The area of human visual system (HVS) based perceptual picture (including images and videos) processing has been an important and very challenging topic in both theory and practice since the very beginning [4]. Its importance and relevance for providing an improved user’s quality of experience (QoE) have become increasingly obvious to the research and professional communities and industries, which prompted efforts to account for perceptual-based visual characteristics in the design and development of future generations of picture coding standards, products, systems and applications such as super- or ultra-high definition (SHD or UHD) imaging and video systems, digital cinema distribution systems, 3-D video/TV, immersive interactive visual systems, and medical imaging and archive systems for telemedicine/telehealth applications [5-7]. It is high time for the visual communication and entertainment services to be transformed from a device-centric “best effort” to user-centric “quality assured” practice for the associated industries to be sustainable in the long run.

In this paper, an overview is provided of state-of-the-art technologies for perceptual visual signal processing, as well as discussions on the issues related to their implementation, optimization and testing. The main frequently used computational modules of a perceptual-based visual processing system are first described in Section 2. Selected perceptual-based visual processing techniques and applications that make use of perceptual models as a means to optimize their performance are presented in Sections 3 to 5, including image/video compression (Section 3), visual signal quality evaluation (Section 4), and computer graphics (Section 5). The most significant research for each topic is highlighted while providing extensive references to related research activities. Finally, a conclusion is presented in Section 6.

2. PERCEPTUAL VISUAL MODELING

A number of perceptual visual models have been developed and have been adopted in the literature. We will discuss two types of perceptual visual modeling in this section.

2.1. Human Visual Attention (VA)

The human visual attention (VA) is the result of several millions of years of evolution by which humans can rapidly
direct their gaze toward objects of interest in the visual field. Two major attention mechanisms include top-down (knowledge/task-driven) [8] and bottom-up (stimulus-driven) [9]. In the former mechanism, attention is under the control of the subject and is related to cognition processing in the human brain. In the latter mechanism, attention is driven by external stimuli to determine which location is sufficiently different from its surroundings to be worthy of one’s attention. Generally, the stimuli involved in top-down control are pattern (e.g., human faces), shape and other cognitive features, while the features involved in bottom-up control include luminance, color, orientation and motion contrast. Moreover, audition, touch, and other sensory features also affect VA [10].

The first explicit bottom-up computational architecture was proposed by Koch and Ullman [11], with the result being a 2-D topographic map that represents the stimulus conspicuity, or saliency, at every location in the visual scene. This general architecture has been further developed and implemented [12]. In this model, the early stages of visual processing decompose the incoming visual input into feature maps of colors, intensities, and orientations. The “center-surround” operation is then implemented on multi-scale feature images, which are obtained by using dyadic Gaussian pyramids. All obtained feature maps are then combined into a saliency map to detect attended regions by using a winner-take-all neural network.

There are many other bottom-up VA models, e.g., Graph-based Visual Saliency model [13] with graph theory for saliency maps from low-level features; a VA model based on Shannon’s self-information measure [14]; machine learning based saliency map for images [15] with features of multi-scale contrast, center-surround histogram and color spatial distribution; the model based on Fourier Transform [16]; the one considering the human visual sensitivity [17]; GAFFE (gaze-attentive fixation finding engine) [18] and FTS (frequency-tuned salient) region detection [19]. For video modeling, motion information can be used [20].

2.2. Just Noticeable Difference
It is well known that the HVS cannot sense all changes in an image/video due to its underlying physiological and psychological mechanisms [4]. The JND (Just Noticeable Difference) is mainly based upon the temporal/spatial contrast sensitivity function (CSF), which describes the sensitivity of the HVS to each frequency component, as determined by psychophysical experiments), background luminance adaptation (LA, referring to how the contrast sensitivity of the HVS changes in function of the background luminance) and contrast masking (CM, referring to the masking effect of the HVS in the presence of two or more simultaneous frequency components). The JND can be determined either in a sub-band domain [21,22] or in the pixel domain [23,24].

The widely-used JND models were formulated in [21,22] based upon CSF, CM and LA in the DCT (Discrete Cosine Transform) domain. The JND model in [25] formulates CSF and contrast masking in 6 band-pass sub-bands, based on a Laplacian pyramid decomposition of images. Although a model operating in pixel domain can be obtained by summing the effects of the visual thresholds in sub-bands [26], such a model can be derived directly with pixels without subband decomposition [23,24]. The key issue of pixel-domain JND estimation is differentiation of edge and textured regions [27,28].

3. PERCEPTION BASED VISUAL SIGNAL CODING
Theories and principles underpinning digital visual signal coding research are widely known to be the Nyquist-Shannon sampling theorem [29], the Shannon source coding theorem or entropy [30], and the rate-distortion theory [31]. Challenges of varying degrees have been mounted ever since the introduction of the aforementioned theories in field of visual signal processing, and become more intense in the past decade or two with noticeable research activities in compressive sampling [32], the perceptual entropy [33] and perceptually lossless coding of digital pictures [34,35], and rate-perceptual-distortion optimization [33,36].

HVS based picture coding can be broadly grouped into two categories, i.e., perceptual lossless coding and perceptual lossy coding. Perceptual lossless coding removes visually redundant information, offers higher compression ratio than what information lossless coding does with no perceivable picture degradation, and defines the neutral point on the perceptual picture quality scale, given the reference. Perceptual lossy coding, on one hand, aims at delivering constant perceptual picture quality at different JND steps as appropriate to applications in quality assured services and, on the other hand, uses rate-perceptual-distortion optimization for constant bit-rate coding services, providing markedly better perceptual picture quality than traditional information lossy coding techniques where raw mathematical distortion measures, such as the mean square error (MSE), are used in rate-distortion optimization.

3.1. Perceptual Lossy Coding
Many approaches to perception based digital picture coding have been proposed, usually based on the existing coding standards, and modifications are made to explore the perceptual aspect of digital picture coding. Early perceptual picture coders considered simple characteristics of the HVS or HVS weighting function in adaptive quantization design [37,38]. Research efforts over the years have refined this perceptual coding approach which anchors the adaptive quantization to a JND detection model incorporating CSF, LA and CM characteristics of the HVS [39,40,23,41,42,28]. JND detection models have been successfully used in motion estimation and handling for video compression [27,43]. Embedding HVS based perceptual distortion measures in rate-perceptual-distortion optimization process of picture coding forms an alternative approach to perceptual coder design where perceptual distortion is minimized for constant bitrate coding or used to maintain a uniform perceptual distortion level for the entire image or
video [44,45,46]. Stand-alone perceptual distortion metrics providing a visual discrimination map to advise perceptual coding strategies have been successfully deployed in adaptive quantization and bit allocation as well as rate-distortion optimization for quality controlled video coding [47-49]. Utilizing cognitive knowledge and experience such as VA or region of interests have also been reported including foveated image and video coding [50-52].

In [43], the visual signal is smoothed within the constraint of JND for better coding performance. In [42], the CSF, LA, and CM characteristics of the HVS are exploited in developing a locally adaptive perceptual-based image codec that is fully compatible with Part 1 of the JPEG2000 standard (so, any JPEG2K decoder can decode the resulting bitstream), achieving an improved perceived visual quality for both lossy and lossless coding. In [53] and [54], the QPs are adjusted according to the visual impact of the signal for the DCT based coding systems such as JPEG and H.264. In the scheme developed in [46], the perceived error is approximated by a multi-channel color CGC (Contrast Gain Control) vision model based perceptual distortion metric for R-D optimization in order to maximize the visual quality of JPEG 2000 coded images. Compared with the JND detection model based approach, the multi-channel CGC vision model [55] has been adopted in [45] and [46] where model parameters are refitted using application specific subjective test data [56].

3.2. Perceptual Lossless Coding

Visually lossless compression is a special type of perception based coding and it is achieved when a compressed visual signal cannot be differentiated from its original, i.e., the visual difference between the compressed picture and its reference is just below or equal to the just not-noticeable-difference (JNND) threshold. There have been few reports on perceptual lossless picture coding and even less have gone through rigorous verification process via, e.g., double blind subjective tests using a sufficiently large sample pool [35].

In [57], different bit streams are generated by using a standard encoder with different given bit rates, and the one with a resultant visual quality (obtained from a quality measurement, e.g., multi-scale SSIM [58]) close to a predefined threshold (e.g., 0.995 for the multi-scale SSIM) is selected as the bit stream for visually lossless. However, the criterion with which the quality score is close to the predefined threshold is not sufficient for visually lossless, because the quality score can also be high for the case where the image with visible distortion on only a small portion of the image. Therefore, most of the existing visually lossless coding methods [36,59-61] are based on the JND and the CGC with visual pruning, exploiting masking properties of the HVS.

In [41], a perceptual-based image coder that can achieve perceptually lossless compression at high compression ratios (15 to 20) is proposed by adapting the quantizer reconstruction levels to the local amount of masking present at the level of each subband transform coefficient and without the need to send any additional side information to the decoder. Other methods modified one of the standard (e.g., JPEG, or JPEG 2000) encoders to account for the perceptual redundancy, where the distortion related parameters in the encoder are adjusted according to the JND model to guarantee that the reconstructed signal is visually lossless. In [35,62], a visual pruning processing is introduced to a multi-channel CGC vision model which is embedded in the JPEG 2000 encoder producing standard complaint bit-stream, forcing perceptual distortion below or at JNND threshold and achieving perceptual lossless coding; in [59] a JND model is incorporated into the JPEG 2000 encoder and the encoded bit stream can be decoded by a JPEG 2000 decoder; in [60] and [61], the visually lossless coding is realized by the DCT based encoder with a JND-adjusted QP and the resultant bit stream is not standard compliant.

However, these encoder manipulation based methods are embedded in the specific encoder (JPEG, or JPEG 2000), and therefore cannot be used directly as part of other standard codec such as the recently proposed H.264 lossless coding. To address this problem in migrating perceptual lossless techniques designed for still image coding to video coding where inter-frame prediction is routinely required, a theoretical framework was recently introduced which is applicable to rate-perceptual-distortion optimization based perceptual video coding implementations [63].

As an example, taking a visual transform domain decomposition model and given an I-frame image, \( \mathbf{x}_t \) and its transform decomposition, \( \mathbf{X}_t \), with a suitable perceptual distortion measure (PDM), the visually filtered coefficient matrix for intra-frame coding which delivers the perceptual lossless compression with maximum perceptual distortion equal to or just under the JNND, \( \mathbf{X}_{t,vf} \), is defined as [62]

\[
\mathbf{X}_{t,vf} = \arg \max_{\mathbf{X}_t} \left\{ \text{PDM} \left( \mathbf{X}_t, \mathbf{X}_t \right) \leq \text{JNND} \right\}
\]  (1)

where \( \mathbf{X}_t \) is visually filtered coefficient matrix from \( \mathbf{X}_t \), (e.g., any interim result from an EBCOT-Embedded Block Coding with Optimal Truncation-coder [44]). The prediction error of the visually filtered transform coefficients of the P-frame at JNND level is defined as

\[
\Delta \mathbf{X}_p = \arg \max_{\Delta \mathbf{X}_p} \left\{ \text{PDM} \left( \mathbf{X}_p, \mathbf{X}_p, \left( \mathbf{X}_p, \Delta \mathbf{X}_p \right) \right) \leq \text{JNND} \right\}
\]  (2)

where \( \mathbf{X}_p \) is the transform decomposition representation of the original P-frame,

\[
\mathbf{X}_p = \left( \mathbf{X}_p, \Delta \mathbf{X}_p \right) = \mathbf{X}_p + \Delta \mathbf{X}_p,
\]  (3)

and \( \Delta \mathbf{X}_p \) is the prediction error of the filtered P-frame transform coefficients. The minimum bitrate \( m_r \) is used to encode the P-frame of the sequence, and defined as

\[
m_r = \min \left( R \left( \Delta \mathbf{X}_p \right), R \left( \mathbf{X}_p \right) \right)
\]  (4)
where $x_p$ is the perceptually lossless intra-frame coded P-frame and $\Delta x_p$ is given in (2). When motion-compensated prediction is used, $x_p$ in Eq. (3) will be replaced by motion-compensated prediction (MP), $x_{mp}$ which obtained from a reference I-frame or a reconstructed previous P-frame.

Another challenging issue is that the existing picture coding standards have adopted complete transforms as mathematical framework for compression, which are not shift-invariant leading to aliasing distortion and constitute a major weakness when used to approximate JND/JNND models in formulating perceptual distortion/quality metrics [36].

4. PICTURE QUALITY EVALUATION
The Mean Squared Error (MSE) and the Peak Signal to Noise Ratio (PSNR) are mathematically tractable and easy-to-compute metrics. However, they can be poor predictor of visual quality, especially when the noise is not additive. The major reason for the overall poor performance of the MSE (or the PSNR) is its assignment of equal importance to all the changes in a visual signal regardless of their perceptual significance.

Some picture quality metrics have been discussed in the last section. Objective image quality assessment can be classified into three categories based on the amount of information used for predicting quality [36]: (1) Full-reference (FR) metrics which use complete reference signal (image/video) information, (2) Reduced-reference (RR) metrics which use only partial information from the reference signal and (3) No-reference (NR) metrics which do not use any reference signal information. With regards to developing an objective PQA (Picture Quality Assessment) algorithm, it can be handled by two broad approaches, as discussed in Sections 4.1 and 4.2.

4.1. Vision Modeling Approach
The vision modeling approach is based on incorporating various HVS characteristics, and aims at simulating the processes of the HVS from the eye to the visual cortex. These metrics are intuitive and appealing since they attempt to account for the properties of the HVS relevant to perceptual quality assessment. The first image and video quality metrics were developed in [64] and [65], respectively. Later other popular HVS-based metrics were developed, including the Visual Differences Predictor (VDP) [66], the Sarnoff JND metric [67], the Moving Picture Quality Metric [68], and the perceptual distortion metric [69].

Although the HVS-based metrics are attractive in theory, they may suffer from some drawbacks. The HVS comprises of many complex processes which work in conjunction rather than independently, to produce visual perception. However, the HVS-based metrics generally utilize results from psychophysical experiments which are typically designed to explore a single dimension of the HVS at a time. In addition, these experiments usually use simple patterns such as spots, bars, and sinusoidal gratings which are much simpler than those occurring in real images. It has been widely acknowledged that a majority, if not all, of PQA metrics in this category adopted vision models devised via threshold vision experiments, which have been contested regarding its extendibility in PQA tasks which commonly fall in supra-threshold vision experiments [36]. It has been found that directly using parameters from the original modeling of HVS experiment where synthetic test patterns were used may not work, re-parameterization of vision model using application specific subjective data, such as those obtained by VQEG for video quality assessment, has been proven to improve PQA performance effectively [56]. In essence, these metrics suffer from drawbacks which mainly stem from the use of oversimplified models describing the HVS. Moreover, the HVS characteristics are not fully understood yet. In addition, due to the complex and highly non-linear nature of the HVS, some of these metrics can be complicated and time-consuming to be used in practice.

4.2. Engineering Approach
Owing to the limitations of the vision-based models discussed above, the engineering based approach has gained popularity during recent years, and it is based primarily on the extraction and analysis of certain features or artifacts in signals, e.g., structural image elements such as contours, or specific distortions (e.g., blocking artifacts) that are introduced by a process, compression technology or transmission link. This does not necessarily mean that such metrics disregard the human vision properties, as they often consider psychophysical effects as well, albeit image content and distortion analysis are the conceptual basis for their design rather than fundamental vision modeling.

With this approach, the assessment of image quality can be considered as a two-stage process: (a) feature extraction, and (b) feature pooling. As for the first stage, the selected features have to form an effective representation of visual quality variations, while the second stage determines the relationship among different features and the perceived visual quality.

A well known FR metric is the structural similarity index measure (SSIM) [70], which is mainly based on the idea of equating the perceived image distortion to the measurement of structural distortion. In SSIM, the mean of quality scores of individual image blocks gives the overall image quality score. Other FR schemes include the Visual Information Fidelity (VIF) index [71] and visual signal-to-noise ratio (VSNR) [72]. The metric known as MSVD (Mean Singular Decomposition) [73] evaluates the quality of each image block based on the error in singular values of
the block, as a result of SVD. Another more recent scheme has been proposed in [74] to measure the change in singular vectors of the reference and distorted images in order to compute the structural changes, and the overall quality is determined by a Minkowski summation. An NR metric using the SVD of local image gradients has been proposed in [75] and used for the proper selection of the parameters of image denoising algorithms. An RR metric that extracts structural features from images, has been proposed in [76]. The overall quality score is then computed as a weighted sum of the features where the weights have been determined by means of subjective experiments.

4.3. Issues Related to Feature Pooling
As for feature (error) pooling, researchers have employed several techniques to fuse the visual features into a quality score such as simple summation based fusion, Minkowski combination, and linear (i.e., weighted) combination. As mentioned in Section 4.2, subjective experiments may be used to compute the weights [76] but such a method is unsuitable for real-time applications. Another method has been developed [77] which involves weighting quality scores with weights determined by local image content, assuming the image source to be a local Gaussian model and the visual channel to be an additive Gaussian model.

Recently, two pooling strategies have been proposed [78] for the SSIM metric. Instead of using a simple mean as the overall quality score, these approaches attempt to weigh the quality scores of different blocks based on visual importance. The first strategy is based on the idea that lower quality regions in images attract more attention than the ones with higher quality; the second strategy uses VA to provide weighting [79]-[80] which is based on the idea that certain regions attract more human attention than the others. The strategy of feature pooling using VA while intuitive may suffer from drawbacks due to the fact that it is not always easy to ascertain regions that attract VA.

5. VISUAL PERCEPTION BASED PERCEPTUAL GRAPHIC RENDERING
There is a growing demand for computer graphics tools that can provide high-end photo-realistic rendering in various applications including gaming, virtual reality, movies, and medical imaging, to name a few. Synthesizing photorealistic images with global illumination is very time-consuming, especially when object models are sophisticated and accurate light simulation is required. Therefore, how to maximize the rendering quality while minimizing the rendering time with limited computational resources, becomes an important research issue.

By exploiting the characteristics of the HVS, significant rendering time can be saved. During the past several decades, various different characteristics of the HVS have been studied and modeled for rendering applications. By using appropriate perceptual models, numerous perceptual image rendering algorithms have been proposed to save the rendering cost.

5.1. VA based perceptual rendering
Since the human’s processing resources are limited, the HVS tends to selectively concentrate on one aspect of the surrounding environment while ignoring the other aspects. The VA models have been exploited in some of the perceptual rendering algorithms. Instead of distributing the rendering computation uniformly over a whole image plane, in VA based perceptual rendering, computation is allocated according to the VA level of image regions. Less amount of computations are performed for regions with lower VA levels, exploiting the fact that human observers are less likely to notice artifacts in non-VA regions. The work in [81] attempted to use a VA model to guide the rendering. The authors proposed a perceptual model which combines a spatio-temporal contrast sensitivity model and a bottom-up VA model [12], and used the proposed model to predict how much error human observers can tolerate at each image pixel position. By allocating less rendering computation to regions with a high error tolerance level, ninety percent of the indirect lighting computation can be saved in radiance caching algorithm.

The work in [82] also attempted to apply a bottom-up VA model in rendering. They improved the accuracy of the VA model of [12], by making use of the 3-D object information such as objects’ distance from image center and size of each object. The scheme in [83] applied the bottom-up VA model in ray tracing. In their scheme, the sampling rate for each pixel is determined by the VA level of that pixel, and higher numbers of sample rays are traced for pixels with a higher VA level.

The top-down attention mechanism has been also used in [84], where a top-down attention model was used in ray tracing. Viewers are given a visual task, e.g., to find out how many teapots exist in a rendered image, and then a task map is generated to predict where people will focus their attention to the given specific task, following the principle of Inattentional Blindness. Using this task map, pixels with higher attention level would receive more samples in the rendering process. In [85], the top-down information was used to guide the allocation of computations in the rendering of dynamic virtual environments, e.g., to predict the object which a user would consider as salient and allocate more rendering computational resources to this object. The work in [86] combined both the bottom-up and top-down attention processes into a VA model and applied it in a ray tracing algorithm. They used the VA model of [12] to generate the VA map and to task information to generate a task map; then they combined the two maps into an importance map and allocated the sample rays accordingly. They also used the importance map of a scene to guide the perceptual rendering [87]. Another top-down stimulus which has also been used for perceptual rendering is the memory schema [88]. It is a knowledge structure related to the human memory, and is used to indicate the context of a scene and
the correlation between objects and the scene. Inconsistent objects which were not expected to be found in a scene were determined as salient, and then models of higher level of detail (LoD) were used for rendering.

5.2. Other aspects of high-level visual perception based perceptual rendering

Besides modeling the VA mechanism, researchers have also worked on perceptual rendering algorithms that take advantage of other aspects of the high-level visual processing. The scheme in [89] took into account the perception of illumination components in global illumination rendering. The global illumination computation is decomposed into several direct and indirect illumination components and illumination components which have lower perceptual importance in perceiving the final image quality, are rendered with a lower number of computational resources. In this case, significant savings in computation can be obtained.

The work in [90] proposed a concept of ‘visual equivalence’, which tries to determine whether two rendered images convey the same perceived information including shape, material and illumination instead of determining whether two images are visually the same. The author of [90] then applied the perceptual ‘visual equivalence’ metric in pre-computed radiance transfer rendering algorithms. Experiments showed that an illumination computation of less accuracy is able to generate images with the same level of perceived information as the reference images. Later in [91], the visual equivalence metric is applied to control of the approximation accuracy of one of the global illumination approximation rendering methods, the virtual point light method. The metric was used to decide where the virtual point light method should be compensated with a path tracing algorithm to achieve a visually equivalent result as the reference image.

6. CONCLUSIONS

Reviewing the field of perceptual processing of digital pictures, significant and concrete progresses have been witnessed over the past decade or two in HVS modeling and adaptation, HVS based picture quality assessment and metrics design, perceptual distortion controlled and perceptual lossless coding of digital pictures including both still images and video, and HVS based perceptual graphic rendering. Embedding HVS based perceptual distortion measures in standard or proprietary picture coders using complete transforms have demonstrated significant improvement in perceived picture quality, making the first step for the constant quality coding. Three issues at least are identifiable, i.e., complete transforms are not shift-invariant and, therefore, are prone to aliasing, JND detection or threshold vision models may not scale well in perceptual lossy coding which operates in supra-threshold environment, and existing HVS models are based on image data in either pixel domain or decomposition domain which requires efficient alternative adaptation for inter-frame (and inter-view) coding of video (and 3-D or multi-view video).

Perceptually driven computer graphics are a developing area.

REFERENCES


