Attention modeling for video quality assessment

balancing global quality and local quality

You, Junyong; Korhonen, Jari; Perkis, Andrew

Published in:
proceedings ICME

Publication date:
2010

Document Version
Early version, also known as pre-print

Link back to DTU Orbit

Citation (APA):
You, J., Korhonen, J., & Perkis, A. (2010). Attention modeling for video quality assessment: balancing global quality and local quality. In proceedings ICME (pp. 914-919). IEEE.

ATTENTION MODELING FOR VIDEO QUALITY ASSESSMENT: BALANCING GLOBAL QUALITY AND LOCAL QUALITY

Junyong You[1], Jari Korhonen[2], Andrew Perkis[1]

1 Centre for Quantifiable Quality of Service (Q2S) in Communication Systems*, Norwegian University of Science and Technology, Trondheim, Norway;
2 Department of Photonics Engineering, Technical University of Denmark, Lyngby, Denmark
Email: junyong.you@q2s.ntnu.no, jark@fotonik.dtu.dk, andrew@iet.ntnu.no

ABSTRACT

This paper proposes to evaluate video quality by balancing two quality components: global quality and local quality. The global quality is a result from subjects allocating their attention equally to all regions in a frame and all frames in a video. It is evaluated by image quality metrics (IQM) with averaged spatiotemporal pooling. The local quality is derived from visual attention modeling and quality variations over frames. Saliency, motion, and contrast information are taken into account in modeling visual attention, which is then integrated into IQMs to calculate the local quality of a video frame. The local quality of a video sequence is calculated by pooling local quality values over all frames with a temporal pooling scheme derived from the known relationship between perceived video quality and the frequency of temporal quality variations. The overall quality of a distorted video is a weighted average between the global quality and the local quality. Experimental results demonstrate that the combination of the global quality and local quality outperforms both sole global quality and local quality, as well as other quality models, in video quality assessment. In addition, the proposed video quality modeling algorithm can improve the performance of image quality metrics on video quality assessment compared to the normal averaged spatiotemporal pooling scheme.

Keywords— Quality metric, visual attention, saliency, motion, temporal pooling

1. INTRODUCTION

Video quality assessment plays an important role in development and optimization of video compression and communication schemes. Subjective quality assessment is considered to be the most reliable way to evaluate the quality of audio and video presentations, but it is time-consuming. Many objective quality metrics for automated quality evaluation of distorted video have been proposed. They can be classified into three categories according to the availability of reference information: full reference, reduced reference, and no reference. Most video quality metrics take into account the attributes of the human visual system (HVS), spatiotemporal video activities, and use certain pooling schemes to combine distortion information over different channels. For example, the perceptual distortion model (PDM) [1] adopted the spatial and temporal mechanisms of the HVS to calculate distortion information in different channels, and then the Minkowski summation was used in pooling spatial and temporal errors between the reference and distorted video sequences. In addition, as an important clue in video, motion information is used to tune the quality degradation on spatial and temporal fidelity between the reference and distorted video sequences in the MOVIE quality model [2], which can evaluate motion quality along computed motion trajectories.

Some video quality metrics are based upon spatial image quality evaluation plus temporal pooling schemes. Image quality assessment is a widely studied issue, and a number of researchers have contributed significant research in the design of image quality assessment algorithms, claiming to have made headway in their respective domains. Under an assumption that the HVS is highly adapted for extracting structural information from a scene, a structural similarity (SSIM) measure [3] can be constructed based on luminance comparison, contrast comparison, and structure comparison between the reference and distorted images. SSIM metric was then extended to measure video distortion by integrating chrominance and motion information [4]. In order to assess the video quality, spatial distortions of all frames are pooled temporally to obtain an overall quality, in which certain video characteristics can be taken into account [5]. Currently, the widely used temporal pooling schemes are Minkowski summation and direct average over frames. However, the temporal mechanism in video quality assessment has not been investigated adequately, and choosing an appropriate temporal pooling scheme for a certain scenario is still an open issue. We have investigated some spatial and temporal pooling schemes for packet loss video streams [6]. The perceived video quality is also influenced by time-varying quality characteristics [7]. For example, the frequency of temporal quality change has a significant influence on the perceived quality. In this work, we present a temporal pooling scheme based on existing conclusions to model the relationship between the perceived video quality and temporal quality changes over time.

Visual attention is an important attribute of the HVS, while its capability in video quality assessment has not been explored adequately. Many psychological and physiological experiments have demonstrated that the human attention is not allocated equally to all regions in the visual field, but focused on certain attention regions [8]. Lu et al. presented a perceptual quality significance map to reflect the modulatory aftereffects of visual attention and evaluated its application in a just-noticeable-difference (JND) model [9]. We have proposed a visual attention based quality

* “Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence” appointed by the Research Council of Norway, funded by the Research Council, NTNU and UNINETT.

978-1-4244-7492-9/10/$26.00 ©2010 IEEE

metric [10], in which video quality is derived by some quality features in extracted attention regions. It shows promising performance in evaluating the quality of video sequences with general distortion types, such as compression and noise. Based on the saliency attention model in [8], Feng et al. [11] investigated a few weighting methods on the pixels in salient regions for mean squared error (MSE), mean absolute difference (MAD), and SSIM metrics for evaluating video quality degradation. However, it was found that incorporating visual attention into quality assessment is not always advantageous [12], especially for packet loss video streams [6].

In this work, we propose to divide the video quality into two components: global quality and local quality. The global quality is determined by general distortions in each video frame that can be calculated by image quality metrics and a direct average over all frames. Based on the global quality in each frame, the local quality will be adjusted due to some particular video attributes, such as visual attention. In addition, the frequency of quality variations over frames also has an impact on the local quality. Our understanding is that the perceived quality of a distorted video is a balancing combination between the global quality and the local quality.

The remainder of this paper is organized as follows. Section 2 analyzes the quality components and the computation. The derivation of local quality based on visual attention and a temporal pooling scheme is presented in Section 3. We evaluate the performance of the proposed algorithm with respect to two publicly available video quality databases in Section 4, and finally, some concluding remarks are given in Section 5.

2. VIDEO QUALITY: GLOBAL AND LOCAL QUALITY

The formation of human perception is a complicated process. In subjective video quality assessment, the neuropsychological mechanism in assessing the quality of a distorted video is still an open issue. Yang et al. [13] presented a video quality metric for assessing spatial quality of temporally interpolated frames by separating the quality into two principle parts, global quality estimator and local distortion estimator. In this work, we propose to divide video quality into two components, global and local quality, in a different approach. The global quality is evaluated by coarse impression when subjects watch a distorted video; whilst the local quality is a complement to the global quality. We assume that the global quality is a result from subjects allocating their attention equally to all regions in a frame and all frames in a sequence. More specifically, the local quality can be affected by many factors, such as distortion in certain regions to which subjects pay more attention, and certain frames with particular visual characteristics or quality levels.

In this work, we employed four IQMs, namely peak signal-to-noise ratio (PSNR), SSIM [3], multi-scale SSIM (MSSIM) [14], and a modified PSNR based on the HVS (PSNR-HVS-M) [15], to compute image quality degradation on each frame. MSSIM is an extension of SSIM, which iteratively applies a low-pass filter in the reference and distorted images and down-samples the filtered images by a factor of 2. At each image scale, the contrast comparison and the structure comparison are calculated, respectively. The luminance comparison is computed only at the highest scale. The overall MSSIM measure is obtained by combining the measures at different scales. PSNR-HVS-M is a modification of PSNR based on a model of visual between-coefficient contrast masking of discrete cosine transform (DCT) basis functions. This model can calculate the maximal distortion that is not visible at each DCT coefficient due to the between-coefficient contrast masking. After obtaining quality value of each frame in a video sequence, the global quality was calculated by averaging directly the quality values over all frames.

Two factors, visual attention and temporal quality change, were taken into account in calculating the local quality. In the next section, a visual attention model will be constructed that can detect attention regions in each frame based on saliency, motion, and contrast information. These four IQMs were used again to compute quality values in the attention regions with integrating derived attention map values. On the other hand, some subjective video quality assessment studies have demonstrated that the frequency of quality variations over time also influence the overall quality [7]. In order to combine local quality values over all frames, a temporal pooling scheme was proposed in this work by adopting two conclusions from [7]: (1) the more frequent quality variation, the worse perceived quality; (2) the frames in the beginning and the end of a video sequence have more significant impact on the overall quality, and the tendency that increasing the quality of frames in the end leads to a better perceived quality. Therefore, we compute the local quality (*LQ*) of a sequence based on the local quality values over all frames using the following temporal pooling scheme:

$$LQ = (1 + \frac{1}{TV}) \cdot \sum_{k} (FLQ \cdot TPF) \qquad (1)$$

where *FLQ* denotes the local quality computed in every frame, which will be presented in detail in Section 3.2, *k* is the frame index, *TV* denotes the total variation of the local quality values over all frames, and *TPF*, as illustrated in Fig. 1, is the filtered result of a function defined in Equation (2) by the Gaussian filter for several times (typically 8).

$$F(k) = \begin{cases} \frac{1}{L}, & k \le \frac{L}{3} \qquad (k\text{: frame index}) \\ \frac{1}{2L}, & \frac{L}{3} < k < \frac{2L}{3} \qquad (L\text{: video length}) \\ \frac{3}{2L}, & k \ge \frac{2L}{3} \end{cases} \qquad (2)$$

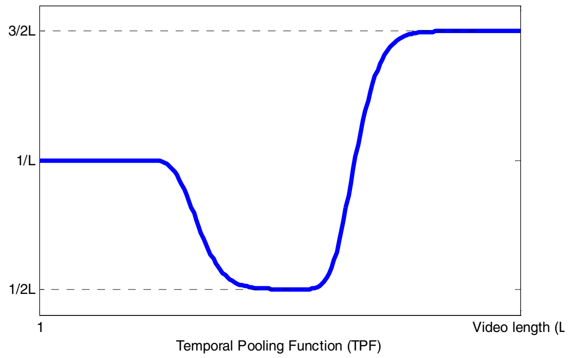


Fig. 1. Temporal pooling function (*TPF*)

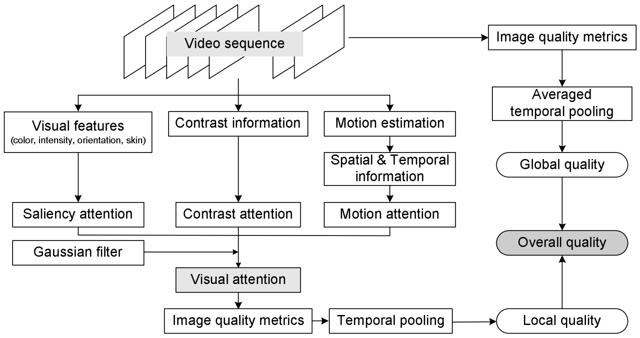


Fig. 2 Flowchart of the proposed video quality model

Finally, the overall quality (*OQ*) of a distorted video is a weighted average between the global quality (*GQ*) and the local quality (*LQ*):

$$OQ = W \cdot GQ + (1 - W) \cdot LQ \tag{3}$$

where *W* denotes the weight. Figure 2 gives the flowchart of the proposed video quality model.

3. LOCAL QUALITY ASSESSMENT BASED ON VISUAL ATTENTION AND TEMPORAL POOLING

3.1. Modeling visual attention

Visual attention has been a widely studied issue in computer vision technology, while its application in video quality assessment is still not explored adequately. Inspired by the behavior and the neuronal architecture of the early primate visual system, a visual attention system has been presented by combining multi-scale image features, such as color, intensity, orientations and other information, into a single topographical saliency map [8]. In this work, we used the SaliencyToolbox developed by Walther and Koch [16] to detect salient regions and computed a saliency map in a video frame in the reference video, based on four visual features: color, intensity, orientation, and skin information. Because there are no rich face and text contents in the video quality databases in our experiments and the skin information has been taken into account in the saliency model, the face and text detection described in our previous work [10] was not employed in this work. Furthermore, motion information has a significant effect in determining attention regions in video sequences [17]. Thus, we adopted the motion attention model in [17] as a component for constructing the attention model in this study. The motion attention model was constructed based on motion vectors, and a spatial window and a temporal sliding window were used in computing the spatial and temporal coherence inductors in the motion attention model. As most sequences in video quality tests contain only single scene, we calculated spatial information (*SI*) and temporal information (*TI*) indices as follows, defined in ITU-T Rec. P.910 [18], in order to determine the sizes of the spatial window and the temporal window.

$$\begin{cases} SI = \max_{time}\{std_{space}[Sobel(F_k)]\} \\ TI = \max_{time}\{std_{space}(F_k - F_{k-1})\} \end{cases} \tag{4}$$

where F_k denotes the *k*-th frame luminance image, *Sobel* is a Sobel filter, and *std* denotes the standard deviation. If a video frame contains more complex contents, a smaller spatial window will be used. Similarly, a shorter temporal window is employed if a video sequence has more complex temporal activities. A motion attention map of a frame can be obtained from the motion attention model.

In addition, it was found that contrast information in the visual field is another important factor in human attention detection [19] and quality assessment [20]. Human usually pay more attention to those regions that have higher contrast levels. Thus, a video frame was divided into different blocks, and the standard deviation of each block was used to denote the contrast information of this block.

After obtaining the saliency map (*S*), the motion attention map (*M*), and the contrast map (*C*) of a frame, these map values were normalized into the interval of [0, 1], respectively. According to our experiments on training sequences, we found that the contrast attention information has a bit less influence on video quality assessment, compared to the saliency and motion attention information. Subsequently, a normalized Gaussian filter (*G*) with the center located at the middle of frame was performed on the weighted average of these above three maps as in Equation (5). Since human usually pay more attention to the regions close to the center of frame, we use such Gaussian filter to assign a weight to the position of attention regions. The weighted attention map (*A*) can depict the distribution of visual attention regions over a video frame.

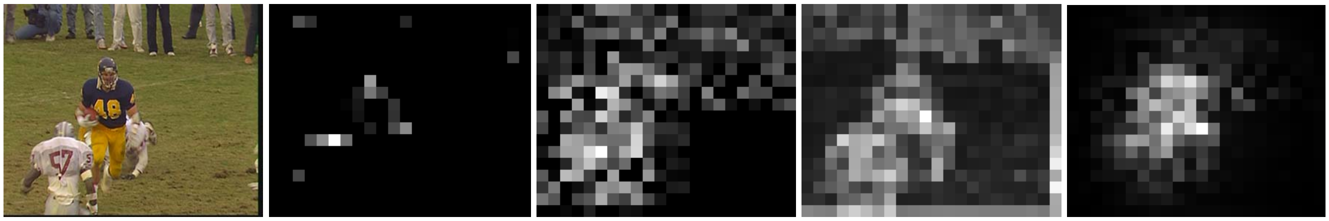
$$A = G \cdot (S + M + 0.5 \cdot C) \tag{5}$$

Figure 3 illustrates a frame image and the attention maps.

3.2. Local quality assessment

The local quality is determined by attention regions and the frequency of quality variations over video frames. The attention regions and the attention map were computed in the reference video. For each frame, a local quality value is computed using image quality metrics based on attention regions. In this work, we have tested three approaches to compute the local quality of a frame as follows.

- We calculated a quality map between a frame in the reference



(a) original image (b) saliency map - *S* (c) motion map - *M* (d) contrast map - *C* (e) attention map - *A*

Fig. 3. Frame image in Football sequence and attention maps

video and the corresponding distorted frame by using PSNR (based on MSE), SSIM, MSSIM, and PSNR-HVS-M, respectively. The local quality value was then computed by a weighted average over all pixels on the quality map, the weight at each pixel being the attention map value.

- As human perception is always decided by a subset of an image in the visual field, rather than the whole image. Therefore, the second approach was to select those regions with higher attention map values and the local quality was a weighted average in the selected regions. In the second approach, we further tested two different methods. The first one was to use the attention map values as the weights, while the other one was to use 1 as the weights in the selected regions, i.e. the direct average of the quality map over the selected regions was taken as the local quality. Furthermore, different thresholds were tested in selecting the candidate attention regions.
- Because the attention model was constructed from different blocks in a video frame, the last approach was to select appropriate attention regions that contain some image blocks first, and then the image quality metrics were applied on these blocks, respectively. The local quality was taken as a direct average over these blocks.

Based on a lot of experiments with respect to adequate subjective quality databases (e.g. the VQEG FR-TV Phase I data sets, etc.), we found that the first method in the second approach has the best performance statistically for the training video datasets. In addition, we found that the best performance was achieved when the threshold was chosen as those 20% regions over all attention regions were participated in the computation of the local quality. Thus, the local quality is computed as follows. After detecting the attention regions and computing the attention map values, the attention map values were sorted in a descending order. Twenty percent of blocks with the highest attention map values over the attention regions were selected as the candidates. The local quality of a frame was computed by the weighted average between the quality map that was calculated by the image quality metrics and the attention map in the attention regions, i.e.

$$FLQ = \underset{\Omega}{AVG}(A \cdot Q) \tag{6}$$

where Ω denotes the selected attention regions, A denotes the attention map values, Q is the quality map value, and AVG denotes the averaging operation.

After computing the local quality of each frame, a local quality curve over all frames in a video sequence can be obtained. The change of quality values between different frames (even two consecutive frames) on this curve might be drastic. However, the formation of human perception is a successive and steady process, in which the perception at the current point is affected by the previous points. Thus, we used the left half of a Gaussian smoothing filter to smooth the local quality curve. Subsequently, the local quality of a video sequence was computed by Equation (1) based on the smoothed curve and the temporal pooling scheme.

4. EXPERIMENT AND DISCUSSIONS

To evaluate the performance of the proposed algorithm, we employed two public subjective video quality databases: EPFL-PoliMI database [21] and LIVE database [22]. These two databases were not participated in determining the best approach as described in Section 3.2 and some other parameters, such as the threshold setting and the derivation of the temporal pooling scheme. Thus, the evaluation results on these two quality databases can provide a fair justification on the performance of the proposed algorithm. In addition, we also applied two other metrics: the VQM proposed in [23] and PSNR, as benchmarks. The VQM is considered to be one of the best objective quality models at present, and PSNR is a widely used metric in evaluating the performance of video coding and transmission schemes.

Because different subjective quality assessment may use different quality scales, a nonlinear regression operation between the metric results (VQ) and the subjective scores (MOS) was performed by the logistic function in Equation (7), as suggested in a Video Quality Experts Group (VQEG) report [24].

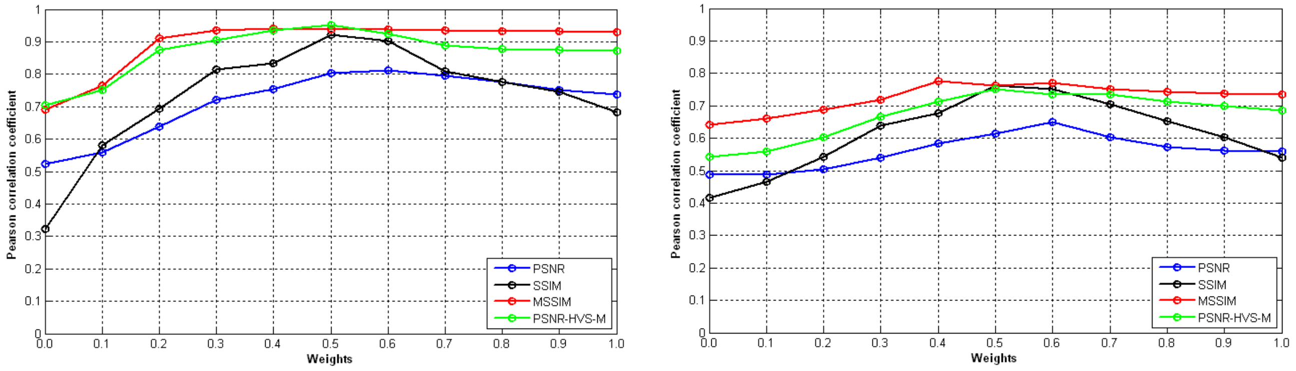
$$MOS_P = \frac{a_1}{1+\exp[-a_2 \cdot (VQ - a_3)]} \tag{7}$$

The nonlinear regression function was used to transform the set of metric values to a set of predicted MOS values, MOS_P, which were compared against the actual subjective scores and then resulted in a criterion: Pearson correlation coefficient.

First, we evaluated the performance of the proposed temporal pooling scheme by combining quality values over all frames. A quality value of each frame was first calculated by PSNR, SSIM, MSSIM and PSNR-HVS-M, respectively. Three temporal pooling schemes on quality values over all frames were tested. The first one was to use the proposed temporal pooling scheme from Equation (1). Second, the direct average of quality values over all frames was taken as the video quality. Finally, the Minkowski summation with the exponent set as 2 was selected as the third

Table 1. Evaluation results of temporal pooling schemes

Database	Pooling scheme	PSNR	SSIM	MSSIM	PSNR-HVS-M
EPFL-PoliMI	Proposed scheme	0.772	0.846	0.945	0.889
	Direct average	0.739	0.682	0.930	0.874
	Minkowski summation	0.721	0.680	0.928	0.868
LIVE	Proposed scheme	0.585	0.553	0.744	0.703
	Direct average	0.560	0.541	0.734	0.685
	Minkowski summation	0.559	0.541	0.734	0.678



(a) Evaluation results on EPFL-PoliMI database

(b) Evaluation results on LIVE database

Fig. 4. Evaluation results of the proposed algorithm in terms of Pearson correlation coefficients

scheme. Table 1 gives the evaluation results of these temporal pooling schemes with EPFL-PoliMI and LIVE databases, respectively, in terms of Pearson correlation coefficient. According to the evaluation results, the proposed pooling scheme is slightly better than either direct average or Minkowski summation, no matter which image quality metric is employed. Actually, the correlation gain is about 1-7% by using the proposed pooling scheme on the training databases, depending on the video dataset.

In our experiments, 11 different weights W={0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0} were used. When the weight is set as 1 or 0, only the sole global or local quality defines the overall quality. Figure 4 gives the Pearson correlation coefficients of the proposed algorithm over different weights, with respect to the EPFL-PoliMI database and the LIVE database, respectively. In addition, Table 2 gives the evaluation results of the VQM and PSNR on these two quality databases, the best correlation of the proposed algorithm, and which image quality metrics achieves the best performance, as well as the respective weights.

According to the evaluation results, the performance of almost all metrics is worse with the LIVE database than with the EPFL-PoliMI database. In our opinion, the reason might be that the LIVE database contains many different video contents and distortion types, whereas the EPFL-PoliMI database considers packet loss artifacts only.

According to Fig. 4, the combination between the global and local quality performs better than either the sole global quality or the sole local quality, and the most appropriate weights locate in the range of [0.4, 0.7]. Thus, the combination of the global quality and local quality is a feasible solution to video quality assessment. There might be some other suitable combination approaches between the global and local quality, which will be studied in future work. In addition, the performance of the sole local quality is worse than the sole global quality regardless of what image quality metrics were used, especially for EPFL-PoliMI database, which partially confirms the assumption that incorporating visual attention into quality assessment is not advantageous in all cases. Our understanding is that this might be because subjects usually evaluate the video quality according to those regions with the most severe degradation, e.g. areas impacted by packet losses, even though these regions are not attention regions.

Compared to the effect on SSIM and PSNR-HVS-M, the effect of the proposed algorithm on MSSIM is not as evident as on SSIM and PSNR-HVS-M. We believe this is because the visual attention model is not very suitable for MSSIM, since MSSIM combines distortions over different scales. According to the comparison between the proposed model and the VQM, our method is not worse than the latter. However, in this work we have used only image quality metrics that do not take into consideration quality features expressing video attributes. The VQM utilizes seven video quality features that were designed particularly for video quality assessment. Therefore, in the future work we will design suitable video quality features that can produce more accurate prediction of the global and the local quality.

5. CONCLUSION

In this paper, we have presented a video quality model for assessing objective video quality based on visual attention and other video attributes. In our model, video quality is divided into two components: global quality and local quality. The global quality is calculated by image quality metrics and direct spatiotemporal averaging method. The local quality is more complicated, determined from visual attention and the frequency of quality variations over video frames. An appropriate attention model was constructed for video quality assessment, and a temporal pooling scheme was derived based on some existing conclusions and a number of training experiments. The experimental results demonstrated that the proposed algorithm can improve the performance of image quality metrics with direct spatiotemporal pooling on video quality assessment. Other suitable

Table 2. Evaluation results of VQM, PSNR, and proposed algorithm

Database	VQM	PSNR	Proposed	Metric	Weight
EPFL-PoliMI	0.956	0.739	0.951	PSNR-HVS-M	0.5
LIVE	0.686	0.560	0.776	MSSIM	0.4

video quality features and temporal characteristics of video content, as well as more suitable visual attention model that can be integrated into the proposed algorithm will be investigated in future work.

REFERENCES

[1] S. Winkler, Digital Video Quality: Vision Models and Metrics, John Wiley & Sons, 2005.

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

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, et al., "Image Quality Assessment: from Error Visibility to Structural Similarity," IEEE Trans. Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[4] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," Signal Processing: Image Communication, vol. 19, no. 2, pp. 121-132, Feb. 2004.

[5] A. Ninassi, O. L. Meur. P. L. Callet, and D. Barba, "Considering Temporal Variations of Spatial Visual Distortions in Video Quality Asssessment," IEEE JSTSP Special Issue, Visual Media Quality Assessment, vol. 3, no. 2, pp. 253-265, Apr. 2009.

[6] J. You, J. Korhonen, and A. Perkis, "Spatial and Temporal Pooling of Image Quality Metrics for Perceptual Video Quality Assessment on Packet Loss Streams," IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp. 1002-1005, Dallas, Texas, USA, Mar. 2010.

[7] M. Zink, O. Künzel, J. Schmitt, and R. Steinmetz "Subjective Impression of Variations in Layer Encoded Videos," Lecture Notes in Computer Science, vol. 2707, pp. 137-154, Jan. 2003.

[8] L. Itti, and C. Koch, "Computational Modeling of Visual Attention," Nature Reviews Neuroscience, vol. 2, no. 3, pp. 194-203, Mar. 2001.

[9] Z. Lu, W. Lin, X. Yang, et al. "Modeling Visual Attention's Modulatory Aftereffects on Visual Sensitivity and Quality Evaluation," IEEE Trans. Image Processing, vol. 14, no. 11, pp. 1928-1942, Nov. 2005.

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

[11] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency based Objective Quality Assessment of Decoded Video Affected by Packet Losses," in Proc. IEEE Int. Conf. Image Processing, pp. 2560-2563, California, USA, Oct. 2008.

[12] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, "Does Where You Gaze on an Image Affect Your Perception on Quality? Applying Visual Attention on Image Quality Metric," in Proc. IEEE Int. Conf. Image Processing, vol. 2, pp. II 169-172, San Antonio, Texas, USA, Sep. 2007.

[13] K-C Yang, A-M Huang, T. Q. Nguten, et al. "A New Objective Quality Metric for Frame Interpolation Used in Video Compression," IEEE Trans. Broadcasting, vol. 54, no. 3, pp. 680-690, Sep. 2008.

[14] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale Structural Similarity for Iimage Quality Assessment", in Proc. IEEE Asilomar Conf. Signals, Systems, and Computers, pp. 1398-1402, Pacific Grove, California, USA, Nov. 2003.

[15] N. Ponomarenko, F. Battisti, K. Egiazarian, et al., "On Between-coefficient Contrast Masking of DCT Basis Functions," in Proc. Int. Workshop Video Processing and Quality Metrics, Scottsdale, Arizona, USA, Jan. 2007.

[16] D. Walther, and C. Koch, "Modeling Attention to Salient Proto-objects," Neural Networks, vol. 19, no. 9, pp. 1395-1407, 2006.

[17] Y-F Ma, L. Lu, H-J Zhang, and M. Li, "A User Attention Model for Video Summarization," in Proc. ACM Int. Conf. Multimedia, pp. 533-542, Juan-les-Pins, France, Dec. 2002.

[18] ITU-T Recommendation P.910, Subjective Video Quality Assessment Methods for Multimedia Applications, ITU, Sep. 1999.

[19] W. Osberger, and A. M. Rohaly, "Automatic Detection of Regions of Interest in Complex Video Sequences," in Proc. SPIE Human Vision and Electronic Imaging VI, vol. 4299, pp. 361-372, 2001.

[20] J. You, F. N. Rahayu, U. Reiter, and A. Perkis, "HVS-based Image Quality Assessment for Digital Cinema," in Proc. SPIE Image Quality and System Performance VII, Jan. 2010.

[21] F. D. Simone, M. Naccari, M. Tagliasacchi, et al. "Subjective Assessment of H.264/AVC Video Sequences Transmitted over a Noisy Channel," in Proc. Int. Workshop Quality of Multimedia Experience, pp. 204-209, San Diego, California, USA, Jul. 2009.

[22] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "A Subjective Study to Evaluate Video Quality Assessment Algorithms," Proc. SPIE Human Vision and Electronic Imaging XV, Jan. 2010.

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

[24] VQEG, "Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II (FR-TV 2)," VQEG, Aug. 2003.