PERCEPTUAL CONTRAST SENSITIVITY BASED VIDEO QUALITY ASSESSMENT IN DCT DOMAIN

Junyong You

Christian Michelsen Research, Bergen, Norway

ABSTRACT

Video quality assessment can be performed by comparing distorted video with the undistorted version by taking human vision system (HVS) into account. A perceptual vision sensitivity model is developed in this paper by systematically integrating visual attention and foveation mechanism into contrast sensitivity function (CSF). The model can accurately estimate a critical frequency beyond which the HVS cannot perceive contrast changes. Subsequently, a video quality metric is proposed by applying the new sensitivity model to a similarity measure between distorted video and reference one in the DCT domain. Experimental results with respect to publicly available video quality databases have demonstrated promising performance of the proposed quality metric.

Index Terms— DCT, perceptual contrast sensitivity, video quality, visual attention

1. INTRODUCTION

Accurate assessment of video quality plays an important role in the chain of digital video services, as unavoidable distortions often occur, e.g., transmission errors over error-prone networks. As humans are the eventual consumers of video services, it is essential to take into account the characteristics of human vision system (HVS) when developing video quality metrics. Sufficient research interests have been attracted in this field and many video quality metrics have been proposed by considering the psychological mechanisms of human perception [1]. As a practical solution, video features characterizing quality distortions can be extracted from degraded video and used to assess how much impact the distortions have on the perceived quality. Pinson and Wolf [2] have proposed a video quality metric (VQM) based on several features representing the degradations on spatial gradient, chrominance, contrast and temporal information. As motion is the most important information in the temporal domain of video sequences, a motion-based video integrity evaluation (MOVIE) has been proposed to predict video quality along motion trajectories [3].

One of the most important characteristics of HVS is contrast sensitivity. Traditionally, the contrast sensitivity function (CSF) shows a typical band-pass filter shape peaking at a certain spatio-temporal frequency (around 4 cycles per degree) with sensitivity dropping off either side of the peak [4]. In addition, due to non-uniform distribution of the photo receptors on the retina, the HVS has the highest resolution around the fixation point in the field of vision and decreases drastically away from that point, as a function of retinal eccentricity [5]. This is called foveated vision mechanism. Due to such a mechanism, the HVS cannot perceive all the details in peripheral areas away from the fixation point, as opposite in the central area. Consequently, viewers need to move their eyes to change the fixation point, such that a complete understanding of the visual stimuli in the field of vision can be acquired. This is strongly related to visual attention mechanism.

Because the HVS cannot perceive all the details in the visual field at once, the brain always selects the most interesting objects and moves to another afterwards. In order to simulate the attention behaviors when viewing visual contents, several computational attention models have been proposed that can predict the distribution of visual attention over an image area [6][7].

As the contrast sensitivity, foveated vision and visual attention mechanisms heavily determines the perception process of visual stimuli, it is crucial to integrate them into video quality assessment. However, existing studies mainly deal with them separately. In [8][9], the spatiotemporal CSF has been integrated into a DCT based scheme to evaluate the difference between distorted video and its reference version. In our earlier work, attention maps produced from computational video attention models have been employed to improve the accuracy of video quality metrics [10][11]. Additionally, foveated vision has also been considered in several video quality metrics, e.g., a foveal signal-to-noise ratio [12], a foveated mean square error measure (FMSE) for video quality [13]. However, existing foveation based quality metrics do not take visual attention mechanism into account, and consequently, they both ignore an important issue of fixation localization. Instead, these metrics usually assume that the fixation points are always located at image center. This is inappropriate as the foveated vision mechanism is essentially determined by the location of fixation and eye movements.

The next Section in this paper first explains a foveation and attention guided contrast sensitivity model, which has been developed in our previous work [14]. By integrating the new contrast sensitivity model into a DCT domain method, an objective video quality metric is presented in Section 3. Experimental results with respect to publicly available video quality databases and performance comparison with state-of-the-art video quality metrics are reported in Section 4. Finally, Section 5 draws concluding remarks.

2. ATTENTION DRIVEN FOVEATED CONTRAST SENSITIVITY

In the traditional foveation based contrast sensitivity function, a contrast threshold (CT) beyond which the HVS cannot perceive contrast changes can be modeled by spatial frequency (f) of visual stimulus and its retinal eccentricity (e) to the fixation point [5], as shown in Eq. (1).

$$CT(f,e) = CT_0 \cdot \exp(\alpha \cdot f \cdot \frac{e+e_2}{e_2}) \quad (1)$$

where CT_0 denotes the minimum contrast threshold, e_2 is the half-resolution eccentricity, and α is the spatial frequency decay constant. A critical frequency (cut-off frequency), f_c, beyond which the contrast will be invisible, can be derived by setting CT to 1.0 (the maximum possible contrast) and solving for f.

$$f_c = \frac{e_2 \cdot \log(1/CT_0)}{\alpha \cdot (e + e_2)} \tag{2}$$

The above sensitivity function describes the situation when viewing still images where temporal changes do not exit. When watching video sequences, the HVS presents a different sensitivity mechanism, which is significantly influenced by the movements of moving objects and the eyes. A common understanding is that the HVS has the highest sensitivity under very low eye movement, e.g., drift or fixation, while it has virtually no sensitivity during the saccade eye movement. Additionally, psycho-visual experiments have demonstrated that the smooth pursuit eye movement can enhance the visual sensitivity, even compared with fixation [15]. More generally, the critical frequency in moving scenarios is actually determined by both the direction and magnitude of the retinal velocity of a moving object, i.e., the difference between the mapped physical movement of the object on the retina and the eye movement [16].

$$f_c' = f_c \cdot \frac{v_c}{|\cos\theta \cdot v_r| + v_c} \tag{3}$$

where v_r and θ denote the magnitude and direction of the retinal velocity, v_c=2deg/sec is the so-called corner velocity.

In addition, other psycho-visual experiments have shown that the visual attention mechanism also significantly affects visual sensitivity. For example, the sensitivity is enhanced at attended stimuli [17][18]. Assuming that there are an attended object and another unattended moving at a same velocity and eccentricity in the visual field, it is naturally that the HVS has a higher sensitivity to the former. In other words, the critical frequency of less attended stimuli can be further decreased such that more redundant information can be removed, while the perception of the overall stimuli is not perceptually degraded. However, it still lacks of quantitative analysis on how much enhancement the attentional intensity impacts on the visual sensitivity. In this work, we can reasonably assume that the impact of attention on visual sensitivity is nonlinear, otherwise the HVS perceives almost nothing in severely less attended areas in the visual field, which is conflict with the late selection theory of attention [19]. Therefore, assuming that the HVS has the highest acuity to the most attended stimulus, i.e., its critical frequency is determined by Eq. (3), the following function is proposed to calculate the critical frequencies with other less attended stimuli in the visual field is performed.

$$f_c'' = f_c' \cdot [\rho + (1-\rho) \cdot AM] \tag{4}$$

where AM (normalized into the range of [0~1]) denotes the attention map depicting the distribution of attention intensity in the visual field, and ρ in the interval [0~1] is a parameter controlling the influence of visual attention on the contrast sensitivity. In our previous work [14], a psycho-visual experiment has been conducted and confirmed that the influence of visual attention on contrast sensitivity can be modeled by Eq. (4). An optimum value 0.7 for ρ has been found from the psycho-visual experiment.

A foveated representation of a video frame can be generated based on the critical frequency derived in Eq. (4) and then modifying the foveal imaging model developed in [5]. As an example, Fig. 1 illustrates a foveated image of a video frame, which is the real representation of the frame generated in the HVS. It can be seen that some areas are blurred, e.g., the peripheral areas away from the fixation point. This is due to the foveated vision mechanism, and the HVS cannot perceive all the details in peripheral areas. Thus, certain information in these areas is lost without being perceived. Consequently, after the eyes have fixated at the most attentive object and perceived necessary information, they must move to another in order to acquire a complete understanding of visual stimuli. This is also the intrinsic function of eye movements. In other words, it can be reasonably assumed that the function of eye movements is to minimize such information loss. The prediction of fixations can be performed based on this assumption. In order to do this, a criterion measuring the information loss between the foveated image and its original version is required. Such information loss is actually dependent on how many high frequency components are contained in the original image. For example, if the original image consists of smooth regions or objects, i.e., containing rarely high frequency components, the foveal operation has less impact on the information loss than that on other images with more high frequency components. Therefore, the following information loss criterion is first defined.

Fig. 1. Illustration of foveated representation of video frame (left: original video frame; right: foveated representation with the proposed sensitivity model; red cross: predicted fixation).

$$IL(O.F) = \frac{2 \cdot \sigma_O \cdot \sigma_F + C}{\sigma_O^2 + \sigma_F^2 + C} \tag{5}$$

where σ_O and σ_F denote the standard deviation of the original image (O) and the foveated version (F), as shown in Fig. 1, and C is a small constant to avoid instability when $\sigma_O^2 + \sigma_F^2 = 0$. This criterion can accurately measure the information loss of a blurred image (i.e., the foveated representation of an image) against its original version with two advantages: 1) boundedness and unique maximum, i.e., $0< IL(O,F)=IL(F,O) \leq 1$ and $IL(O,F)=1$ if and only if $O=F$; 2) if O is smoother, F will also be smoother, and $IL(O,F)$ is closer to 1 accordingly. Furthermore, eye movements are significantly influenced by the attention mechanism, as viewers always tend to fixate at the most attentive object. Thus, the information loss criterion is then adjusted based the attention map:

$$PIL = IL \cdot AM \tag{6}$$

PIL is actually determined solely by the velocity of eye movement, as other parameters, e.g., the physical of moving objects, eccentricity, can both be derived from a video sequence itself. Therefore, fixation prediction can be converted to an optimization problem of finding out the optimum velocity of eye movement in order to minimize PIL. The initial condition is set as that the fixation localizes at the center in the first video frame and then fixation prediction can be performed frame by frame. In addition, the attention map AM can be derived from a computational video attention model. Readers can refer to [14] for the details about the generation of foveated images, derivation of information loss criterion, video attention model, and fixation prediction.

With the predicted fixation, velocities of moving objects in a video sequence (this can be derived from video motion analysis, e.g.,

optical flow), and attention map, the critical frequency can be estimated for each location (or image block) by Eq. (4). This will be employed in a DCT based video quality metric.

3. DCT BASED VIDEO QUALITY ASSESSMENT

DCT can decompose an images into different frequency bands. According to [20], the continuous spatial frequency f at a location can be converted to the discrete normalized frequency by $\min\{0.5, \beta \cdot f\}$, and the conversion factor β is determined by the viewing distance d_v between the eyes and video/image signals and the deviated distance d between the location and the fixation point.

$$\beta = \tan^{-1}\left(\frac{d}{d_v} + \frac{1}{2 \cdot d_v}\right) - \tan^{-1}\left(\frac{d}{d_v} - \frac{1}{2 \cdot d_v}\right) \tag{7}$$

In an $N \times N$ DCT domain, the continuous spatial frequency f is related to the k-th discrete frequency band, and can be calculated as following.

$$f_d = \frac{k}{2N \cdot \beta} \tag{8}$$

Based on the critical frequency f_c'' estimated from the foveated contrast sensitivity model, if distortion occurs with a video sequence, the HVS can only perceive the distortion within the frequency bands lower than f_c''. For those frequency bands higher than f_c'', any distortions with them are not perceptible. Taking a simple example, if an area far away from the fixation point with less attentive degree and high retinal velocity, the corresponding critical frequency is relatively low, and distortions occurred in such an area will be more difficult to be perceived. As an opposite, distortions in an area with low moving speed, high attentive degree and close to the fixation can be easily perceived by viewers, because the critical frequency is high enough such that the distortions are perceptible in all frequency bands. Consequently, an efficient perceived video quality metric can be derived from this mechanism.

DCT coefficients can represent image characteristics, as DC coefficient indicates the energy of an image block and AC coefficients represent its structure. For natural video sequences, adjacent pixels often have high correlations. As explored in the widely used image quality metric SSIM [21], the HVS is very sensitive to structural changes. The distortion on video structure can be measured by the autocorrelation between a distorted video and its reference. According to the Wiener-Khinchin theorem, such autocorrelation can be represented by the DCT of the power spectrum of an image. The local variance of a sub-band of DCT coefficients is strongly related to the autocorrelation in the pixel domain. Thus, the local variance of patches in the DCT domain is employed to simulate the similarity between a distorted video and its reference version, as following in Eq. (9).

$$S_k = \begin{cases} \dfrac{2 \cdot \sigma_k^R(i,j) \cdot \sigma_k^D(i,j) + C}{\sigma_k^R(i,j)^2 + \sigma_k^D(i,j)^2 + C}, & \text{if } f_d \le f_c^* \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where $\sigma_k^R(i,j)$ and $\sigma_k^D(i,j)$ denote the local variance in the k-th DCT band ($1 \le k \le 15$ for $N = 8$) at location (i, j) in an image patch of a reference video frame (R) and distorted frame (D), respectively, and C is a small constant to keep stability of denominator. In this work, the size of an image patch is set to be same as the DCT block, typically 8.

In addition, DC coefficient, i.e., the 0-th band, indicates the energy of an image block while it is not related to image structure. In this work, the covariance indicating cross correlation between a distorted image patch and its reference is employed to measure the perceived difference on DC coefficient.

$$S_0 = \frac{\sigma_0^{R,D}(i,j) + C}{\sigma_0^R(i,j) + \sigma_0^D(i,j) + C} \tag{10}$$

where $\sigma_0^{R,D}(i,j)$ denotes the covariance of the DC coefficients between the reference frame and distorted frame.

The perceived distortion between a distorted image patch and its reference in the DCT domain is measured in Eqs. (9) and (10). In order to evaluate the perceived quality at frame level, a pooling scheme over DCT bands and spatial image patches has been performed. Considering that the foveated contrast sensitivity is already taken into account in Eq. (9), a direct mean over DCT bands is employed, i.e.,

$$Q_{\text{patch}} = \text{mean}(S_k), \quad k = 0, \cdots, 2 \cdot N - 1 \tag{11}$$

Additionally, viewers often tend to perceive image areas with low quality with more severity than high quality areas. Thus, a widely used Minkowski summation with the exponent as 2 is performed on the quality indices over all image patches to derive the perceived quality at frame level.

$$Q_{\text{frame}} = \text{mean}\left(\sqrt{Q_{\text{patch}}^2}\right) \tag{12}$$

A temporal pooling of quality indices over all the frames is required to derive an overall quality of a distorted video. In this work, a 2-step scheme is proposed taking eye movement into account. As the HVS has virtually no sensitivity during the saccadic eye movement, a video sequence is first divided into different clips by removing those frames over which saccade occurs. Saccade can be detected by the velocity of eye movement as explained in Section 2. The temporal pooling is performed on the clips in which only fixation, drift and smooth pursuit eye movements exist. According to our experiments, each consecutive duration of detected saccade is quite short, e.g., in the order of 1-3 frames. Thus, only small parts of video frames are excluded from the temporal pooling due to saccades. Inspired by the observation that large temporal variations of distortions are often more annoying than limited variations at eye fixation level [22], a short-term pooling is proposed within each divided clip.

$$Q_{\text{clip}} = \text{mean}(Q_{\text{frame}}) \cdot \{1 + 3 \cdot \max(D') \cdot g[sign(D')]\} \tag{13}$$

where D' denote the gradient of quality values in the clip, $sign(D')$ is the number of sign changes of D', and g is the Gaussian function centered at 1.0.

In addition, it has been observed from several perceptual video quality studies that the frames in the beginning and the end of a video sequence have more significant impact on the overall video quality, and the tendency of increasing the quality in the end often leads to a better overall quality [11][23]. Taking into account this observation and the influence of temporal variation of quality values on overall quality evaluation, the following long-term temporal pooling is performed to derive an overall perceived video quality.

$$VQ = \left(1 + \frac{1}{TV}\right) \cdot \sum (Q_{\text{clip}} \cdot GPF) \tag{14}$$

where TV denotes the total variation of quality values over all the clips, and GPF is function defined in Eq. (15) and then filtered by the Gaussian filter for 8 times.

$$F(n) = \begin{cases} 1/L, & n \le L/3 \\ 1/(2L), & L/3 < n < 2L/3 \\ 3/(2L), & n \ge 2L/3 \end{cases} \quad \begin{matrix} (n\text{: clip index}) \\ (L\text{: number of clips}) \end{matrix} \tag{15}$$

Table I. Evaluation results on LIVE video quality database

Criteria	PSNR	VQM	MOVIE	FMSE	DVQM	Proposed
Pearson	0.567	0.688	0.814	0.613	0.653	***0.819***
RMSE	9.05	7.97	6.38	8.74	8.36	***6.21***
E-RMSE	6.33	5.37	***3.8***	6.08	5.44	5.75

Table II. Evaluation results on IRCCyN/IVC video quality database

Criteria	PSNR	VQM	MOVIE	FMSE	DVQM	Proposed
Pearson	0.44	0.487	0.487	0.527	0.529	***0.757***
RMSE	0.744	0.723	0.722	0.703	0.695	***0.493***
E-RMSE	0.539	0.517	0.516	0.5	0.604	***0.485***

4. EXPERIMENTS

In order to evaluate the performance of the proposed quality metric, two other well-referenced video quality models, namely VQM [2] and MOVIE [3], have been included. Additionally, a foveation based video quality metric, FMSE [13], and a DCT based quality metric, DVQM [8], have also been implemented in our experiments. Furthermore, the traditional quality metrics, PSNR, was served as a benchmark. Two publicly available video quality databases with annotated subjective scores have been employed to evaluate the above quality models, including the LIVE database (resolution of 768×432 at frame rates of 25ftps and 50fps) [24] and the IRCCyN/ IVC database (resolution of 720×576) [25]. The IRCCyN/IVC database also contains eye-tracking results, which can deduce ground-truth fixation points and fixation density map.

Subjective quality ratings are often compressed at the ends of the rating scales. Therefore, in order to compensate this weakness, as suggested in the recent VQEG report [26], a mapping using the following cubic polynomial function between individually obtained objective video quality measures (VQ) from the above quality models and subjective scores (MOS or DMOS) has been performed.

$$(D)MOS_P = a_0 + a_1 \cdot VQ + a_2 \cdot VQ^2 + a_3 \cdot VQ^3 \qquad (16)$$

where the mapping coefficients $\{a_0, a_1, a_2, a_3\}$ are obtained by fitting the function to maximize the correlation between VQ and (D)MOS, *(D)MOS*$_P$ denotes the mapped video quality. As suggested in [26], the performance of an objective quality metric can be evaluated quantitatively by the three criteria between *(D)MOS*$_P$ and the (D)MOS values, including Pearson correlation, root mean square error (RMSE), and epsilon-insensitive RMSE (E-RMSE) based upon the 95% confidence interval of the subjective scores.

Tables I and II give the evaluation results of the video quality metrics regarding the two databases, respectively. According to the results, the proposed video quality metric shows significantly better performance than the others. A major reason is the accurate simulation of visual contrast sensitivity by taking into about attention and foveation mechanisms, and then an appropriate integration into video quality metric. This can be demonstrated by the comparison that the proposed metric outperforms FMSE and DVQM. Although FMSE takes the foveation mechanism into account, it assumes that the fixation always located at image center. Consequently, the contrast sensitivity is inappropriately modeled. DVQM is based on integrating spatiotemporal contrast sensitivity into DCT based video quality assessment, while it ignores the foveation and attention mechanisms that can heavily determine video quality perception.

However, the E-RMSE result with the LIVE database shows that the proposed metric does not produce a sufficiently robust prediction of perceived video quality. In our opinion, a potential reason lies in inaccurate estimation of fixation locations that can cause deviation of contrast sensitivity modeling in some video frames. In order to demonstrate the influence of fixation localization and attention modeling on the proposed video quality metric, the following experiment has been performed.

The fixation points produced from 30 participants on 20 video sequences have been recorded in the IRCCyN/IVC database. The recorded fixation points can be taken as ground-truth data and an overall single fixation point for each video frame can be derived based on a fixation density map (FDM). The FDM is generated as in the following two steps: 1) For each effectively recorded fixation point, a 2D Gaussian filter with the same resolution as the video frame is generated. The filter is centered at the fixation point and its variance is derived from the viewing angle (2º approximating the size of the fovea) and the viewing distance. 2) All the Gaussian filters corresponding to all the effective fixation points are summed and then normalized into the range of [0~1], that is taken as the FDM. The FDM can be employed to represent the attention map and the point achieving the peak value in the FDM is chosen as the overall single fixation point.

Consequently, the critical frequency in each video frame can be derived as in Section 2 using the ground-truth fixation point and FDM as the attention map. Video quality of the IRCCyN/IVC database is then recomputed and compared against the subjective quality scores. The respective evaluation results are: 0.809 (Pearson), 0.47 (RMSE), and 0.475 (E-RMSE). Thus, the results confirm that accurate localization of fixation points and prediction of attention distribution can significantly improve the accuracy of foveation based video quality assessment.

5. CONCLUSION

This paper has proposed a perceptual contrast sensitivity model by appropriately integrating visual attention mechanism and motion perception into foveated sensitivity. A systematic approach for attention map derivation, fixation prediction and contrast sensitivity has been developed. Consequently, the contrast sensitivity model has been integrated into a DCT based video quality metric. Simulative experiments with respect to publicly available video quality databases have demonstrated that the proposed video quality metric outperforms other compared video quality models. In future work, we aim to improve the prediction accuracy of attention distribution and fixation, as our experiment has also demonstrated that using ground-truth data on attention distribution and fixations can significantly improve the performance of the proposed quality metric.

6. REFERENCES

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

[2] M. Pinson, and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," *IEEE Trans. Broadcast*., vol. 50, no. 3, pp. 312-322, Sep. 2004.

[3] K. Seshadrinathan, and A. C. Bovik, "Motion Tuned Spatiotemporal Quality Assessment of Natural Videos," *IEEE Trans. Image Process*., vol. 19, no. 2, pp. 335-350, Feb. 2010.

[4] C. A. Burbeck and D. H. Kelly, "Spatiotemporal Characteristics of Visual Mechanisms: Excitatory-inhibitory Model," *J. Opt. Soc. Am.*, vol. 70, no. 9, pp. 1121–1126, 1980.

[5] W. S. Geisler, and J. S. Perry, "A Real-time Foveated Multi-resolution System for Low-bandwidth Video Communication," in *Proc. SPIE Human Vision and Electron. Imaging*, vol. 3299, San Jose, CA, USA, Jan. 1998, pp. 294-305.

[6] L. Itti, and C. Koch, "Computational Modeling of Visual Attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194-203, Mar. 2001.

[7] J. Harel, C. Koch, and P. Perona, "Graph-based Visual Saliency", in *Advances in Neural Information Process. Systems,* 2007, pp. 681–688.

[8] F. Xiao, "DCT-based Video Quality Evaluation," MSU Graphics and Media Lab (Video Group), 2000, available online: http://compression.ru/video/quality_measure/vqm.pdf.

[9] A. B. Watson, J. Hu, and J. F. McGowan III, "DVQ: A Digital Video Quality Metric based on Human Vision," *J. Electronic Imaging*, vol. 10, no. 1, pp. 20-29, Jan. 2001.

[10] J. You, A. Perkis, M. M. Hannuksela, and M. Gabbouj, "Perceptual Quality Assessment based on Visual Attention Analysis," in *Proc. ACM Multimedia*, Beijing, China, Oct. 2009, pp. 561-564.

[11] J. You, J. Korhonen, A. Perkis, and T. Ebrahimi, "Balancing Attended and Global Stimuli in Perceived Video Quality Assessment," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1269-1265, Dec. 2011.

[12] 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.

[13] S. Rimac-Drlje, M. Vranješ, and D. Žagar, "Foveated Mean Squared Error - A Novel Video Quality Metric," *Multimedia Tools Appl.*, vol. 49, no. 3, pp. 425-445, Sep. 2010.

[14] J. You, "Video Gaze Prediction: Minimizing Perceptual Information Loss," in *Proc. IEEE Int. Conf. Multimedia and Expo*, pp. 438-443, Melbourne, Australia, Jul. 2012.

[15] A. C. Schütz, D. I. Braun, D. Kerzel, and K. R. Gegenfurtner, "Improved Visual Sensitivity During Smooth Pursuit Eye Movements," *Nature Neurosci*., vol. 11, no. 10, pp. 1211-1216, Oct. 2008.

[16] M. P. Eckert, and G. Buchsbaum, "The Significance of Eye Movements and Image Acceleration for Coding Television Image Sequences," in *Digital Images and Human Vision*, A. B. Watson (Ed.), pp. 89-98, Cambridge, Massachusetts, MIT Press, 1993.

[17] F. Pestilli, and M. Carrasco, "Attention Enhances Contrast Sensitivity at Cued and Impairs It at Uncued Locations," *Vision Res*., vol. 45, no. 14, pp. 1867-1875, Jun. 2005.

[18] I. Motoyoshi, "Attentional Modulation of Temporal Contrast Sensitivity in Human Vision," *PLoS ONE*, vol. 6, no. 4, e19303, Apr. 2011.

[19] H. E. Pashler, The Psychology of Attention, MIT Press, Cambridge, MA, 1998.

[20] Z. Wei and K. Ngan, "Spatio-temporal Just Noticeable Distortion Profile for Grey Scale Image/Video in DCT Domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 337–346, 2009.

[21] Z. Wang, A. C. Bovik, H. R., Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process*., vol. 13, no. 4, pp. 600-612, Apr. 2004.

[22] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering Temporal Variations of Spatial Visual Distortions in Video Quality Assessment," *IEEE J. Select. Topics Signal Process*., vol. 3, no. 2, pp. 253–265, Apr. 2009.

[23] M. Zink, O. Künzel, J. Schmitt, and R. Steinmetz, "Subjective Impression of Variations in Layer Encoded Videos," *Lecture Notes Comput. Sci*., vol. 2707, pp. 137–154, Jan. 2003.

[24] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. L. Cormack, "Study of Subjective and Objective Quality Assessment of Video," *IEEE Trans. Image Process*., vol. 19, no. 6, pp. 1427-1441, Jun. 2010.

[25] IRCCyN/IVC Eye-tracker SD Database, available online: http://ivc.univ-nantes.fr/en/databases/Eyetracker_SD_2008_11/?article555.

[26] Video Quality Expert Group, "Report on the Validation of Video Quality Models for High Definition Video Content," Jun. 2010, available online: http://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx.