs in other pictures.

C. *Image and Video retargeting*

With the recent explosion of commonly available device types (tablet, smart phone, large displays, . . .), formats (3D, HD, Ultra HD, . . .) and services (video streaming, image database browsing, . . .), the visual dimension of multimedia contents viewed by a human observer can vary enormously, resulting in the stimulation of very different fractions of his or her visual field. Depending on display capacity and the purpose of the application, contents often need to be repurposed to generate smaller versions, with respect to image size, resolution, frame rate, ... A common way to achieve this goal is to dramatically down-sample the picture homogeneously, as in thumbnail modes. This often yields poorly rendered pictures since important objects of the scene may be no longer recognizable. Alternatively, content repurposing techniques perform content-aware image resizing, for example by seam carving [57]. Saliency based image re-targeting (or content repurposing or reframing techniques) algorithms have been proposed following this idea: identify important regions of interest and compute the reduced picture centered on these parts [58], [59] (see figure 5 for an illustration). More recently, dynamic (*i.e.* time changing) thumbnails have been introduced using a dynamic computational model of visual attention [60]. Rubinstein and colleagues [61] have evaluated many image re-targeting algorithms both objectively and subjectively and demonstrated the value of saliency based cropping approaches.

D. *Image and Video quality assessment*

Perceptual objective image quality assessment uses an algorithm that evaluates the quality of pictures or video as a human observer would do based on the properties of the human visual system. Visual attention is one of the features that can be considered based on the rationale that an artifact is likely more annoying in a salient region than in other areas [62]. Most of objective quality assessment methods can be decomposed in two steps. Image distortion is first locally (pixel-base, block-based, . . .) evaluated resulting in a distortion map. In the second step, a pooling function is used to combine the distortion map values into a single quality score value. An intuitive idea to improve quality assessment methods using visual attention information is to give greater weight at the pooling stage to degradation appearing in salient areas than in non-salient areas [63], [64]. Initial approaches consisted in weighting the distortion map using local saliency values before computing a linear or non linear mean. More recent studies, based on eye tracking data, demonstrated that this simple weighting is not very effective [65], [66] in the case of compression artifacts. Nevertheless, such approaches can lead to significantly improved performance in the case of non-uniformly located distortions such as those due to transmission impairments [67]. Alternative weighting methods have been introduced for compression artifacts with varying success [68], [69]. In ref. [70], more complex combinations of saliency map and distortion are introduced, assuming that weights should be a function of both saliency value and distortion level. You *etal* [71], [72] revisit the problem at the distortion level for video content. Distortion visibility can be balanced according to the human contrast sensitivity function. As the latter is spatially non uniform, gaze estimation should be considered to properly apply it.

Another open issue is which parts of the original content and its distorted version should be used for estimating the saliency map. Artifacts themselves may affect the deployment of visual attention; they may, for instance, attract attention [73]. Moreover, objective quality measures are expected to correlate with the outcomes of quality assessment experiments performed by observers. To obtain comparison data, observers need to perform specific tasks. Such tasks are likely to affect the visual attention deployment compared to a free-viewing [74]–[76].

E. Medical imaging

Over the past twenty years, digital medical imaging techniques (Computed Tomography, Magnetic Resonance Imaging, Ultrasound, Computed Radiography/Digital Radiography, Fluoroscopy, Positron Emission Tomography, Single Photon Emission Computed Tomography, . . .) have revolutionized healthcare practice, becoming a core source of information for clinicians to render diagnostic and treatment decisions. Practical analysis of medical images requires two basic processes: visually inspecting the image (involving visual perception processes, including detection and localization tasks), and performing an interpretation (requiring cognitive processes). Unfortunately, interpretation is not error-free and can be affected by the observer's level of expertise and by technological aspects. Moreover, a side effect of the dramatic increase in the availability and use of medical images is a shortage of qualified image reading experts. It is likely that the time per image that is available for interpretation will continue to decrease in the future. Expertise in medical image reading therefore needs to be seen under the two aspects: accuracy and speed [77]. Understanding how clinicians read images, how they develop expertise throughout their careers, and why some people are better at interpreting medical images than others are crucial questions that are related to visual attention.

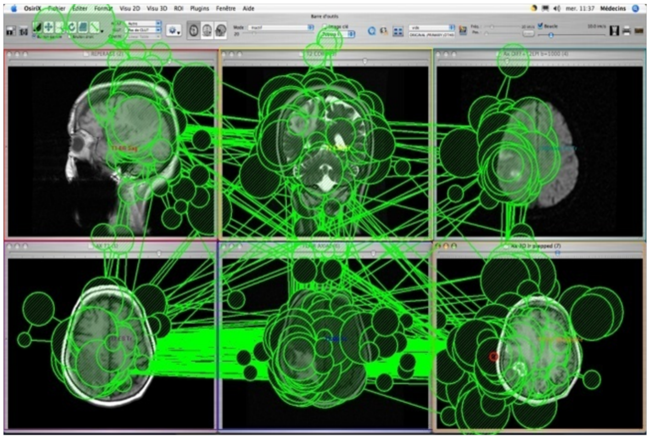

Fig. 6. Scanpath and gaze fixations on multiple MRI sequences. Shown are different MRI sequences (gray) taken from one patient's head, overlaid with eye movement data of a clinical expert (green). Lines are saccades and shaded circles signify fixations, with the diameter proportional to viewing time. It is seen that the image reader uses several sequences, implying comparison of different source of information from the different representations in several panels.

Such knowledge represents great potential to develop better training programs and create new tools that could enhance and speed up the learning process. A longitudinal study [77] of pathology residents during their development of expertise in reading slides of breast biopsies used eye tracking experiments at the beginning of each of their three years of residency, documenting changes of their scan paths as they increased the level of their experience. The data showed that search patterns changed with each successive year of experience. Over time, residents spent significantly less time per slide, made fewer fixations, and performed less examination of non-diagnostic areas. Similar findings have been obtained in radiology on multi-slice images such as Computer Tomography scans (CCT) [78] or multi sequences Magnetic Resonance Imaging (MRI) [79]. Figure 6 shows an example of scanpath and gaze fixations in the case of multiple MRI sequences.

F. Stereoscopic 3D: new opportunities for visual attention

A key factor required for the wide-spread adoption of services based on stereoscopic images will be the creation of a compelling visual experience for the end-user. Perceptual issues and the importance of considering 3D visual attention to improve the overall 3D viewing experience in 3DTV broadcasting have been discussed extensively [80]. Integrating visual attention at source and channel coding level represents limited adaption compared to 2D case. More interestingly, content production offers new original opportunities to make use of insights in visual attention mechanisms, especially dealing with perceptual concept such as visual comfort. Comfortable viewing conditions, *e.g.,* zone of comfortable viewing, of stereoscopic content is linked to several factors such as accommodation-vergence conflict, range of depth of focus and range of fusion [81], [82]. A seminal study by Wopking [83] suggests that visual discomfort increases with high spatial frequencies and disparities, partially because the limits of stereoscopic fusion increase as a result of the decreased spatial frequency. More generally, it appears that blurring can have a positive impact on visual comfort because it reduces the accommodation-vergence conflict, limiting both the need for accommodation and the effort to fuse [84], [85]. Simulating depth-of-field (DOF) is a way to take advantage of the retinal defocusing property in order to improve visual comfort, by artificially blurring images to a degree that corresponds to the relative depth from fixated objects. As reported by Lambooij *et al.* [86], "three essential steps are required for proper implementation of a simulated DOF: localization of the eye positions, determination of the fixation point and implementation of blur filters to non-fixated layers." This procedure has been applied in virtual reality environments but has drawbacks in more general contexts since it affects depth cue integration between retinal disparity and areas with high amounts of blur [87]. Blurring effects can also be used for 3D content to direct the viewer's attention towards a specific area of the image that could meet a comfortable viewing zone. In gaming and in the computer graphics community, visual attention modeling has attracted a growing interest. Visual attention models have been used to produce a more realistic behavior of a virtual character, to improve interactivity in 3D virtual environments, and to improve visual comfort when viewing rendered 3D virtual environments [88]–[90].

Due to geometry issues related to depth rendering, adaptation from a cinema environment to the home environment is far from being an automatic, straightforward process for 3D content production. Automated content-based post-production or post-processing tools to help adapt 3D content to television are expected to be developed. 3D visual attention models can be employed to provide the area of interest and convergence plane to drive the content repurposing of stereoscopic content. In addition, the adaptation of the scene depth can be used to improve visual comfort. To reduce both visual discomfort and fatigue, the convergence plane is usually continuously set to the main area of interest, as the latter is moving across different depth levels. A way to reduce eye strain is to modify the convergence plane of the main area of interest to place it on the display plane, *i.e.,* by adapting the content disparity. Such visual attention based adaptive rendering of 3D stereoscopic video has been proposed using a 2D visual attention model [91].

IV. Conclusion

Visual attention is attracting a high level of interest in the vision science community. In this paper, we have demonstrated that this research interest is highly penetrating the Information and Communication Technology (ICT) field with some successful outcomes although there are still challenges ahead. One caveat is that, as in any trans-disciplinary approach, one has to assure that concepts from one research field are properly used when appropriated by another. For instance, in the image processing community, the terms "salience" and "importance" (or Visual Salience and Region of Interest/Importance) have sometimes been considered synonymous, while, as stated, they should be distinguished. Both denote the most visually "relevant" parts of the scene. However, the concepts differ as they may refer to two different mechanisms of visual attention: bottom-up *vs.* top-down. While the interaction between ICT and vision science is intensifying, the ICT community needs to assure carefully that the proper tools (models, validation protocols, databases, . . .) are used for the proper needs.

Acknowledgment

Work supported in part by Office of Naval Research grant N000141010278 and NIH grant R01EY016281.

References

[1] D. S. Reich, F. Mechler, K. P. Purpura, and J. D. Victor, "Interspike intervals, receptive fields, and information encoding in primary visual cortex," *J. Neurosci.*, vol. 20, no. 5, pp. 1964–74, March 2000.

[2] P. Reinagel and R. C. Reid, "Temporal coding of visual information in the thalamus," *J. Neurosci.*, vol. 20, no. 14, pp. 5392–400, Jul 2000.

[3] N. Brenner, S. P. Strong, R. Koberle, W. Bialek, and R. R. d. R. v. Steveninck, "Synergy in a neural code," *Neural Computation*, vol. 12, no. 7, pp. 1531–1552, 2000.

[4] H. von Helmholtz, *Handbuch der physiologischen Optik*. Leipzig: Voss, 1867.

[5] W. M. Wundt, *Grundzüge de physiologischen Psychologie*. W. Engelman, 1874.

[6] W. James, *The Principles of Psychology*. New York: Henry Holt, 1890.

[7] D. E. Broadbent, *Perception and Communication*. London: Pergamon, 1958.

[8] U. Neisser, *Cognitive psychology*. New York: Appleton-Century-Crofts, 1967.

[9] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[10] Z. J. He and K. Nakayama, "Surfaces versus features in visual search," *Nature*, vol. 359, pp. 231–233, 1992.

[11] J. M. Wolfe, "Guided search 2.0 – a revised model of visual search," *Psychonomics Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.

[12] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo, "Modelling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, October 1995.

[13] J. Wolfe and T. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nat. Rev. Neurosci.*, vol. 5, pp. 495–501, Jun 2004.

[14] D. Noton and L. Stark, "Scanpaths in eye movements," *Science*, vol. 171, pp. 308–311, 1971.

[15] W. H. Zangemeister, K. Sherman, and L. Stark, "Evidence for global scanpath strategy in viewing abstract compared with realistic images," *Neuropsychologia*, vol. 33, no. 8, pp. 1009–10 025, 1995.

[16] M. Shepherd, J. M. Findlay, and R. J. Hockey, "The relationship between eye movements and spatial attention," *The Quarterly Journal of Experimental Psychology*, vol. 38, no. 3, pp. 475–491, 1986.

[17] W. X. Schneider and H. Deubel, "Visual attention and saccadic eye movements: Evidence for obligatory and selective spatial coupling," *Studies in Visual Information Processing*, vol. 6, pp. 317–324, 1995.

[18] H. Deubel and W. X. Schneider, "Saccade target selection and object recognition: Evidence for a common attentional mechanism," *Vision research*, vol. 36, no. 12, pp. 1827–1837, 1996.

[19] J. Hoffman and B. Subramaniam, "The role of visual attention in saccadic eye movements," *Perception and Psychophysics*, vol. 57, no. 6, pp. 787–795, 1995.

[20] E. Kowler, E. Anderson, B. Dosher, and E. Blaser, "The role of attention in the programming of saccades," *Vision Research*, vol. 35, no. 13, pp. 1897–1916, 1995.

[21] R. M. McPeek, V. Maljkovic, and K. Nakayama, "Saccades require focal attention and are facilitated by a short-term memory system," *Vision research*, vol. 39, no. 8, pp. 1555–1566, 1999.

[22] A. Yarbus, *Eye Movements and Vision*. New York: Plenum Press, 1967.

[23] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, pp. 219–227, 1985.

[24] E. Niebur and C. Koch, "Control of selective visual attention: Modeling the "where" pathway," in *Advances in Neural Information Processing Systems*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996, vol. 8, pp. 802–808.

[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based fast visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[26] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Neuroscience*, vol. 2, pp. 194–203, 2001.

[27] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[28] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, pp. 1395–1407, Nov 2006.

[29] S. Mihalas, Y. Dong, R. von der Heydt, and E. Niebur, "Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects," *Proceedings of the National Academy of Sciences*, vol. 108, no. 18, pp. 7583–8, 2011, pMCID: PMC3088583.

[30] D. Parkhurst, K. Law, and E. Niebur, "Modelling the role of salience in the allocation of visual selective attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[31] J. Wang, D. M. Chandler, and P. Le Callet, "Quantifying the relationship between visual salience and visual importance," in *Proceedings of SPIE*, vol. 7527, Feb. 2010, pp. 75 270K–75 270K–9. [Online]. Available: http://spiedigitallibrary.org/proceedings/resource/2/psisdg/7527/1/75270K_1?isAuthorized=no

[32] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimal object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, Jun 2006, bu ; cv ; td, pp. 2049–2056.

[33] A. Torralba and A. Oliva, "Statistics of natural image categories," *Network: Computation in Neural Systems*, vol. 14, pp. 391–412, 2003.

[34] A. Torralba, A. Oliva, M. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object selection," *Psychological Review*, vol. 113, no. 4, pp. 766–86, 2006.

[35] J. M. Wolfe, G. A. Alvarez, R. Rosenholtz, Y. I. Kuzmova, and A. M. Sherman, "Visual search for arbitrary objects in real scenes," *Attention, Perception, & Psychophysics*, vol. 73, no. 6, pp. 1650–1671, 2011.

[36] C. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur, "Everyone knows what is interesting: Salient locations which should be fixated," *Journal of Vision*, vol. 9, no. 11, pp. 1–22, October 2009.

[37] U. Engelke, H. Liu, J. Wang, P. Le Callet, I. Heynderickx, H. Zepernick, and A. Maeder, "A comparative study of fixation density maps," *IEEE Transactions on Image Processing*, 2012.

[38] O. L. Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior Research Methods*, pp. 1–16, 2012. [Online]. Available: http://link.springer.com/article/10.3758/s13428-012-0226-9

[39] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, 2007, pp. 1–8.

[40] J.-S. Lee and T. Ebrahimi, "Perceptual video compression: A survey," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 684–697, 2012.

[41] A. Maeder, J. Diederich, and E. Niebur, "Limiting human perception for image sequences," *Proceedings of the SPIE*, vol. 2657, pp. 330–337, 1996.

[42] D. Parkhurst and E. Niebur, "Variable resolution displays: a theoretical, practical and behavioral evaluation," *Human Factors*, vol. 44, no. 4, pp. 611–29, 2002.

[43] S. Daly, K. Matthews, and J. Ribas-Corbera, "As plain as the noise on your face: Adaptive video compression using face detection and visual eccentricity models," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 30–46, 2001. [Online]. Available: +http://dx.doi.org/10.1117/1.1333679

[44] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304 –1318, Oct. 2004.

[45] A. Eleftheriadis and A. Jacquin, "Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit-rates," *Signal Processing: Image Communication*, vol. 7, no. 3, pp. 231 – 248, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/092359659500028U

[46] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, "Low bit-rate coding of image sequences using adaptive regions of interest," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 8, pp. 928 –934, Dec. 1998.

[47] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, Nov. 2009, pp. 3417 –3420.

[48] Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communication of H.264/AVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 1, pp. 134 –139, Jan. 2008.

[49] C.-W. Tang, C.-H. Chen, Y.-H. Yu, and C.-J. Tsai, "A novel visual distortion sensitivity analysis for video encoder bit allocation," in *2004 International Conference on Image Processing, 2004. ICIP '04*, vol. 5, Oct. 2004, pp. 3225 – 3228 Vol. 5.

[50] Y. Feng, G. Cheung, W.-t. Tan, and Y. Ji, "Hidden markov model for eye gaze prediction in networked video streaming," in *2011 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2011, pp. 1 –6.

[51] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Efficient video coding based on audio-visual focus of attention," *Journal of Visual Communication and Image Representation*, vol. 22, no. 8, pp. 704–711, Nov. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S104732031000146X

[52] F. Boulos, B. Parrein, P. Le Callet, D. Hands *et al.*, "Perceptual effects of packet loss on h. 264/AVC encoded videos," in *Fourth International workshop on Video Processing and Quality Metrics for consumer electronics, VPQM*, 2009.

[53] H. Boujut, J. Benois-Pineau, O. Hadar, T. Ahmed, and P. Bonnet, "Weighted-MSE based on saliency map for assessing video quality of h.264 video streams," in *Proc. SPIE 7867, Image Quality and System Performance VIII, 78670X*, Jan. 2011, pp. 78 670X–78 670X. [Online]. Available: http://dx.doi.org/10.1117/12.876471

[54] U. Engelke, R. Pepion, P. Le Callet, and H.-J. Zepernick, "Linking distortion perception and visual saliency in H.264/AVC coded video containing packet loss," in *Proceedings of SPIE*, vol. 7744, Jul. 2010, pp. 774 406–774 406–10. [Online]. Available: http://spiedigitallibrary.org/proceedings/resource/2/psisdg/7744/1/774406_1?isAuthorized=no

[55] Y. Dhondt, P. Lambert, and R. Van de Walle, "A flexible macroblock scheme for unequal error protection," in *2006 IEEE International Conference on Image Processing*, Oct. 2006, pp. 829 –832.

[56] F. Boulos, W. Chen, B. Parrein, and P. Le Callet, "Region-of-interest intra prediction for h. 264/AVC error resilience," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, 2009, p. 3109–3112.

[57] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, Jul. 2007. [Online]. Available: http://doi.acm.org/10.1145/1276377.1276390

[58] O. Le Meur, X. Castellan, P. Le Callet, and D. Barba, "Efficient saliency-based repurposing method," in *2006 IEEE International Conference on Image Processing*, Oct. 2006, pp. 421 –424.

[59] V. Setlur, T. Lechner, M. Nienhaus, and B. Gooch, "Retargeting images and video for preserving information saliency," *IEEE Computer Graphics and Applications*, vol. 27, no. 5, pp. 80 –88, Oct. 2007.

[60] M. P. Da Silva, V. Courboulay, and P. Le Callet, "Real time dynamic image re-targeting based on a dynamic visual attention model," in *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, Jul. 2012, pp. 653 –658.

[61] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," in *ACM SIGGRAPH Asia 2010 papers*, ser. SIGGRAPH ASIA '10. New York, NY, USA: ACM, 2010, p. 160:1–160:10. [Online]. Available: http://doi.acm.org/10.1145/1866158.1866186

[62] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 50 –59, Nov. 2011.

[63] W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating higher level perceptual factors," in *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*. IEEE, 1998, pp. 414–418.

[64] R. Barland and A. Saadane, "Blind quality metric using a perceptual importance map for JPEG-20000 compressed images," in *2006 IEEE International Conference on Image Processing*, Oct. 2006, pp. 2941 –2944.

[65] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barbba, "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric," in *IEEE International Conference on Image Processing, 2007. ICIP 2007*, vol. 2, Oct. 2007, pp. II –169 –II –172.

[66] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971 –982, Jul. 2011.

[67] U. Engelke, M. Barkowsky, P. Le Callet, and H.-J. Zepernick, "Modelling saliency awareness for objective video quality assessment," in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, Jun. 2010, pp. 212 –217.

[68] E. Larson, C. Vu, and D. Chandler, “Can visual fixation patterns improve image fidelity assessment?” in *15th IEEE International Conference on Image Processing, 2008. ICIP 2008*, 2008, pp. 2572–2575.

[69] J. You, A. Perkis, and M. Gabbouj, “Improving image quality assessment with modeling visual attention,” in *2010 2nd European Workshop on Visual Information Processing (EUVIP)*, Jul. 2010, pp. 177 –182.

[70] J. Redi, H. Liu, P. Gastaldo, R. Zunino, and I. Heynderickx, “How to apply spatial saliency into objective metrics for JPEG compressed images?” in *2009 16th IEEE International Conference on Image Processing (ICIP)*, Nov. 2009, pp. 961 –964.

[71] J. You, L. Xing, A. Perkis, and T. Ebrahimi, “Visual contrast sensitivity guided video quality assessment,” in *2012 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2012, pp. 824 –829.

[72] J. You, J. Korhonen, and A. Perkis, “Attention modeling for video quality assessment: Balancing global quality and local quality,” in *2010 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2010, pp. 914 –919.

[73] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, “Do video coding impairments disturb the visual attention deployment?” *Signal Processing: Image Communication*, vol. 25, no. 8, p. 597–609, 2010.

[74] A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, and A. Tirel, “Task impact on the visual attention in subjective image quality assessment,” in *Proceedings of European Signal Processing Conference*, France, Sep. 2006, p. invited paper. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00342685

[75] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, “Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric,” *Signal Processing: Image Communication*, vol. 25, no. 7, p. 547–558, 2010.

[76] J. Redi, H. Liu, R. Zunino, and I. Heynderickx, “Interactions of visual attention and quality perception,” in *Proc. SPIE 7865, Human Vision and Electronic Imaging XVI, 78650S*, Feb. 2011, pp. 78 650S–78 650S. [Online]. Available: http://dx.doi.org/10.1117/12.876712

[77] E. A. Krupinski, “On the development of expertise in interpreting medical images,” in *Proc. SPIE 8291, Human Vision and Electronic Imaging XVII, 82910R*, Feb. 2012, pp. 82 910R–82 910R. [Online]. Available: http://dx.doi.org/10.1117/12.916454

[78] A. Venjakob, T. Marnitz, J. Mahler, S. Sechelmann, and M. Rötting, “Radiologists’ eye gaze when reading cranial CT images,” in *Proc. SPIE 8318, Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment, 83180B*, Feb. 2012, pp. 83 180B–83 180B. [Online]. Available: http://dx.doi.org/10.1117/12.913611

[79] C. Cavaro-Ménard, J.-Y. Tanguy, and P. Le Callet, “Eye-position recording during brain MRI examination to identify and characterize steps of glioma diagnosis,” in *Proceedings of SPIE*, vol. 7627, Mar. 2010, pp. 76 270E–76 270E–8. [Online]. Available: http://spiedigitallibrary.org/proceedings/resource/2/psisdg/7627/1/76270E_1?isAuthorized=no

[80] Q. Huynh-Thu, M. Barkowsky, and P. Le Callet, “The importance of visual attention in improving the 3D-TV viewing experience: Overview and new perspectives,” *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 421 –431, Jun. 2011.

[81] S. Pastoor, “Human factors of 3D displays in advanced image communications,” *Displays*, vol. 14, no. 3, pp. 150–157, Jul. 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0141938293900365

[82] S. Nagata, “The binocular fusion of human vision on stereoscopic displays— field of view and environment effects,” *Ergonomics*, vol. 39, no. 11, pp. 1273–1284, 1996, PMID: 8888639. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00140139608964547

[83] M. Wopking, “Viewing comfort with stereoscopic pictures: An experimental study on the subjective effects of disparity magnitude and depth of focus,” *Journal of the Society for Information Display*, vol. 3, no. Nr.3, pp. 101–103, 1995.

[84] J. L. Semmlow and D. Heerema, “The role of accommodative convergence at the limits of fusional vergence.” *Investigative Ophthalmology & Visual Science*, vol. 18, no. 9, pp. 970–976, Jan. 1979. [Online]. Available: http://www.iovs.org/content/18/9/970

[85] K. Talmi and J. Liu, “Eye and gaze tracking for visually controlled interactive stereoscopic displays,” *Signal Processing: Image Communication*, vol. 14, no. 10, pp. 799–810, Aug. 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596598000447

[86] M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn, “Visual discomfort and visual fatigue of stereoscopic displays: A review,” *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 30 201–1–30 201–14, 2009.

[87] G. Mather and D. R. Smith, “Depth cue integration: stereopsis and image blur,” *Vision Research*, vol. 40, no. 25, pp. 3501–3506, Jan. 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0042698900001784

[88] S. Hillaire, A. Lecuyer, R. Cozot, and G. Casiez, “Using an eye-tracking system to improve camera motions and depth-of-field blur effects in virtual environments,” in *IEEE Virtual Reality Conference, 2008. VR ’08*, Mar. 2008, pp. 47 –50.

[89] S. Hillaire, A. Lécuyer, G. Breton, and T. R. Corte, “Gaze behavior and visual attention model when turning in virtual environments,” in *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology*, ser. VRST ’09. New York, NY, USA: ACM, 2009, p. 43–50. [Online]. Available: http://doi.acm.org/10.1145/1643928.1643941

[90] S. Hillaire, A. Lecuyer, R. Cozot, and G. Casiez, “Depth-of-field blur effects for first-person navigation in virtual environments,” *IEEE Computer Graphics and Applications*, vol. 28, no. 6, pp. 47 –55, Dec. 2008.

[91] C. Chamaret, S. Godeffroy, P. Lopez, and O. Le Meur, “Adaptive 3D rendering based on region-of-interest,” in *SPIE Stereoscopic Displays and Applications XXI*, vol. 7524. SPIE, Feb. 2010. [Online]. Available: http://dx.doi.org/10.1117/12.837532

Patrick Le Callet received both an M.Sc. and a PhD degree in image processing from Ecole polytechnique de l’Université de Nantes. He was also a student at the Ecole Normale Superieure de Cachan where he sat the “Aggrégation” (credentialing exam) in electronics of the French National Education. He worked as an Assistant Professor from 1997 to 1999 and as a full time lecturer from 1999 to 2003 at the Department of Electrical Engineering of Technical Institute of the University of Nantes (IUT). Since 2003 he teaches at Ecole polytechnique de l’Université de Nantes (Engineering School) in the Electrical Engineering and the Computer Science departments where is now a Full Professor. Since 2006, he is the head of the Image and Video Communication lab at CNRS IRCCyN, a group of more than 35 researchers. He is mostly engaged in research dealing with the application of human vision modeling in image and video processing. His current centers of interest are 3D image and video quality assessment, watermarking techniques and visual attention modeling and applications. He is co-author of more than 140 publications and communications and co-inventor of 13 international patents on these topics. He also co-chairs within the VQEG (Video Quality Expert Group) the “Joint-Effort Group” and “3DTV” activities. He is currently serving as associate editor for IEEE transactions on Circuit System and Video Technology, SPRINGER EURASIP Journal on Image and Video Processing, and SPIE Electronic Imaging.

Ernst Niebur graduated with an MS degree (Diplom Physiker) from the Universität Dortmund, West Germany. He received a Post-Graduate Diploma in Artificial Intelligence from the Swiss Federal Institute of Technology (EPFL), Switzerland, and the Ph.D. degree (Dr ès sciences) in physics from the Université de Lausanne, Switzerland. His dissertation topic was a detailed computational model of the motor nervous system of the nematode C. elegans.

Niebur was a Research Fellow and a Senior Research Fellow at the California Institute of Technology, Pasadena, and an Adjunct Professor at Queensland University of Technology, Brisbane, Australia. He joined the faculty of Johns Hopkins University in 1995 where he is currently a Professor of Neuroscience in the School of Medicine, and of Brain and Psychological Sciences in the School of Arts and Sciences. He uses computational neuroscience to understand the function of the nervous system at many levels.

Niebur was the recipient of a Seymour Cray (Switzerland) Award in Scientific Computation in 1988, an Alfred P. Sloan Fellowship in 1997, and a National Science Foundation CAREER Award in 1998.