Perceptual Video Compression: A Survey

Jong-Seok Lee, *Member, IEEE*, and Touradj Ebrahimi, *Member, IEEE*

***Abstract*—With the advances in understanding perceptual properties of the human visual system and constructing their computational models, efforts toward incorporating human perceptual mechanisms in video compression to achieve maximal perceptual quality have received great attention. This paper thoroughly reviews the recent advances of perceptual video compression mainly in terms of the three major components, namely, perceptual model definition, implementation of coding, and performance evaluation. Furthermore, open research issues and challenges are discussed in order to provide perspectives for future research trends.**

***Index Terms*—Coding efficiency, focus of attention, human visual system, perceptual video compression, quality of experience (QoE), region of interest, visual perception.**

I. Introduction

In these days, popularity of multimedia content production and distribution has been increasing drastically with the support of advances in related technologies. We are witnessing a remarkable evolution of video communication technologies and an exponentially increasing volume of video contents produced by both professionals and amateurs. Accordingly, video compression has been a very popular research topic for several decades in order to provide efficient solutions for storing and delivering a large amount of video data. The main goal of video compression is to efficiently condense visual data as much as possible while minimizing the loss of visual quality due to compression. Effort for achieving such a goal has led the development of several video coding standards such as MPEG-1, MPEG-2, MPEG-4, H.263, and H.264/AVC.

Coding efficiency of video compression is achieved by removing statistical redundancy and perceptual redundancy. In modern video compression methods including the aforementioned standards, the statistical redundancy removal has been extensively exploited as their core technology, such as motion compensation (MC), intra-frame prediction, entropy coding, and so on. In addition, reducing the perceptual redundancy has been also considered significantly, which includes quantization matrices attenuating high frequency components, chroma subsampling, deblocking filtering, etc. However, the human visual system (HVS) has many important perceptual properties that can be exploited to improve coding efficiency further without significant perceived quality degradation. For example, human observers do not focus on the whole scene, but only a small region around a point of fixation is captured at a high spatial resolution and the peripheral region at lower resolutions, in which intelligent perceptual redundancy reduction in the peripheral region may not be noticeable. Coding techniques based on such an opportunity, which fall in the scope of *perceptual video compression*, have received a great deal of attention. Although the detailed technologies and exploited perceptual properties vary significantly across individual techniques, their main idea is to maximize perceived quality rather than conventional notions of quality that are usually represented by the peak signal-to-noise ratio (PSNR) or the mean square error (MSE). The basic difference between PSNR (or MSE) and perceptual quality measures is that the former looks at distortions from the signal point of view, which leads to poor correlation with perceived quality [96], whereas the latter takes the perceptual point of view. Thus, unlike conventional compression, perceptual compression considers quality measurement that somehow imposes different weights on distortions occurring in different spatial, temporal, or spatio-temporal portions of video data (e.g., regions-of-interest (ROI) vs. non-ROI in ROI coding, salient pixels vs. the rest in saliency-based coding, pixels having different distances from a fixation point in an image in foveated coding, pixels associated with different error visibility values in just-noticeable-distortion-based coding, etc.).

Some difficulties have obstructed the evolution of perceptual video compression. One reason is the lack of understanding of HVS and its perceptual mechanisms that are used as bases of developing perception-aware coding techniques. Although HVS is a popular topic in many areas including biology, psychophysics, psychology, and neuroscience, its functionality and mechanisms are not fully understood yet. Moreover, application- and context-dependent quality expectations of users have sometimes prevented researchers from reaching generally applicable perceptual compression techniques. Nevertheless, perceptual video compression has great potential as a solution to facilitate multimedia content management due to its efficiency for data rate reduction.

Considering these facts, this paper reviews recent advances of perceptual video compression and compares developed technologies for employing perceptual properties in coding with the aim of highlighting promising techniques and prospecting future research trends and challenges. Although perceptual video compression has been continuously researched for the past decade, a complete review of the developed methods is rarely found in literature except for a brief overview given in [17].

In general, designing a perceptual video compression algorithm requires answering the following three major questions:

Manuscript received November 01, 2011; revised May 15, 2012 and August 05, 2012; accepted August 06, 2012. Date of publication August 23, 2012; date of current version September 12, 2012. This work was supported in part by the Ministry of Knowledge Economy, Korea, under the IT Consilience Creative Program (NIPA-2012-H0201-12-1001), in part by Yonsei University Research Fund, and in part by the COST Action IC1003 European Network on Quality of Experience in Multimedia Systems and Services (Qualinet). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Weisi Lin.

J.-S. Lee is with the School of Integrated Technology, Yonsei University, 406-840 Incheon, Korea (e-mail: jong-seok.lee@yonsei.ac.kr).

T. Ebrahimi is with the Multimedia Signal Processing Group, Institute of Electrical Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland (e-mail: touradj.ebrahimi@epfl.ch).

Digital Object Identifier 10.1109/JSTSP.2012.2215006

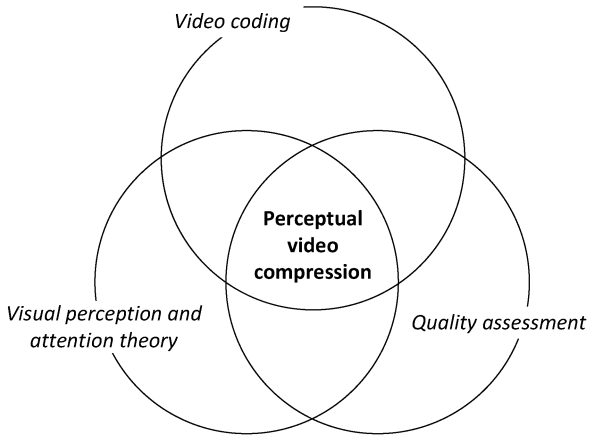

Fig. 1. Research fields related to perceptual video compression.

- What kind of perceptual models can be used, i.e., which parts in video data are more (or less) susceptible to coding distortion in quality perception than others?
- How is the considered perceptual model incorporated in encoding?
- How is the effectiveness of the algorithm (e.g., perceived quality of resulting video sequences) validated?

Therefore, the research of perceptual video compression lies in the intersection of three important research topics, namely, video coding, visual perception and attention theory, and quality assessment, as illustrated in Fig. 1. First, psychological, biological, and psychophysical studies on human visual perception and attention mechanisms provide how perceptual features can be employed for efficient coding. Second, perceptual video compression methods are usually built upon the traditional video coding techniques such as block-based processing, motion estimation (ME), and transformation. Compatibility to existing standard decoders is often regarded as an important requirement. Finally, quality assessment is used for evaluation of the performance of developed encoding techniques, as well as building blocks in encoding (e.g., measuring ME errors). While simple measures such as PSNR are frequently used, they are not well-suited for "perceptual" quality evaluation of "perceptual" coding results. Therefore, subjective and objective quality assessment studies play an important role in developing and evaluating perceptual compression techniques.

In order to provide a brief snapshot of existing techniques, Table I lists representative samples of perceptual compression along with their key ideas. It can be observed that various perceptual models and techniques are used in the three components, i.e., perceptual model definition, video coding implementation, and algorithm validation. In the following, existing work is extensively reviewed from the viewpoint of each of these three components. Section II briefly presents background knowledge about human visual perception mechanisms, which are used as the basis of establishing perceptual models. In Section III, techniques modeling perceptual mechanisms for compression are reviewed. Then, Section IV describes how the models are used for encoding. Methods validating and evaluating perceptual compression algorithms are presented in Section V. Future research revenues and challenges are prospected in Section VI and, finally, concluding remarks are given in Section VII.

II. Human Visual Perception Mechanisms

A. Contrast Sensitivity

It has been shown that the acuity of HVS varies with both the spatial and temporal frequencies of stimuli. A signal whose contrast below a certain threshold with respect to the signal frequency is not detected by HVS. The *contrast sensitivity* is defined as the reciprocal of this threshold. An empirical model considering spatial and temporal frequencies was given in [43], which takes the form of a low-pass filter with respect to the spatial frequency. Thus, human observers are less sensitive to error in high frequency areas than in low frequency areas, which has been popularly used in image and video compression.

B. Masking

Often, the presence of a stimulus reduces or even eliminates the visibility of another stimulus, which is called *visual masking* [6], [28]. Masking can occur both spatially and temporally. In the spatial domain, an error signal in a textured region is more difficult to detect than one in a smooth area. In the temporal domain, a larger difference in luminance between image frames usually leads to a larger temporal masking effect.

C. Fovea

In a human eye, a circular region of about 1.5 mm in diameter on the retina, called *fovea*, exists on the visual axis line. The density of sensor cells in the retina is the highest in this region and decreases rapidly with respect to the angle with the visual axis, called *eccentricity*. The fovea takes up approximately 50% of the visual cortex in the brain. It can capture the scene projected onto it at a high resolution, which covers only a small visual angle of about 2° around the center of gaze, whereas the peripheral vision delivers information at a low resolution decreasing logarithmically with eccentricity. Based on this observation, a model of the contrast sensitivity of HVS as a function of the spatial frequency and the eccentricity was obtained in [30], where the sensitivity decreases as the eccentricity increases and, in addition, such decay is faster for a higher frequency component. The model can be used to find the critical (or cut-off) frequency (i.e., the maximum visible frequency component) for a given eccentricity [91].

D. Visual Attention

Attention is a cognitive process allocating resources of HVS, i.e., it makes HVS focus on a particular aspect of the environment and ignore other things. *Change blindness*, failure of detection of large changes to objects and scenes, and *inattentional blindness*, failure of perception of unattended objects, are examples of phenomena occurring due to visual attention [78].

According to its status, attention can be distinguished to *overt attention* and *covert attention*. The former refers to directing the focus of attention towards a stimulus source with eye movements, whereas the latter is to attend to peripheral locations of interest without moving eyes. Covert attention allows us to

TABLE I
SELECTED EXAMPLES OF RECENTLY DEVELOPED PERCEPTUAL VIDEO COMPRESSION ALGORITHMS (CHRONOLOGICALLY ORDERED)

Ref.	Perceptual information*	Baseline codec**	Implementation***	Evaluation***
[23]	Face	H.263	RC assigning uneven quantization step sizes according to importance	Subjective comparison with H.263 for fixed bit rates
[53]	Fixation point obtained by eye-tracking	H.263	New encoder/decoder using foveation filtering, motion estimation using foveal MAD, RC using foveal MSE, etc.	Objective comparison of each step with that of H.263 in terms of MAD, foveal MAD, PSNR, and foveal PSNR
[39]	Spatio-temporal bottom-up saliency model	Codec-independent	Non-uniform blurring by using Gaussian pyramid, then encoding with MPEG-1 or DivX	Comparison of saliency maps and human subjects' eye fixations
[31]	Motion of arteries	3D SPIHT	Compressing non-ROI coefficients via parametric modeling of their values	(1) Objective comparison with 3D SPIHT in terms of RMSE of ROI (2) Subjective comparison with 3D SPIHT by trained cardiologists
[98]	Luminance adaptation, texture masking, and skin color	H.263	RC using a weighted distortion measure according to sensitivity	(1) Objective comparison with H.263 RC in terms of foreground PSNR (2) Subjective comparison with H.263 RC
[9]	Moving objects	Codec-independent	Blurring of non-ROI, then encoding with MPEG-1 or MPEG-4	Objective comparison in terms of semantic PSNR with frame rate reduction, frame size reduction, and static background
[82]	Moving objects (fast segmentation)	MPEG-4	RC allocating more bits to ROI in proportion to its PSNR, size, and pixel value variance	Objective comparison with MPEG-4 RC in terms of PSNR of ROI and non-ROI
[88]	Bottom-up saliency [39] and skin color (implemented as an embedded system)	Codec-independent	Blurring of non-ROI, then encoding with MPEG-1 or MPEG-4	Subjective comparison with the standard MPEG-1 or MPEG-4 and the case using only bottom-up saliency
[83]	Motion attention, visual sensitivity and visual masking	H.264/AVC	Application of varying QPs to different MBs according to their attention indexes using FMO	Subjective comparison with H.264/AVC using fixed QPs or RC
[58]	Face (fast detection)	H.264/AVC	RC using a weighted distortion measure according to importance of each MB and ROI-based computational power allocation	(1) Objective comparison with H.264/AVC RC in terms of PSNR (2) Comparison of computational complexity with H.264/AVC RC
[4]	Bottom-up saliency [39] and face	Codec-independent	Foveation filtering of images [76], then encoding with MPEG-4	(1) Comparison of attention maps and human subjects' gaze patterns (2) Subjective comparison with the case without foveation for fixed quantization coefficients (3) Subjective comparison with the case using only bottom-up saliency for fixed quantization coefficients
[19]	Motion activity by visual rhythm analysis and face	H.264/AVC	Application of varying QPs to different MBs according to their importance by using FMO	Objective comparison with H.264/AVC RC in terms of PSNR of each slice
[14]	Foveated JND (spatial JND, temporal JND, and foveation using bottom-up saliency [39] and face)	H.264/AVC	RC using a weighted distortion measure according to sensitivity	(1) Objective comparison with H.264/AVC, and the methods in [98] and [58] in terms of PSNR of ROI and foveated peak signal-to-perceptible-noise ratio (2) Subjective comparison with H.264/AVC for fixed bit rates
[57]	Guidance map generated from the bottom-up saliency model [39]	H.264/AVC	Constrained optimization of quantization steps for bit allocation according to the guidance map	(1) Comparison of guidance maps and human subjects' eye fixations (2) Objective comparison with H.264/AVC RC and the method in [39] in terms of eye-tracking weighted PSNR
[50]	Sound-emitting region	H.264/AVC	Application of varying QPs to different MBs according to their distances from the sound source by using FMO	Subjective comparison with H.264/AVC using constant QPs
[103]	Texture vs. non-texture region	Codec-independent	Encoding of non-texture regions with H.264/AVC and entropy coding of side information for texture warping/synthesis	(1) Objective comparison with H.264/AVC in terms of an artefact-based video metric (2) Subjective comparison with H.264/AVC

* JND: just-noticeable-distortion
** 3D SPIHT: 3D wavelet compression based on set partitioning in hierarchical trees
*** FMO: flexible macroblock ordering; MAD: mean absolute distortion; MB: macroblock; MSE: mean square error; PSNR: peak signal-to-noise ratio; QP: quantization parameter; RC: rate control; RMSE: root mean square error; ROI: region-of-interest

quickly scan the field of view for interesting locations, for instance, to spot one's name in a list quickly.

Depending on factors driving attention, attention can be classified into two categories: *bottom-up* (also called *reflexive* or *exogenous*) attention and *top-down* (also called *voluntary* or *endogenous*) attention. Low-level salient features induce bottom-up attention automatically, e.g., abrupt change or prominent appearance of color, shape, motion, orientation, contrast, size, and so on. On the other hand, goal-oriented cognitive control is responsible for top-down attention. A priori knowledge and expectation are typically involved together with such control. For example, road signs on the road would draw attention of car drivers, which are less interesting to small kids in comparison to running cars. In the context of multimedia content, there are some factors that universally tend to draw top-down attention and are often used in engineering applications of visual attention. It has been shown that human observers naturally tend to look at the center of the given scene [32], which has been also criticized in [39], [57]. Human faces are one of the most important and agreed sources drawing top-down attention, which was also proved in psychological studies [10], [36]. Bottom-up and top-down attention mechanisms are thought to interact with each other, which is supported by neurophysiological studies (e.g., [22]) and modeled by neuroscience studies (e.g., [45]).

Computational models of visual attention have been explored extensively due to their usefulness in many applications of computer vision, human-computer interaction, robotics, image and video processing, marketing, design, etc. A thorough review of such models can be found in [29].

E. Multimodality of Attention

In humans' attention, both auditory and visual sensory modalities are often involved simultaneously and influence each other, which is observed in both top-down and bottom-up attention mechanisms [79]. For instance, visual attention is attracted toward the location of an acoustic cue and thus the visual processing capability at the location is enhanced, which is called *cross-modal facilitatory effect* [80]. In [85], it was shown that, even when people are performing a visual task, a novel auditory stimulus can capture their visual attention. This cross-modal orienting is automatic, i.e., it occurs even when detailed information about the visual target is given to subjects in order to prevent uninformative auditory cues from orienting attention [61]. The multimodal information processing capability of the human brain has been supported by many neurological and neuroimaging studies (e.g. [59], [75]).

III. Perceptual Models

In this section, we review the models of perceptual mechanisms used in existing perceptual compression techniques by classifying them into four categories. The first category of the methods requires manual efforts to define which region is important and needs to be kept with high quality in compression, either in advance (Section III-A) or in real-time (Section III-B). The second category is based on visual bottom-up and/or top-down attention mechanisms in order to automatically identify perceptual importance of each pixel or region (Section III-C). In the third category, unequal human visual sensitivity over the scene is derived to identify in which region a large amount of coding artifact can be hidden without being noticed by viewers (Section III-D). Finally, the methods in the fourth category exploit multimodal interaction in attention (Section III-E).

A. A Priori Manual ROI Selection

In some applications, accurate identification of ROIs in video sequences is a critical issue and thus experienced specialists are sometimes involved in manual selection of ROIs. Medical applications are a good example. In [60], a system for transmission of ultrasound videos over mobile WiMAX was proposed, where a medical specialist selects the most diagnostically important areas in ultrasound videos for employing unequal error protection strategies providing higher protection to the diagnostically relevant data. Similarly, a scheme for pediatric video transmission over mobile wireless networks was implemented in [72] by applying higher quality levels for ROIs than non-ROIs, where ROIs were identified by medical experts *a priori*.

B. User Input-Based Attended Region Selection

An accurate way to find important regions in image frames is to get such information from the user while he/she is watching the content. The user can explicitly specify the point or region of fixation by using a pointing device such as a mouse, or an eye-tracker can be used to identify the gaze direction while minimizing the user's effort. Obviously, the latter is feasible only when an eye-tracking equipment is available at the user side.

The rate-shaping method based on a discrete cosine transform (DCT) domain foveation model presented in [37] assumes that the fixation point of the user is explicitly specified by the mouse click and sent to the transmission module through a feedback channel. The real-time foveated multiresolution system developed in [30] also receives the foveation point from a pointing device such as a mouse or an eye-tracker.

A difficulty in using a pointing device is that there could exist transmission and processing delay between the moment of the specification at the receiver side and that of the creation of the perceptually coded content at the transmitter side. The work in [46] tried to resolve this issue by employing a perceptual attention window instead of a fixation point in the developed real-time perceptual transcoding system.

Due to the necessity of user-specific encoding, the approach of using user inputs cannot be implemented in broadcasting applications. Moreover, it is disadvantageous in that a feedback channel is required to send the information about the user attention in real-time. However, the approach can exploit more accurate information about the attended spatial location in comparison to automatic approaches described later. Thus, some particular applications can benefit from this merit. For example, in an interactive teleoperation or telemedicine application, accurate specification of important regions to be encoded with high quality would be critical while reducing the data rate is also important in a resource-constrained network, and the user at the receiver side can be considered as cooperative enough to provide such information. Surveillance is another example of such applications, where a remote user receiving a surveillance video stream may want to watch a certain region with high quality during manual observation. The region can be pointed by the user and the server can apply adaptive coding that encodes the pointed region with high quality.

Other pointing devices can be also applied, e.g., finger touch on a touchscreen, laser pens, motion sensing controllers, and camera-based gesture recognition devices, although their usage is very rare in perceptual video compression.

C. Visual Attention

Visual attention models mimic the human attention mechanism to distinguish perceptually important and unimportant regions in a given scene. Thus, they provide useful information on which part should be encoded with higher quality than other parts. An assumption of perceptual video compression is that attended or unattended regions in a given scene are consistent across observers. If this assumption is violated considerably, a perceptually encoded video sequence that maintains good quality for a group of observers may be perceived as of degraded quality for another group of observers. Fortunately, existing studies strongly support the assumption of general agreement across human observers in viewing behavior and attended regions in given scenes (e.g., [32], [34], [81], [99]).

There exist many computational models for attention in literature [29]. However, this section focuses on particular models that have been successfully applied to perceptual video compression.

1) Bottom-Up Attention: The neurobiological attention model by Itti [39] is one of the most popular approaches in perceptual compression. This model is based on a nonlinear

integration of low level cues of conspicuity in terms of color, flicker, intensity, orientation, and motion in order to imitate center-surround response characteristics of low-level visual processing neurons in the primate brain.

In [33], a spatiotemporal saliency detection model called *phase spectrum of quaternion Fourier transform* (PQFT) was proposed, which uses the phase spectrum information of the quaternion Fourier transform of intensity, color, and motion features. The result of the model is represented as a hierarchical tree-structured sensitivity of an image frame, which is used in *multiresolution wavelet domain foveation* for compression.

2) Top-Down Attention: Top-down attention has been employed more frequently for perceptual compression than bottom-up attention due to its intuitiveness. For example, one of the most popular top-down cues is a human face, which was used in [1], [11], [18], [23], [25], [42], [58], [73], [86], [93]. Especially, top-down attention has been considered in particular applications where the video content does not vary much and thus semantically important regions can be regarded as known *a priori*. Sign language and video conferencing (or videophone) are among such applications, where face, mouth, hand or, more broadly, skin regions are considered as ROIs. Examples of sign language applications are found in [1], [73], while video conferencing applications were presented in [11], [25], [86].

The fact that a moving object draws attention is also popularly used as a top-down cue [9], [41], [66], [77], [82]. For this, the region corresponding to a moving object is segmented in an image frame. Sometimes, it is necessary to distinguish interesting and uninteresting moving regions as well as global motion. For example, different motion characteristics of narrower arteries and others such as chest and heart were exploited to segment the former in image frames for coding of angiogram videos [31]. The model presented in [19] is based on visual rhythm analysis of image frames. Visual rhythm is an abstraction of a video that captures the temporal change of pixel values along a specific sampling line. The temporal evolution of the visual rhythm along the four sampling lines of each image (horizontal, vertical, diagonal, and anti-diagonal) is used to distinguish six different object and camera motions and to construct a rectangular ROI in each image.

3) Combined Attention: There have been efforts to combine the bottom-up and top-down models for developing more realistic attention models. For this, the saliency model in [39] is popularly used as the bottom-up model to be combined with top-down cues such as center [15], skin/face information [4], [88] and/or captions [13].

The way to integrate the two perceptual attention models is an important issue. In [15], the bottom-up saliency map is multiplied by a Gaussian function having the center located at the scene center. In [13], the bottom-up saliency map and two top-down attention maps (faces and captions) are multiplied using user-defined weights. A Bayesian framework was proposed in [4] in order to integrate face and low-level attention cues that are probabilistically modeled. In [88], Itti's saliency model was extended at two levels by using skin/face information. In the feature-level integration, the intermediate conspicuity map for each low-level cue in the saliency model is modulated by the top-down gains computed from the statistical skin representation model. In the result-level integration, all conspicuity maps are again weighted by using the top-down gains and fused to produce the final saliency map. This approach was also used for carotid ultrasound videos by defining an atherosclerotic plaque area (instead of a face region) as the top-down attended area [87].

D. Visual Sensitivity

Visual sensitivity-based approaches are motivated by the existence of a sensitivity threshold, i.e., an error, distortion, or noise level below this threshold is not visually detectable. Thus, the coding distortion is allowed until its level is less than the threshold, which will lead to bit rate reduction without noticeable quality degradation. The threshold value for a pixel location is dependent on the relationship of the pixel with spatially or temporally neighboring pixels in terms of luminance, color, semantics, etc. Often, the sensitivity analysis is combined with the aforementioned attention-based ROI identification approach in order to consider different sensitivity between attended and unattended regions.

Based on the fact that complex texture regions are detail-irrelevant and thus quite a large amount of errors can be hidden in those regions, texture analysis/synthesis-based coding has been developed [3], [5], [27], [64], [65], [68], [103], [105]. Basically, an image frame is segmented into texture and non-texture regions, and the texture region is "analyzed" to build a model that can be used at the encoder side to "synthesize" the region, while the non-texture region is encoded with a conventional encoder.

In [98], two sensitivity factors were considered, namely, luminance adaptation and texture masking. The former refers to the fact that human observers are more sensitive to luminance contrast rather than its absolute value. Thresholds for these two types of sensitivity are obtained [21] and combined to form the so-called *stimulus-driven sensitivity map*. Then, skin color detection is used to modulate this map so as to increase the sensitivity value in a skin region, which produces a cognition-integrated perceptual sensitivity map.

The method in [69] combines a spatial masking model and an importance map considering visual attention. In the former, each 8×8 image block is classified into one of flat, edge, and texture regions in order to ensure minimal quantization errors in flat regions and allow large quantization errors in texture regions. The latter is generated by considering various factors such as contrast, size, shape, location, foreground/background, and motion.

The *visual distortion sensitivity index* (VDSI) proposed in [84] combines the motion information and texture structure in a scene. The idea is that, if a moving object that tends to attract human attention has random texture, it can hide a large amount of coding distortion. The *texture randomness index* is first measured for each macroblock (MB) and, if the *motion attention index* of the block is high enough, the VDSI value takes its maximum, and otherwise, the VDSI is equal to the texture randomness index. A similar approach was shown in [2], where a texture masking model was combined with ROIs identified by

using domain-specific information (e.g., ball, referee, players' faces, etc. in soccer videos).

In the method presented in [83], the whole scene is divided into four regions based on their masking thresholds. The region coded with the highest quality is determined by a motion attention model using motion intensities and directions. Then, DCT coefficients with low spatiotemporal sensitivities in the motion-unattended region are picked and their masking thresholds are estimated, which are based on various factors such as spatial frequencies of DCT coefficients, eye movement, and luminance masking. According to the threshold magnitudes of the MBs, the other three regions are divided, which are encoded with different quality.

In [14], [16], a plausible measure of visual sensitivity, just-noticeable-distortion (JND), was improved by considering the relationship between visibility and eccentricity, which is called *foveated JND* (FJND). The model consists of three terms, i.e., spatial JND, accounting for spatial masking and luminance contrast [21], temporal JND, explaining larger temporal masking for larger inter-frame luminance difference, and foveation, determined by the bottom-up saliency map [39] and skin color.

E. Cross-Modal Attention

While the aforementioned methods consider only visual information, there have been attempts to take into account cross-modal attention recently [47], [49], [50]. In [47], [49], [50], it was assumed that the sound-emitting region serves as an ROI in a scene. An audio-visual source localization algorithm using a mono audio signal was applied to automatically find the sound-emitting region.

Since the sound-emitting region is a subset of moving objects in a scene, a question that might arise is the relative importance of the sound-emitting and silent motions. The subjective study presented in [48] demonstrated unequal amounts of attention received by sound-emitting and silent moving regions. More specifically, the foreground region was defined as either only sound-emitting objects or all moving (both sound-emitting and silent) objects, and the quality of the background region was degraded by blurring in each case. The subjective evaluation results showed that, even if the silent moving objects are blurred in the former case, the perceived quality is not significantly degraded in comparison to the latter case where all moving objects are preserved with high quality. In other words, not all moving objects receive visual attention equally and sound-emitting objects tend to attract more attention than others.

IV. Implementation of Video Compression

Broadly speaking, there exist two different approaches to incorporate perceptual models into encoding: pre-processing and embedded encoding. In the former approach, pre-processing is applied to the images according to the non-uniform distribution of the permissible amount of quality degradation and then the resulting images are encoded with existing encoders. The latter approach controls non-uniform distortion allocation within an encoder. Depending on the encoder type, this approach can be further divided into non-scalable coding and scalable coding.

A. Pre-Processing

Pre-processing of each image frame is the simplest way to apply a non-uniform distortion distribution in a scene, which is sometimes called *offline foveation* [67].

A commonly used pre-processing method is spatial blurring. The strength of blurring in each region is so adjusted that less important regions are more strongly blurred. Since the high frequency components are removed by blurring, less bits are allocated to encode the less important regions. A simple way is to divide the scene into the foreground and background, and only the background region is blurred [9], [88]. However, this may create clearly visible boundaries between the foreground and background regions. In order to solve this problem, a pyramid of blurring filters (e.g., Gaussian) can be used to apply blurring in a gradual manner [39], [42], [49]. In [24], a temporal pre-processing method, which basically updates color maps of the background region only periodically, was additionally applied to the spatially blurred image frames.

A more elaborate pre-processing can be done via foveation filtering [76], as used in [4]. In order to simulate the foveation phenomenon in human eyes (Section II-C), each image frame is filtered by a low-pass filter whose cut-off frequency decreases as the distance of the considered pixel from the fixation point increases.

An advantage of the pre-processing approach is that any of existing encoders can be easily used without modification to encode the pre-processed images. Moreover, blurring prior to encoding can reduce the blockiness artifact that may happen in the region encoded with low quality in the encoder-embedded approach. However, due to the separate quality control and encoding, the coding gain obtained by this approach is regarded as the lower bound of the expected gain that can be reached via perceptual compression.

B. Non-Scalable Coding

There have been efforts to design new perceptual video encoder-decoder pairs. In [30], a real-time foveated multiresolution pyramid codec was developed for low bandwidth video communications. The encoder performs a chain of foveated low-pass pyramid generation, interpolation of foveation region boundaries, ME, quantization, and zero-tree and arithmetic coding, which can be done in real-time to enable interactive video communication between the user specifying the foveation point and the server sending the foveated video stream. The work in [53] presented an end-to-end foveated encoder/decoder employing several foveated video processing algorithms such as foveation filtering, ME, motion compensation, rate control (RC), and post-processing. In the texture analysis/synthesis-based approach, the encoder includes a standard encoder for coding of non-texture regions and an additional module typically performing entropy coding of the side information describing texture regions [3], [5], [27], [64], [65], [68], [103], [105].

However, compliance of a perceptually coded video stream to an existing decoder has been considered as important in order to allow efficient deployment of the encoding methods in existing systems without necessity of replacing decoders.

Some video coding standards inherently support application of different encoding parameters for different regions, which

can be used for assigning spatially uneven quality. The latest video coding standard, H.264/AVC, provides a scheme called *flexible macroblock ordering* (FMO), where a frame can be divided into multiple slices encoded separately. Different quantization parameter (QP) values can be assigned to different slices so that perceptually more important slices are encoded with smaller QP values for higher quality. This idea has been used in several systems, e.g., [1], [2], [19], [47], [50], [51], [83], [84].

The method presented in [76] shares a similar idea to that in [53] (i.e., spatially varying foveation cut-off frequencies), but its foveation processing in the spatial or DCT domain was designed so that it can be incorporated in existing video coding standards such as H.263. The bit rate reduction method proposed in [57] formulates the bit allocation problem considering spatial saliency [39] as a constrained optimization problem, in which the bit rate is minimized such that the distortion weighted by saliency is fixed. The solution of the problem is given as adjustment of the quantization step of each MB, where the step is inversely proportional to the perceptual importance of the MB. In [90], encoding parameters in H.264/AVC are differently assigned for ROI and non-ROI, e.g., the search range for ME is set to a large value for ROI and a small value for non-ROI, and some of the available prediction modes are not considered for non-ROI to save computation. A similar strategy for mobile devices was presented in [41], which controls parameters such as the ME accuracy (e.g., full, half, or quarter pixel), distortion metric (e.g., sum of absolute differences (SAD), sum of squared errors, or Hadamard SAD), search range, and mode selection in H.264/AVC according to the available battery power.

While the aforementioned methods focus on achieving maximal bit rate reduction while preventing significant perceived quality degradation with respect to conventional, the perceptual RC approach tries to improve perceptual quality as much as possible while targeting the same fixed bit rate. A popularly used idea is to employ a perceptual distortion measure that reflects unequal importance of each frame, each visual object, and/or each MB. The perceptual MB-level RC scheme in [98] uses a distortion measure weighted by a perceptual sensitivity considering luminance adaptation, texture masking, and skin detection in the rate-distortion (R-D) model used for obtaining the optimal quantization step size for each MB, which was implemented in the H.263 platform. In the RC scheme based on H.263+ [18], different distortion weights determined manually were used for ROI and non-ROI MBs in the distortion measure and a fuzzy logic controller was used to estimate the optimal QP of each MB. The estimated QP is further refined by considering temporal information, i.e., transition between ROI and non-ROI of a MB in adjacent frames. In [55], the rate-quantization (R-Q) and R-D models for H.263 were constructed by using the *foveal SNR* (FSNR) that considers the foveation characteristics in distortion measurement. The H.263-compatible scheme in [86] uses a weighted MSE (i.e., a ten times larger weight for a face region) for visual object-level RC that eventually leads to MB-level RC. In [14], the distortion measure of each MB was weighted according to its average FJND value for MB-level quantization adjustment. In [58], the importance of a MB based on face detection was used as a weighting factor in the linear R-Q model of H.264/AVC to obtain the optimized QP for the MB. In addition, coding parameters such as mode decision, number of reference frames, accuracy of motion vectors, and ME search range were adjusted at the MB-level to allocate more resources to the ROI.

In another type of perceptual RC, importance of each region is used to determine the number of target bits for the region. A modified H.261 encoder called *foreground/background coding* was implemented in [11], which allocates more bits in face regions than the rest of the scene. The modifications include employment of two quantizers (MQUANT), one for foreground and the other for background, and a RC scheme to adjust the two quantizers periodically. The method in [89] is based on H.263, and skips encoding of non-ROI MBs to re-allocate the saved bits to ROI. In [82], an error-prone wireless video transmission scenario was considered, where decrease of the channel throughput due to retransmission of data can be viewed as a dynamic RC problem. The proposed solution was that, at a given bit rate budget, the target bits are distributed between the moving foreground and static background in proportion to the PSNR, size, and pixel value variance of each region. In [23], the quantization step size of each MB was scaled by its sensitivity derived from face detection and incorporated into the R-D model, from which the base quantization step size was obtained.

The object-based coding approach, enabled in MPEG-4 [38], is also a candidate for implementing perceptual compression due to its inherent capability to assign different quality levels to different objects and background. A scene is treated as a composition of multiple visual objects (VOs) that are separately encoded and decoded. Thus, some of the objects can be encoded with higher quality than the other objects and background. While there is a significant amount of work dealing with strategies allocating an appropriate number of bits to each VO, it is usually assumed that VOs are already segmented and their relative importance is given, and only a few methods proposing complete perceptual compression procedures are found. In [15], combination of Itti's bottom-up saliency model and the center prior were used in obtaining each VO's weighting value proportional to its attention value. In the RC method proposed in [7], the size, location, motion, and coding complexity of each VO were considered for computing the priority of the object for bit allocation. However, both of the systems assumed that results of object segmentation are available *a priori*. In fact, the difficulty and computational complexity of object segmentation are disadvantages of the object-based coding approach, especially for real-time applications. Moreover, the shape information of the VOs needs to be additionally included in the bit stream, which may consume a considerable number of bits.

Not only coding but also transcoding or rate shaping can benefit from exploiting perceptual properties of the HVS. In [37], a rate shaper was proposed based on a DCT-domain foveation model in order to implement a lightweight rate reduction scheme for MPEG-1 bit streams. DCT coefficients smaller than their corresponding amplitude thresholds that depend on the eccentricity are eliminated.

C. Scalable Coding

Scalable coding is an alternative to non-scalable coding for dealing with difficulties of simultaneous video transmission to

multiple end-users having heterogeneous network conditions and terminal characteristics. Scalability in multiple dimensions (e.g., temporal, spatial, and quality dimensions) provides flexibility and adaptability of video transmission. By taking out part of the encoded bit stream, the bit rate can be adaptively adjusted according to the available resources.

In [93], a wavelet-based scalable coding method called *foveation scalable video coding* was proposed, which orders a bit stream so as to place information of the attended area in a scene (i.e., face) at the beginning of the bit stream. Then, truncation of the bit stream due to limited network resources still keeps the important information around the foveation point at the cost of quality degradation in the peripheral region.

A similar approach was implemented in [13] based on the scalable extension of H.264/AVC (called SVC) [74], where ROI and non-ROI are distinguished by using a bottom-up saliency map, face, and captions, and then the non-ROI data in the enhancement layer has a higher priority to be discarded when bit stream truncation is required. Especially, a model called *scalable visual sensitivity profile* [102] was employed to obtain saliency maps with multiple scalable levels, which can be easily incorporated in spatial scalability layers in SVC.

In [12], it was argued that correct receipt of ROI data in a scalable bit stream by using the aforementioned approach does not necessarily guarantee correct decoding of the data due to error propagation between ROI and non-ROI. In order to reduce this undesirable effect, a scalable coding method was proposed, where prediction of a MB in ROI is obtained as a weighted sum of unrestricted ME and confined ME within the ROI reference. The weight controls the trade-off between error resilience and coding efficiency.

V. Evaluation and Validation

The ultimate goal of a coding method is to improve coding gain and/or visual quality in comparison to other methods, and this is also the most important criterion for evaluation of perceptual compression methods. In addition, some real-time applications are validated by minimum additional computational complexity due to computation of the perceptual models. Sometimes, employed attention or saliency models are validated by comparing them with human subjects gaze patterns, which is only an indirect way for evaluation of perceptual compression.

A. Coding Gain

The performance of a perceptual compression method is frequently validated by showing that an additional coding gain in terms of bit rate or file size is obtained without significant quality degradation in comparison with conventional coding methods. Here, a rigorous proof of the same perceptual quality is possible only via statistical analysis of subjective quality assessment results, which will be explained more in Section V-B.

When different quality factors are used for different regions (e.g., non-uniform QPs in a frame), it is often necessary to send additional bits to describe the boundaries between the regions having different quality and the assigned quality factors. For example, when the FMO scheme with arbitrary slice shapes in H.264/AVC is used (i.e., Type 6), the information about which MBs belong to which slices needs to be updated when the slice assignment changes. When a low bit rate condition is considered, the overhead due to this information is not negligible in comparison to the image data, which eventually degrades the coding efficiency. A solution for this is not to update the assignment at every frame but only intermittently, for instance, at every n frames $(n > 1)$ [40], [47] or only when the slice assignment changes for a significant amount of MBs [50], because slice grouping errors at an accuracy of a few MBs would not cause significant quality degradation. A similar issue arises for texture analysis/synthesis-based methods, i.e., the side information enabling a decoder to reconstruct texture regions must be sent along with the conventionally encoded bit stream describing non-texture regions. Thus, the main concern of the methods in this category has been how to generate compact descriptions of texture regions, e.g., [3], [5], [103].

B. Quality Assessment

In general, quality assessment can be done either subjectively or objectively. Subjective quality assessment is the most accurate and reliable way to measure perceived quality of the given content. However, it is usually time-consuming and expensive. Therefore, objective quality metrics have been developed to automatically predict the quality perceived by human observers.

1) Subjective Evaluation: Many environmental and contextual factors influence results of a subjective quality assessment experiment. Thus, it is important to carefully define the goal of the experiment, select appropriate test material, and design the test procedure and environment in order to exclude unwanted external factors and obtain reproducible and reliable results. For this, there has been effort to standardize subjective test activities (e.g., [70], [71]).

The effectiveness of a perceptual compression algorithm that produces lower bit rates in comparison to conventional coding can be subjectively demonstrated by asking subjects to provide ratings for conventionally and perceptually encoded video sequences and showing insignificance of their quality score difference (or even significant superiority of the latter), e.g., [4], [47], [49], [50], [83], [84], [88]. Alternatively, it can be shown that, for fixed bit rate conditions, the quality scores for perceptual compression are higher than those for conventional compression (e.g., [14], [23], [98]).

In applications of specific domains, subjective evaluation by specialists of the domains is appropriate to show the effectiveness of the methods. The method in [72] was designed for telemedicine and, thus, a medical expert evaluated *diagnostic losslessness* of encoded video sequences. In order to evaluate a motion-based perceptual coding method for angiogram videos [31], trained cardiologists were involved in measuring diagnostic difference of keyframe images of original and compressed videos.

A question that may be raised regarding the effectiveness of perceptual compression in general is whether the impairment introduced in perceptually less important regions draws attention, which would be an undesirable side effect of perceptual compression. In [67], it was shown that foveation filtering does not alter the gaze pattern of human observers significantly. However, repeated viewings of the same content can change the gaze

pattern via increase of top-down influences on attentional selection [8] and, consequently, the background region having poor quality may be attended after multiple viewings [67]. In [51], it was shown that content-dependence is involved in quality perception for repeated viewings; talking faces presented with audio signals were found to be strong attractors of visual attention so that quality degradation in the background region was not noticeable even in repeated viewings, which was not true for non-speech contents.

2) Objective Evaluation: Objective quality metrics can be classified into three categories according to the availability of the reference (original) signal: full-reference metrics, when the original signal is fully accessible, reduced-reference metrics, when only partial information of the original signal is available, and no-reference metrics, when the original signal is not accessible.

Many perceptual compression techniques employ PSNR as a metric to demonstrate their performance. The PSNR value over the whole scene tends to be lower by perceptual compression than conventional compression for a fixed bit rate condition. However, that of the ROI or perceptually important region is expected to be improved.

PSNR can be replaced by metrics developed for better correlation with subjective quality. A few metrics designed for evaluation of perceptual compression take the basic form of PSNR, while weighted versions of MSE are used in order to consider perceptual properties in the weights. Examples include the foveal PSNR (FPSNR) [56], which weights distortions with the local bandwidth decreasing with eccentricity, foveated wavelet quality index [91], which combines a wavelet domain visual sensitivity model and a structural distortion measure, peak signal-to-perceptible noise ratio (PSPNR) [20] and foveated peak signal-to-perceptible noise ratio [14], which consider only errors greater than JND and foveated JND thresholds, respectively, semantic PSNR (SPSNR) [9], which imposes different weights on semantically segmented regions, and eye-tracking-weighted PSNR (EWPSNR) [57], where a Gaussian distribution around each eye fixation point is used for weighting.

When the coding techniques and corresponding metrics for evaluation are based on the same or similar perceptual models (e.g., PSPNR, FPSNR, and SPSNR), comparison of the conventional and developed coding methods based on the metrics may be biased. Furthermore, the metrics usually use only partial perceptual properties, and thus may not be generalized for other techniques using different perceptual properties. In this sense, using eye-tracking data for original video sequences (e.g., EWPSNR) would be a fair way, although it requires expensive eye-tracking experiments.

The aforementioned metrics are all full-reference metrics, and reduced-reference or no-reference metrics have been less used in evaluating perceptual compression methods. In [76], a no-reference blockiness metric [92] was used to assess blocking artifacts promoted by the proposed DCT domain foveation algorithm.

Apart from the metrics specifically designed for evaluation of perceptual compression, there are several generic objective metrics developed in the field of visual quality assessment [35], [96]. For example, the structural similarity index map (SSIM) [94], focusing on identifying degradation of structural information to which human observers are sensitive, has been used in [24], [60], [97]. However, the issue how accurate generic metrics are for evaluation of perceptual compression is still unanswered.

C. Computational Complexity

Computation of perceptual models may impose significant computational complexity overhead. Thus, theoretical and empirical analysis of the overhead has been sometimes considered as important, especially in real-time applications.

In [33], some existing saliency models were criticized for their high computational complexity, and the developed PQFT was shown to be able to work in real-time, requiring less than 1 ms per image in C/C++ implementation. The foveated coding method in [76] tried to reduce the computational complexity of the foveation process by using techniques such as an approximated foveation model and lookup tables. For the method that combinationally uses the bottom-up saliency model and the top-down attention to faces [4], it was shown that the computational bottleneck of each attention model is the pyramidal image representation and the optical flow computation for the former and the discrete symmetry transform used in eye detection in a face for the latter, for which available optimized implementations may be adopted. Fast segmentation of moving regions in real-time was considered in [82] for the motion attention-based RC algorithm. It was shown that the complexity of ROI segmentation procedure takes only 2.3% to 2.43% of the total encoding time. In [11], a videophone application was considered, which defined the face region as the ROI. To ensure real-time operation of the method, a fast face segmentation using the skin color distribution was developed, which took less than 1 μs for a CIF-size image. In another real-time conversational video application where the face region was considered as the ROI [58], not only a fast face detection algorithm based on the skin tone information was developed, but also the computational complexity was reduced by disabling computationally complex encoding options at the non-ROI part. Real-time skin region segmentation was implemented for ROI-based RC in [11], [18].

D. Performance Comparison

To our best knowledge, there has been no systematic performance comparison of various perceptual compression methods. Although conducting such comparison is out of scope of this paper, it is still informative to summarize the evaluation results reported in the original papers of various methods (Table II), in order to provide a rough idea about achievable benefits of perceptual compression methods in terms of improvements of compression gain and perceived quality. The columns of the table detail the experimental conditions (number of contents, frame size, frame rate, and bit rate range), the coding gain improvement against conventional methods at the same perceived quality, the subjective and objective quality improvement against conventional methods for the same bit rate conditions, and computational complexity, respectively. In the table, we include only the cases where proper performance comparison was conducted, e.g., the coding gain improvements shown are at the same perceived quality verified via subjective tests. In many cases, the advantage appears as quality improvement, by up to around 2 dB in PSNR, at the perceptually important regions. Traditionally, the

TABLE II
PERFORMANCE EVALUATION RESULTS REPORTED IN THE ORIGINAL PAPERS (CHRONOLOGICALLY ORDERED). FIGURES ARE AVERAGE VALUES OVER DIFFERENT CONDITIONS AND CONTENTS UNLESS MENTIONED

Ref.	Condition	Coding gain for constant perceived quality	Subjective quality improvement for fixed bit rate*	Objective quality improvement for fixed bit rate	Complexity
[98]	6 contents, QCIF, 10 fps, 32-128 kbps		DMOS improvement by 12.3 out of 100 against H.263 TMN8	PSNR improvement by up to 1.76 dB+ at ROI (skin region) against H.263 TMN8	Perceptual sensitivity calculation taking 8% of time for full-search ME
[82]	3 contents, QCIF, 10 fps, 32/64 kbps			PSNR improvement by 2.01 dB at ROI (moving objects) against MPEG-4	ROI segmentation taking 2.34% of total encoding time
[83]	4 contents, CIF, 30 fps, 500-3000 kbps	Improvement by 8.98% against JM			
[58]	3 contents, QCIF, 15/30 fps, 8-244 kbps			PSNR loss by 0.29 dB (which is "negligible") against JM	ME/MC time reduction by 59% against JM; Decoding complexity reduction by 24%
[19]	3 contents, QCIF, 10 fps, 32/48/64 kbps			PSNR improvement by 1.19 dB at ROI (motion+face) against JM	ROI determination taking only 5.07 ms per frame
[14]	5 contents, CIF, 30 fps, 128 kbps			PSNR improvement by 2.24 dB, 0.69 dB, 0.58 dB at ROI (face) against JM, [58], [98]; FPSNR improvement by 1.03 dB, 0.77 dB, 0.43 dB against JM, [58], [98]	
	5 contents, CIF, 30 fps, 50-500 kbps		DMOS improvement by 5.58 out of 100 against JM		
[57]	50 contents, HD (1920×1080), 30 fps, 260-10000 kbps			EWPSNR improvement by 0.97 dB against JM	
[50]	6 contents, SD (720×480 or 576) and HD (1920×1080), 25/30 fps, 285-6109 kbps	Improvement by 10.1% for SD and 41.9% for HD against JM			

*DMOS: differential mean opinion score

+Measured from the graphs in the original paper

performance of perceptual video compression has been tested under low bit rate conditions (consequently, small resolutions such as QCIF or CIF). Recently, high definition (HD) contents have been considered, where a higher gain seems to be obtained.

VI. FUTURE TRENDS AND CHALLENGES

Although many perceptual video compression algorithms have been successfully applied, there are open issues and challenges that still require more research in the future, which are described in this section.

A. Quality Assessment

As seen earlier, the use of quality assessment in perceptual video compression is two-fold. One is to objectively measure the amount of distortion introduced by lossy operations such as quantization (e.g., in R-D optimization). The other is to subjectively and/or objectively evaluate the perceived quality of videos produced by a perceptual compression algorithm. Despite its importance, however, rigorous subjective and objective quality assessment has been often neglected for both purposes. Regarding objective quality assessment, PSNR is by far the most popular measure in both cases. Therefore, incorporating perceptually designed objective quality metrics in perceptual compression algorithms will be a promising way to improve their performance. In addition, thorough statistical subjective analysis of the performance of perceptual compression algorithms will be necessary for reliable verification of their effectiveness.

B. Improving Perceptual Models

The difficulty of applying perceptual models can be explained by taking examples from [57], [84]. In many cases, a moving object tends to draw visual attention, which supports use of motion information as a top-down attention cue. Nevertheless, if the object has complex texture, relatively severe distortion can still be hidden in that region without human observers' notice. In some cases, a moving object making regular motion may belong to the background of a scene and may not draw attention. However, once the object is attended, and if it is smoothly textured, distortion in that region would be easily noticeable.

In order to design a more robust and effective visual perception model, it would be necessary to properly integrate various factors of visual saliency and sensitivity in a unified framework.

C. Environment Dependence

Nowadays, the range of multimedia communication environments are becoming wider and wider, from mobile devices to HD and immersive environments. Different environments have different physical parameters such as display properties, lighting conditions, viewing distance and angle, and sizes of attended and peripheral regions, as well as contextual parameters such as levels of immersion and concentration. Consequently, perceptual factors, e.g., the eye movement pattern over the scene, the resolution of peripheral vision, visibility of coding artifacts, and motion perception, will vary across the environments. For example, the perceived quality tends to change with respect to the size of the video frame size [52], [62]. In addition, a larger amount of coding artifacts can be hidden in the peripheral region without significant overall quality degradation in foveated coding for a larger spatial resolution [51]. Therefore, it would be desirable to investigate environment dependence of effectiveness of perceptual compression algorithms.

D. Perceptual Scalability

More than one type of device are sometimes involved simultaneously in multimedia consumption due to the popularity of network-based content distribution applications. The effectiveness of a perceptual compression method on one type of device may not be guaranteed on another one. Therefore, coping with the heterogeneous viewing environments would be challenging but beneficial. Perceptual scalability could be a solution to produce scalable bit streams from which video data encoded with suitable parameters of perceptual compression can be adaptively extracted according to each of the target environments existing over the network.

The same idea can be also applied to video transmission applications performing dynamic quality adaptation that deals with variations in end-users' device capabilities and network constraints for ensuring the maximal quality of experience (QoE) to users [52]. While such adaptation can be done via the conventional three-dimensional scalability (i.e., frame size, frame rate, and frame quality) (e.g., [44], [101]), perceptual scalability could be considered as another adaptation dimension [54].

E. Bimodality

Multimedia content usually accompanies both audio and visual signals, which influence each other in various ways. Besides the audio-guided visual focus of attention mentioned in Section II-E, there are several other types of audio-visual interaction that can be used for perceptual video compression, e.g., selective attention to a modality [26] and mutual interaction of audio and visual quality [100]. Moreover, joint audio-visual data compression, which tries to find an appropriate balance of bit allocation to each modality for a given bandwidth budget [95], would be an interesting extension of perceptual video compression for low bit rate communication conditions such as mobile environments.

F. 3D Video

Recently, 3D video content is receiving a significant amount of attention as a new type of multimedia. Since 3D video sequences require higher data rates than conventional 2D videos and, thus, benefit from perceptual compression would be very appreciated for efficient 3D video content distribution. Several perceptual properties are newly introduced in 3D videos, which can be exploited for perceptual compression. For example, attention tends to be affected by depth perception and thus ROIs may be defined by considering depth values (e.g., close objects to the viewers) [63], [104], binocular suppression allows quality degradation in one of the left or right views without noticeable 3D quality deterioration, etc.

VII. Concluding Remarks

We have presented an extensive survey of perceptual video compression by categorizing and analyzing existing approaches. Three important stages in developing perceptual video compression algorithms were defined, i.e., perceptual model definition, implementation of coding, and performance evaluation. Various perceptual models have been used to exploit characteristics of HVS, such as application-specific manual definition of ROIs, sensitivity-based models, and top-down/bottom-up visual and cross-modal focus of attention models. Implementation of a perceptual compression algorithm usually takes the form of modification of existing standard coding while the decoder compatibility is maintained, if possible. Evaluation and validation of an algorithm are an important step to demonstrate the effectiveness of the algorithm. Subjective and/or objective quality assessment is usually the primary concern, and sometimes computational complexity is also examined especially for real-time applicability.

Then, open issues for future research were also discussed. Dealing with newly arising multimedia experience such as HD and 3D is necessary to provide efficient solutions for future multimedia content production and distribution. Coping with various factors affecting the performance of perceptual video compression, such as its dependence on environment and personal variability, still remains challenging. Finally, more profound understanding of visual and audio-visual perceptual mechanisms will be desirable through multi-disciplinary research. Then, such understanding will need to be incorporated effectively in compression algorithms, which will provide efficient solutions for many multimedia applications dealing with continuously increasing volumes of video data.

Appendix
Abbreviations

DCT	Discrete cosine transform
DMOS	Differential mean opinion score
EWPSNR	Eye-tracking weighted peak signal-to-noise ratio
FMO	Flexible macroblock ordering
FPSNR	Foveated peak signal-to-noise ratio
HD	High definition
HVS	Human visual system
JND	Just-noticeable-distortion
MAD	Mean absolute distortion
MB	Macroblock
MC	Motion compensation
ME	Motion estimation
MSE	Mean square error
PSNR	Peak signal-to-noise ratio
PQFT	Phase spectrum of quaternion Fourier transform
QoE	Quality of experience
QP	Quantization parameter
RC	Rate control
R-D	Rate-distortion
RMSE	Root mean square error
ROI	Region-of-interest
R-Q	Rate-quantization
SAD	Sum of absolute differences

SD	Standard definition
SPSNR	Semantic peak signal-to-noise ratio
VDSI	Visual distortion sensitivity index
VO	Visual object

REFERENCES

[1] D. Agrafiotis, N. Canagarajah, D. R. Bull, and M. Dye, "Perceptually optimised sign language video coding based on eye tracking analysis," *Electron. Lett.*, vol. 39, no. 24, pp. 1703–1705, Nov. 2003.

[2] D. Agrafiotis, S. J. C. Davies, N. Canagarajah, and D. R. Bull, "Towards efficient context-specific video coding based on gaze-tracking analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 4, pp. 22:1–22:15, Dec. 2007.

[3] J. Ballé, A. Stojanovic, and J.-R. Ohm, "Models for static and dynamic texture synthesis in image and video compression," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1353–1365, Nov. 2011.

[4] G. Boccignone, A. Marcelli, P. Napoletano, G. D. Fiore, G. Iacovoni, and S. Morsa, "Bayesian integration of face and low-level cues for foveated video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 12, pp. 1727–1740, Dec. 2008.

[5] M. Bosch, F. Zhu, and E. J. Delp, "Segmentation-based video compression using texture and motion models," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1366–1377, Nov. 2011.

[6] B. Breitmeyer and H. Öğmen, *Visual Masking: Time Slices through Conscious and Unconscious Vision*, 2nd ed. Oxford, U.K.: Oxford Univ. Press, 2006.

[7] X. Cai, F. H. Ali, and E. Stipidis, "Object-based video coding with dynamic quality control," *Image Vis. Comput.*, vol. 28, no. 3, pp. 285–297, Mar. 2010.

[8] R. Carmi and L. Itti, "The role of memory in guiding attention during natural vision," *J. Vision*, vol. 6, no. 9, pp. 898–914, Aug. 2006.

[9] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Semantic video analysis for adaptive content delivery and automatic description," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1200–1209, Oct. 2005.

[10] M. Cerf, J. Harel, W. Einhauser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2007, vol. 20.

[11] D. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, pp. 551–934, Jun. 1999.

[12] Q. Chen, X. Yang, L. Song, and W. Zhang, "Robust video region-of-interest coding based on leaky prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 9, pp. 1389–1394, Sep. 2009.

[13] Q. Chen, G. Zhai, X. Yang, and W. Zhang, "Application of scalable visual sensitivity profile in image and video coding," in *Proc. IEEE Int. Symp. Circuits Syst.*, Seattle, WA, May 2008, pp. 268–271.

[14] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 806–819, Jun. 2010.

[15] Z. Chen, J. Han, and K. N. Ngan, "Dynamic bit allocation for multiple video object coding," *IEEE Trans. Multimedia*, vol. 8, no. 6, pp. 1117–1124, Dec. 2006.

[16] Z. Chen, M. Li, and Y.-P. Tan, "Perception-aware multiple scalable video streaming over WLANs," *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 675–678, Jul. 2010.

[17] Z. Chen, W. Lin, and K. N. Ngan, "Perceptual video coding: Challenges and approaches," in *Proc. ICME*, Singapore, Jul. 2010, pp. 784–789.

[18] M.-C. Chi, M.-J. Chen, C.-H. Yeh, and J.-A. Jhu, "Region-of-interest video coding based on rate and distortion variations for H.263+," *Signal Process.: Image Commun.*, vol. 23, pp. 127–142, 2008.

[19] M.-C. Chi, C.-H. Yeh, and M.-J. Chen, "Robust region-of-interest determination based on user attention model through visual rhythm analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 7, pp. 1025–1038, Jul. 2009.

[20] C.-H. Chou and C.-W. Chen, "A perceptually optimized 3-D sub-band codec for video communication over wireless channels," , vol. 6, no. 2, pp. 143–156, Apr. 1996.

[21] C.-H. Chou and Y.-C. Li, "A perceptually tuned sub-band image coder based on the measure of just-noticeable-distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 6, pp. 467–476, Dec. 1995.

[22] M. Corbetta and G. L. Shulman, "Control of goal-oriented and stimulus-driven attention in the brain," *Nature Reviews Neurosci.*, vol. 3, no. 3, pp. 201–215, Mar. 2002.

[23] S. Daly and J. Ribas-Corbera, "As plain as the noise on your face: Adaptive video compression using face detection and visual eccentricity models," *J. Electron. Imaging*, vol. 10, no. 1, pp. 30–46, Jan. 2001.

[24] C. Dikici and H. I. Bozma, "Attention-based video streaming," *Signal Process.: Image Commun.*, vol. 25, no. 10, pp. 745–760, Nov. 2010.

[25] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, "Low bit-rate coding of image sequences using adaptive regions of interest," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 8, pp. 928–934, Dec. 1998.

[26] J. Driver, "A selective review of selective attention research from the past century," *British J. Psychol.*, vol. 92, pp. 53–78, Feb. 2001.

[27] A. Dumitraş and B. G. Haskell, "An encoder-decoder texture replacement method with application to content-based movie coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 825–840, Jun. 2004.

[28] J. T. Enns and V. D. Lollo, "What's new in visual masking," *Trends in Cognitive Sci.*, vol. 4, no. 9, pp. 345–352, Sep. 2000.

[29] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. Applied Percept.*, vol. 7, no. 1, pp. 6:1–6:39, Jan. 2010.

[30] W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication," in *Proc. SPIE*, San Jose, CA, Jan. 1998, vol. 3299, pp. 294–305.

[31] D. Gibson, M. Spann, and S. I. Woolley, "A wavelet-based region of interest encoder for the compression of angiogram video sequences," *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 2, pp. 103–113, Jun. 2004.

[32] R. B. Goldstein, R. L. Woods, and E. Peli, "Where people look when watching movie: Do all viewers look at the same place?," *Comput. Biol. Med.*, vol. 37, no. 7, pp. 957–964, Jul. 2007.

[33] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[34] U. Hasson, O. Landesman, B. Knappmeyer, I. Vallines, N. Rubin, and D. J. Heeger, "Neurocinematics: The neuroscience of film," *Projections*, vol. 2, no. 1, pp. 1–26, 2009.

[35] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Process.: Image Commun.*, vol. 25, no. 7, pp. 469–481, Aug. 2010.

[36] O. Hershler and S. Hochstein, "At first sight: A high-level pop out effects for faces," *Vis. Res.*, vol. 45, no. 13, pp. 1707–1724, Jun. 2005.

[37] C.-C. Ho, J.-L. Wu, and W.-H. Cheng, "A practical foveation-based rate-shaping mechanism for MPEG videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 11, pp. 1365–1372, Nov. 2005.

[38] *Information Technology-Coding of Audio/Visual Objects, Part 2: Visual*, ISO/IEC 14496-2:2004, 2004.

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

[40] A. K. Kannur and B. Li, "An enhanced rate control scheme with motion assisted slice grouping for low bit rate coding in H.264," in *Proc. ICIP*, San Diego, CA, Oct. 2008, pp. 2100–2103.

[41] A. K. Kannur and B. Li, "Power-aware content-adaptive H.264 video encoding," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 925–928.

[42] L. S. Karlsson and M. Sjöström, "Improved ROI video coding using variable Gaussian pre-filters and variance in intensity," in *Proc. ICIP*, Genoa, Italy, Sep. 2006, vol. 2, pp. 313–316.

[43] D. H. Kelly, "Motion and vision. II. Stabilzed spatio-temporal threshold surface," *J. Opt. Soc. Amer.*, vol. 69, no. 10, pp. 1340–1349, Oct. 1979.

[44] A. Khan, L. Sun, and E. Ifeachor, "QoE prediction model and its application in video quality adaptation over UMTS networks," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 431–442, Apr. 2012.

[45] E. I. Knudsen, "Fundamental components of attention," *Annu. Rev. Neurosci.*, vol. 30, pp. 57–78, 2007.

[46] O. V. Komogortsev and J. I. Khan, "Predictive real-time perceptual compression based on eye-gaze-position analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 4, no. 3, pp. 23:1–23:16, Sep. 2008.

[47] J.-S. Lee and T. Ebrahimi, "Efficient video coding in H.264/AVC by using audio-visual information," in *Proc. IEEE Int. Conf. Multimedia Signal Process.*, Rio de Janeiro, Brazil, Oct. 2009, pp. 1–6.

[48] J.-S. Lee, F. D. Simone, and T. Ebrahimi, "Influence of audio-visual attention on perceived quality of standard definition multimedia content," in *Proc. Int. Workshop Quality of Multimedia Experience*, San Diego, CA, Jul. 2009, pp. 13–19.

[49] J.-S. Lee, F. D. Simone, and T. Ebrahimi, "Video coding based on audio-visual attention," in *Proc. Int. Conf. Multimedia and Expo*, New York, Jun. 2009, pp. 57–60.

[50] J.-S. Lee, F. D. Simone, and T. Ebrahimi, "Efficient video coding based on audio-visual focus of attention," *J. Vis. Commun. Image R.*, vol. 22, no. 8, pp. 704–711, Nov. 2011.

[51] J.-S. Lee, F. D. Simone, and T. Ebrahimi, "Subjective quality evaluation of foveated video coding using audio-visual focus of attention," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1322–1331, Nov. 2011.

[52] J.-S. Lee, F. D. Simone, N. Ramzan, E. Izquierdo, and T. Ebrahimi, "Quality assessment of multidimensional video scalability," *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 38–46, Apr. 2012.

[53] S. Lee and A. C. Bovik, "Fast algorithms for foveated video processing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 2, pp. 149–162, Feb. 2003.

[54] S. Lee, A. C. Bovik, and Y. Y. Kim, "High quality, low delay foveated visual communications over mobile channels," *J. Vis. Commun. Image R.*, vol. 16, no. 2, pp. 180–211, Apr. 2005.

[55] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 977–992, Jul. 2001.

[56] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 129–132, Mar. 2002.

[57] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image Vis. Comput.*, vol. 29, no. 1, pp. 1–14, Jan. 2011.

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

[59] E. Macaluso and J. Driver, "Multisensory spatial interactions: A window onto functional integration in the human brain," *Trends in Neurosci.*, vol. 28, no. 5, pp. 264–271, May 2005.

[60] M. G. Martini and C. T. E. R. Hewage, "Flexible macroblock ordering for context-aware ultrasound video transmission over mobile WiMAX," *Int. J. Telemed. Appl.*, vol. 2010, pp. 1–14, 2010.

[61] V. Mazza, M. Turatto, M. Rossi, and C. Umilta, "How automatic are audiovisual links in exogenous spatial attention?," *Neuropsychologia*, vol. 45, no. 3, pp. 514–522, 2007.

[62] J. D. McCarthy, M. A. Sasse, and D. Miras, "Sharp or smooth? Comparing the effects of quantization vs. frame rate for streamed video," in *Proc. SIGCHI Conf. Human Factors in Comput. Syst.*, Vienna, Austria, Apr. 2004, pp. 535–542.

[63] S. Nasir, C. T. E. R. Hewage, Z. Ahmad, M. Mrak, S. Worrall, and A. Kondoz, "Quality-driven coding and prioritization of 3D video over wireless networks," in *High-Quality Visual Experience*, ser. Signals and Communication Technology, M. Mrak, M. Grgic, and M. Kunt, Eds. New York: Springer, 2010, ch. 21, pp. 477–495.

[64] P. Ndjiki-Nya, D. Bull, and T. Wiegand, "Perception-oriented video coding based on texture analysis and synthesis," in *Proc. ICIP*, Cairo, Egypt, Nov. 2009, pp. 2273–2276.

[65] P. Ndjiki-Nya, B. Mokai, G. Blättermann, A. Smolic, H. Schwarz, and T. Wiegand, "Improved H.264/AVC coding using texture analysis and synthesis," in *Proc. ICIP*, Barcelona, Spain, Sep. 2003, pp. 849–852.

[66] E. Nguyen, C. Labit, and J.-M. Odobez, "A ROI approach for hybrid image sequence coding," in *Proc. ICIP*, Austin, TX, Nov. 1994, vol. 3, pp. 245–249.

[67] M. Nyström and K. Holmqvist, "Effect of compressed offline foveated video on viewing behavior and subjective quality," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 6, no. 1, pp. 1–14, 2010.

[68] B. T. Oh, Y. Su, A. Segall, and C.-C. J. Kuo, "Synthesis-based texture coding for video compression with side information," in *Proc. ICIP*, San Diego, CA, Oct. 2008, pp. 1628–1683.

[69] W. Osberger, A. J. Maeder, and N. Bergmann, "A perceptually based quantization technique for MPEG encoding," in *Proc. SPIE*, San Jose, CA, Jan. 1998, vol. 3299, pp. 148–159.

[70] "Methodology for the subjective assessment of the quality of television pictures," Rec. ITU-R BT.500-11, 2002.

[71] "Subjective video quality assessment methods for multimedia applications," Rec. ITU-R P.910, 1999.

[72] S. P. Roi, N. S. Jayant, M. E. Stachura, E. Astapova, and A. Pearson-Shaver, "Delivering diagnostic quality video over mobile wireless networks for telemedicine," *Int. J. Telemed. Appl.*, vol. 2009, pp. 1–9, 2009.

[73] D. M. Saxe and R. A. Foulds, "Robust region of interest coding for improved sign language telecommunication," *IEEE Trans. Inf. Technol. Biomed.*, vol. 6, no. 4, pp. 310–316, Dec. 2002.

[74] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[75] R. Sharma, V. I. Pavlović, and T. S. Huang, "Toward multimodal human-computer interface," *Proc. IEEE*, vol. 86, no. 5, pp. 853–869, May 1998.

[76] H. R. Sheikh, B. L. Evans, and A. C. Bovik, "Real-time foveation techniques for low bit rate video coding," *Real-Time Imaging*, vol. 9, pp. 27–40, 2003.

[77] L. Shen, Z. Liu, Z. Zhang, and H. Yan, "Motion attention based frame-level bit allocation scheme for H.264," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, Kunming, Yunnan, China, Nov. 2009.

[78] D. J. Simons and C. F. Chabris, "Gorillas in our midst: Sustained inattentional blindness for dynamic events," *Perception*, vol. 28, no. 9, pp. 1058–1074, 1999.

[79] C. Spence, "Crossmodal spatial attention," *Ann. New York Acad. Sci.*, vol. 1191, pp. 182–200, Mar. 2010.

[80] C. Spence and J. Driver, "Audiovisual links in exogenous covert spatial orienting," *Percept. Psychophys.*, vol. 59, no. 1, pp. 1–22, 1997.

[81] L. Stelmach, W. Tam, and P. Hearty, "Static and dynamic spatial resolution in image coding: An investigation of eye movements," in *Proc. SPIE*, San Jose, CA, Feb. 1992, vol. 1453, pp. 147–152.

[82] Y. Sun, I. Ahmad, D. Li, and Y.-Q. Zhang, "Region-based rate control and bit allocation for wireless video transmission," *IEEE Trans. Multimedia*, vol. 8, no. 1, pp. 1–10, Feb. 2006.

[83] C.-W. Tang, "Spatiotemporal visual considerations for video coding," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 231–238, 2007.

[84] C.-W. Tang, C.-H. Chen, Y.-H. Yu, and C.-J. Tsai, "Visual sensitivity guided bit allocation for video coding," *IEEE Trans. Multimedia*, vol. 8, no. 1, pp. 11–18, Feb. 2006.

[85] D. J. Tellinghuisen and E. J. Nowak, "The inability to ignore auditory distractors as a function of visual task perceptual load," *Percept. Psychophys.*, vol. 65, no. 5, pp. 817–828, 2003.

[86] L. Tong and K. R. Rao, "Region-of-interest based rate control for low-bit-rate video conferencing," *J. Electron. Imaging*, vol. 15, no. 3, pp. 1–12, 2006.

[87] N. Tsapatsoulis, C. Loizou, and C. Pattichis, "Region of interest video coding for low bit-rate transmission of carotid ultrasound videos over 3G wireless networks," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Lyon, France, Aug. 2007, pp. 3717–3720.

[88] N. Tsapatsoulis, K. Rapantzikos, and C. Pattichis, "An embedded saliency map estimator scheme: Application to video coding," *Int. J. Neural Syst.*, vol. 17, no. 4, pp. 289–304, 2007.

[89] H. Wang, Y. Liang, and K. El-Maleh, "Real-time region-of-interest video coding using content-adaptive background skipping with dynamic bit reallocation," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 45–48.

[90] M. Wang, T. Zhang, C. Liu, and S. Goto, "Region-of-interest based dynamical parameter allocation for H.264/AVC encoder," in *Proc. Picture Coding Symp.*, Chicago, IL, May 2009, pp. 1–4.

[91] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1397–1410, Oct. 2001.

[92] Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. ICIP*, Vancouver, BC, Canada, Oct. 2000, vol. 3, pp. 981–984.

[93] Z. Wang, L. Lu, and A. C. Bovic, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 243–254, Feb. 2003.

[94] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process.: Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[95] S. Winkler and C. Faller, "Perceived audiovisual quality of low bitrate multimedia content," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 973–980, Oct. 2006.

[96] S. Winkler and M. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660–668, Sep. 2008.

[97] L. Yang, L. Zhang, S. Ma, and D. Zhao, "A ROI quality adjustable rate control scheme for low bitrate video coding," in *Proc. Picture Coding Symp.*, Chicago, IL, May 2009, pp. 1–4.

[98] X. Yang, W. Lin, Z. Lu, X. Lin, S. Rahardja, E. Ong, and S. Yao, "Rate control for videophone using local perceptual cues," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 4, pp. 496–507, Apr. 2005.

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

[100] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, "Perceptual-based quality assessment for audio-visual services: A survey," *Signal Process.: Image Commun.*, vol. 25, no. 7, pp. 482–501, Aug. 2010.

[101] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, "Three dimensional scalable video adaptation via user-end perceptual quality assessment," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 719–727, Sep. 2008.

[102] G. Zhai, Q. Chen, X. Yang, and W. Zhang, "Scalable visual significance profile estimation," in *Proc. ICASSP*, Las Vegas, NV, Apr. 2008, pp. 268–271.

[103] F. Zhang and D. R. Bull, "A parametric framework for video compression using region-based texture models," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1378–1392, Nov. 2011.

[104] Y. Zhang, G. Jiang, M. Yu, Y. Yang, Z. Peng, and K. Chen, "Depth perceptual region-of-interest based multiview video coding," *J. Vis. Commun. Image R.*, vol. 21, no. 5–6, pp. 498–512, 2010.

[105] C. Zhu, X. Sun, F. Wu, and H. Li, "Video coding with spatio-temporal texture synthesis and edge-based inpainting," in *Proc. ICME*, Hannover, Germany, Jun. 2008, pp. 813–816.

Jong-Seok Lee (M'06) received his Ph.D. degree in electrical engineering and computer science in 2006 from KAIST, Korea, where he also worked as a postdoctoral researcher and an adjunct professor. From 2008 to 2011, he worked as a research scientist in the Multimedia Signal Processing Group at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Currently, he is an assistant professor at the School of Integrated Technology, Yonsei University, Korea. His research interests include multimedia signal processing and multimodal human-computer interaction. He is author or co-author of over 60 publications. He was the chair of the First Spring School on Social Media Retrieval held in 2010 and an organizing committee member of its second edition in 2011. He is a voting member of Multimedia Communication Technical Committee of the IEEE Communication Society.

Touradj Ebrahimi (M'92) received his M.Sc. and Ph.D., both in electrical engineering, from Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 1989 and 1992, respectively. From 1989 to 1992, he was a research assistant at the Signal Processing Laboratory of EPFL. During the summer 1990, he was a visiting researcher at the Signal and Image Processing Institute of the University of Southern California, Los Angeles, California. In 1993, he was a research engineer at the Corporate Research Laboratories of Sony Corporation in Tokyo. In 1994, he served as a research consultant at AT&T Bell Laboratories. He is currently a Professor heading Multimedia Signal Processing Group at EPFL, where he is involved with various aspects of digital video and multimedia applications. He is author or co-author of over 100 papers and holds 10 patents.