Visual Saliency and Distortion Weighting Based Video Quality Assessment

Lin Zhu, Li Su, Qingming Huang, and Honggang Qi

Graduate University of Chinese Academy of Sciences, Beijing, China
{lzhu,lsu,qmhuang,hgqi}@jdl.ac.cn

Abstract. Video quality assessment (VQA) is very important in many video processing applications. For example, the rate-distortion (RD) optimization in video coding needs an efficient distortion metric to assess the RD cost of candidate coding parameters. However, most existing metrics employ little visual perceptual information, or some are too complex to meet real-time requirement. In this paper we propose a new model called saliency and distortion weighted structural similarity index with temporal pooling strategy (SDTW-SSIM). In the proposed model, spatial and temporal saliency is obtained from the referenced video. Besides, a distortion weighting map is employed to give a full description of visual attention. To better present the perceptual properties of videos, both frame and sequence level saliency features are taken into account. Experimental results show that, compared with state-of-the-art methods, the proposed method performs well on both computational efficiency and assessment accuracy.

Keywords: Video Quality Assessment, Visual Attention, Motion Estimation, Distortion Weighting, Structural Similarity.

1 Introduction

With the rapid development of multimedia technology, video service is getting more and more popular. Therefore, video quality measurement plays a fundamentally important role in video processing applications. A straightforward way of evaluating video quality is achieved by subjective testing [1]. However, it has to follow strict evaluation conditions, and is laborious and expensive. Thus, objective quality assessment metrics that can reflect the perceived video quality are necessary. For example, most recently video coding algorithms use the rate-distortion optimization (RDO) to remove the redundant information, during which, objective video quality is calculated to evaluate the video coding distortion.

Traditional distortion metrics usually calculate the video's mean squared error (MSE) or peak signal-to-noise ratio (PSNR), as a result often deviating from the human perceptual feelings. In order to automatically assess the quality of videos in a perceptually consistent manner, human visual system (HVS) has been introduced into this field by modeling its physiological and psychological features [2, 3]. Considering

W. Lin et al. (Eds.): PCM 2012, LNCS 7674, pp. 546–555, 2012.

that the HVS is an extremely complicated system and there still lacks full understanding of it currently, improving methods are required.

With the knowledge that natural image signals are highly structured, a measure of structural similarity (SSIM) [4] to approximate the perceived image quality has been employed, which outperforms many state-of-the-art perceptual image quality metrics. The visual information fidelity (VIF) [5], which is based on visual statistics, models images as realizations of Gaussian Scale Mixtures in the wavelet domain. Though VIF and several recently proposed methods deliver better consistency with perceptual image evaluations, such as the feature similarity (FSIM) index [6] and the information content weighted SSIM (IW-SSIM) index [7], the highly computational complexity prevents them from real-time video applications. A video SSIM (VSSIM) [8] metric has been proposed to measure the quality of the distorted video in three levels, namely the local region level, the frame level, and the sequence level. However, the motion information along temporal trajectory and visual attention of HVS has not been made full used. More recently, a motion-based video integrity evaluation (MOVIE) index [9] has been proposed. MOVIE is shown to match human visual perception of video quality quite closely, but it is complicated to meet real-time video assessment.

Visual attention (VA) [10] is one of the most essential visual phenomena of HVS, which shows that the salient regions in visual field are highly focused by human eyes. In [11], a saliency detection method is incorporated with several video quality metrics and could improve the evaluating accuracy. While, the saliency based weighting strategy is only executed in frame level. In practice, the sequence level pooling stage is often done in simplistic or ad-hoc ways. It lacks theoretical principles as the basis for the development of reliable computational models. On the other hand, the saliency map in [11] is only obtained from the referenced videos. The differences between original and reconstructed ones are ignored. An intuitive idea shows that more emphasis should be put at high distorted regions, which can be done by using non-uniform weighting approach [12].

In order to deal with the aforementioned issues, in this paper we propose a VA based video quality assessment (VQA) approach, as shown in Fig. 1. The proposed VA map combines both visual saliency and distortion attention information. And it is finally employed as weight through the frame and sequence level quality pooling procedure. Correspondingly, we call the metrics proposed in these two levels: saliency and distortion weighted SSIM (SDW-SSIM), and SDW-SSIM with temporal pooling strategy (SDTW-SSIM). In our method, block-based motion estimation (BME) is applied to provide the motion information for temporal analysis. By considering visual factors, the proposed metric performs more consistency with human visual perception, meanwhile, it is efficient enough to meet real-time applications.

The rest of the paper is organized as follows. Firstly, the proposed visual saliency detection methods in frame level and sequence level are depicted in Section 2. Then, in Section 3, pooling strategies considering distortion weighting are developed to obtain the VQA index. In Section 4, experimental results are presented and analyzed. Finally, the work is summarized in Section 5.

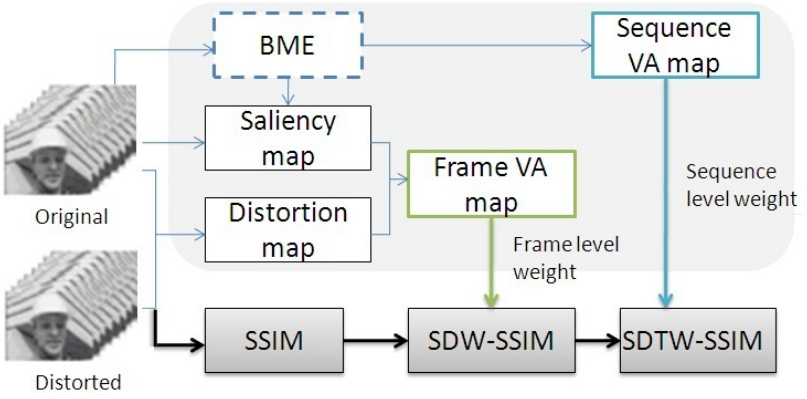

Fig. 1. Framework for the proposed VQA system

2 Spatial and Temporal Visual Saliency Analysis

2.1 Frame Level Saliency Map

Among existing saliency detecting methods, a Fourier transform based approach is proposed, which uses phase spectrum extracted from spectral domain to construct the corresponding salient areas in spatial domain [13]. Standing out from other models in computer vision, it is independent of prior knowledge and parameters and performs fast enough to meet real-time requirements. Furthermore, it can be extended from a two-dimensional Fourier transform to a quaternion Fourier transform (QFT) [14].

Firstly, each frame in the video sequence is represented as a quaternion image with four features. Different features affect the final saliency detection result [11, 13]. Here, we choose luminance, chrominance and motion information to construct the quaternion, considering that this kind of composition principle is similar to human visual perception process. Chrominance is represented by the H channel in HSV color space, and BME is applied to provide the motion information. Define the tth frame in a video sequence as $F(t)$, t=1, 2, …, N, where N is the total frame number of the video. $I(t)$, $H(t)$ are the luminance and chrominance components of frame $F(t)$.

After performing BME approach on luminance channel for each frame $F(t)$, we get the horizontal and vertical motion vector $V_x(t)$, $V_y(t)$. Then, the new quaternion image can be conducted as

$$q(t) = I(t) + H(t)\mu_1 + V_x(t)\mu_2 + V_y(t)\mu_3 \tag{1}$$

where

$$\mu_i^2 = -1, \quad i = 1,2,3$$

$$\mu_1 \perp \mu_2, \quad \mu_2 \perp \mu_3, \quad \mu_3 \perp \mu_1, \quad \mu_3 = \mu_1\mu_2$$

According to the theory demonstrated in [13], the saliency map $SM(t)$ can be obtained through the following ways. Firstly, QFT is applied to get the frequency domain representation $Q(t)$ of the quaternion image $q(t)$. Then, in order to construct the saliency map, the inverse QFT is employed to the phase spectrum $p(t)$ extracted from $Q(t)$, i.e.

$$Q(t) = QFT(q(t)) \tag{2}$$

$$p(t) = P(Q(t)) \tag{3}$$

$$SM(t) = g(t) * \| QFT^{-1}(e^{\mu \cdot p(t)}) \|^2 \tag{4}$$

where $g(t)$ is a Gaussian filter function, μ is a unit pure quaternion, and implementation details of QFT are available in [14].

This improved QFT method presents integrated saliency of intensity, color, and motion features. We choose to execute this method under the resolution of 64×64, and achieve a fast and visual friendly result.

2.2 Sequence Level Saliency Analysis

Most existing video saliency detection methods developed from image processing are lacking in temporal visual attention analysis. Considering that the changes of objects' moving speed are more likely to cause human concern, here we attempt to model this phenomenon as saliency along temporal trajectory.

Since time dimension features can be associated with moving information of objects in the scene, we use motion vectors as the modeling basis. We merge the previously obtained horizontal and vertical motion vectors into one component $V(t)$, we obtain

$$V(t) = \sqrt{V_x(t)^2 + V_y(t)^2} \tag{5}$$

With $V(t)$, the extent of changes between moving velocity in current frame and previous adjacent ones is calculated, and then the temporal salient degree $SV(t)$ is summed up as follows

$$SV(t) = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} | V_{ij}(t) - \frac{1}{3} \sum_{k=t-3}^{t-1} V_{ij}(k) | \tag{6}$$

where i and j denote the location of pixels, H and W are the height and width of the frame, respectively.

3 Pooling Strategy For VQA Index

3.1 Frame Level Pooling Strategy: SDW-SSIM

In section 2, saliency is obtained from the referenced videos only. Considering that HVS tends to pay more attention to the low quality region [12], here we employ distortion as another factor that affects VA. In this section, we incorporate the visual saliency map and distortion weight with SSIM metric in frame level.

For the distortion weight, here we use square error measure,

$$DM(t) = [I_r(t) - I_d(t)]^2 \tag{7}$$

where $I_r(t)$ and $I_d(t)$ denote the referenced and distorted frame pixels in the luminance channel, respectively.

The frame quality index (FQI) at the *t*th frame is merged as follows,

$$FQI(t) = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}\left[SM_{ij}(t)\cdot DM_{ij}(t)\cdot SSIM_{ij}(t)\right]}{\sum_{i=1}^{H}\sum_{j=1}^{W}\left[SM_{ij}(t)\cdot DM_{ij}(t)\right]} \tag{8}$$

Saliency map and distortion map are the two factors that are considered here as visual attention information. While in the saliency detection stage, luminance, color and motion information are included, distortion map stresses the corresponding distorted areas. All together, they are used to simulate the HVS observation mechanism. One advantage of the proposed frame level strategy is that it is sample and efficiency, making it appropriate for the RDO process of real-time video coding.

In order to give an intuitive perception of the frame level weighting scheme, here we present some saliency and distortion map detected from the referenced and distorted frames. As being seen from Fig. 2(c), the brighter areas show more significance for visual observation. Since the saliency detection procedure combines luminance, color and motion information, areas apt to catch VA in the referenced video are successfully detected. In order to make up the deficiency of saliency maps which is obtained from the referenced video only, the distortion maps provide information of the distorted video that is different from the referenced one. The corresponding distortion maps are shown in Fig. 2(d), in which, deeper distortion areas are brighter. Thus, the combination of these two points of view gives the VA a more comprehensive description in Fig. 2(e), and the following test demonstrates that the VA scheme proposed can significantly improve the performance of SSIM index.

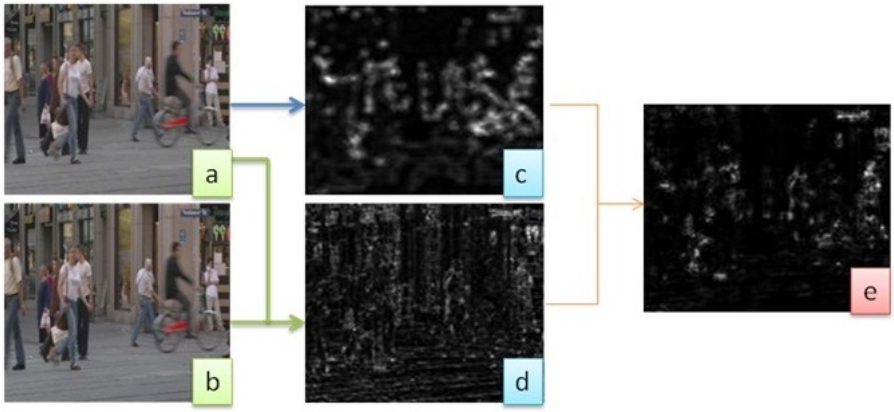

Fig. 2. (a) frame selected from the referenced video; (b) distorted frame; (c) saliency map; (d) distortion map; (e) VA map

3.2 Sequence Level Pooling Strategy: SDTW-SSIM

After each frame quality index have been generated with the saliency and distortion weighted method, the following stage is to pool the quality index sequences along time-domain to obtain the final VQA index. By taking into account temporal visual saliency in sequence level pooling, we obtain the proposed VQA metric SDTW-SSIM differently from the general mean value method.

The temporal salient degree $SV(t)$ is described in (6). The VQA index of the overall video is finally generated as

$$VQI = \sum_{i=1}^{N} \left[SV(t) \cdot FQI(t) \right] / \sum_{i=1}^{N} SV(t) \tag{9}$$

where N is the total frame number of the video. The introducing of this sequence level pooling strategy makes the objective VQAs more consistency with the characteristics of human visual process, thus, the performance of the VQAs can be improved.

4 Experiments

In this section, several VQAs together with the proposed SDW-SSIM method are tested on the LIVE VQA database [1]. First, three evaluation metrics are employed to compare the obtained objective assessment index with the subjective score provided within the database. Second, we present the time cost to make a comparison of the complexity between different metrics. Third, the scatter plot of the proposed method is compared with three other ones.

The LIVE VQA database [1] consists of 10 referenced videos of natural scenes and 150 distorted videos with the associated differential mean opinion score (DMOS). Each referenced video corresponds to 15 distorted ones with a wide range of distortions, created by using four common distortion types.

4.1 Performance Comparisons

The performance of the proposed method SDW-SSIM and SDTW-SSIM are compared with other VQAs. For PSNR, SSIM [4], VIF [5], weighted signal-to-noise ratio (WSNR) [15], FSIM [6], IW-SSIM [7], VS-SSIM [11] and SDW-SSIM, we first generate the quality index of each frame, and then make an average to obtain the final VQA index. The performance statistics of MOVIE in Table 1 is obtained from [16], which is also executed on the LIVE VQA database [1] under the same experimental conditions.

To evaluate the performance of the objective quality assessment models, we employ the metrics recommended in the VQEG Phase I FR-TV test [17]. Spearman rank-order correlation coefficient (SROCC) between the objective and subjective scores is presented as a measure of prediction monotonicity. After non-linear regression analysis, the Pearson linear correlation coefficient (LCC) and the root-mean-square error (RMSE) are calculated to measure the prediction accuracy. Higher LCC, SROCC and lower RMSE values indicate better evaluating performance.

Table 1. Performance comparisons

Algorithm	LCC	SROCC	RMSE
PSNR	0.5331	0.5032	9.2871
SSIM	0.5373	0.5184	9.2582
VIF	0.5869	0.5566	8.8878
WSNR	0.6783	0.6439	8.0658
FSIM	0.7238	0.7061	7.5739
IW-SSIM	0.7487	0.7411	7.2767
VS-SSIM	0.6101	0.5885	8.6976
MOVIE	0.8116	0.7890	-
SDW-SSIM	**0.7773**	**0.7630**	**6.9063**
SDTW-SSIM	**0.7868**	**0.7711**	**6.7756**

Table 1 shows, by employing visual saliency and distortion attention weighting with SSIM on the frame level pooling procedure only, the proposed SDW-SSIM metric shows reasonably good results compared with the other VQAs listed. Furthermore, SDTW-SSIM combines both frame and sequence level VA weight, and consequently provides even better performance. It still performs a slightly worse comparing with the MOVIE index. The time-consuming optical flow and complex HVS model have been used in the MOVIE index, as a result, making it too complicated to meet real-time requirement. While the statistical data of time cost in the following test shows that, the complexity of the proposed metric is quite acceptable.

4.2 Complexity Comparisons

In order to compare the complexity between different VQAs in practical applications, here, we also evaluate the running speed of some VQAs selected in Table 2. All these tests were run at Matlab R2010a software on Windows XP platform. Experimental environment is a DELL PC, with 2.00 GB memory. Because the released software of MOVIE index is a C++ implementation, the Matlab version is not available, so the MOVIE index is not listed in our test of this part. In Table 2, we record the time cost for different VQAs while evaluating the visual quality of the randomly selected video sequence pa3_25fps.yuv from LIVE VQA database [1].

Table 2. Time cost comparisons

Algorithm	Time(seconds)
PSNR	5.3
SSIM	45.5
VIF	861.8
WSNR	73.4
FSIM	477.3
IW-SSIM	408.9
VS-SSIM	180.9
SDTW-SSIM	184.2
SDTW-SSIM-noBME	**59.6**

Different from image applications, video processing is more sensitive to the complexity of the algorithm being used. The proposed SDTW-SSIM index also shows superiority in this respect as can be noticed from Table 2. Note that IW-SSIM and FSIM supply more flexibility than many VQAs listed in Table 1. However, the two schemes are more time-consuming. In addition, considering that BME module has been integrated into quite a few video processing systems already, for instance, the advanced video coding standard H.264, as a result, the actual time cost of SDTW-SSIM can be further reduced accordingly. The SDTW-SSIM-noBME in Table 2 represents the rest cost after removing the time used for the BME process. It shows that there is only a slight increase comparing with SSIM, while the evaluation performance of SSIM is much lower as can be seen from Table 1.

4.3 Scatter Plot Comparisons

For each scatter plot in Fig. 3, vertical and horizontal axes are for subjective and objective measurements, respectively. Each sample point represents one test video sequence. As being observed in Fig. 3, the sample points of the proposed SDTW-SSIM metric distribute more closely to the fitted line, which means it outperforms other VQAs. The same conclusion has been drawn from the above mentioned Table 1, the proposed metric has higher prediction monotonicity and accuracy.

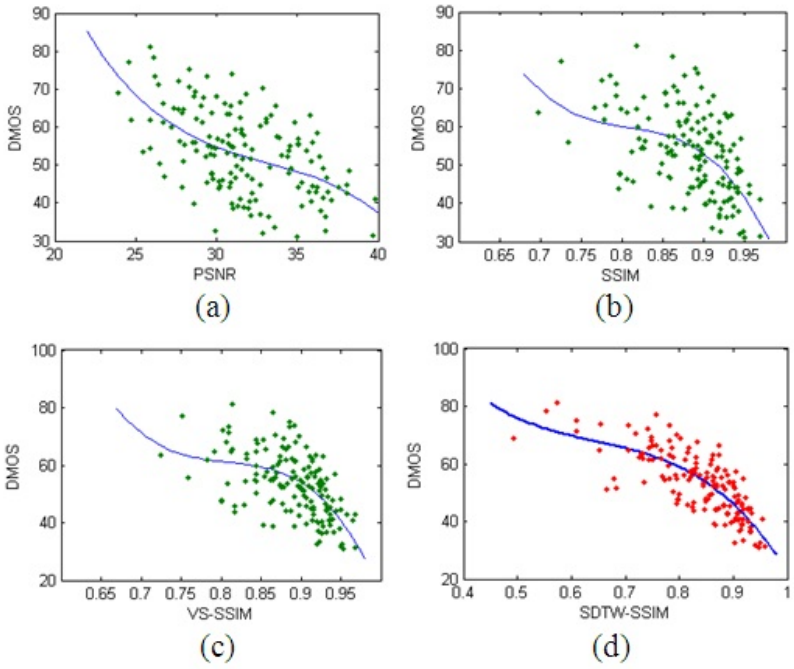

Fig. 3. Scatter plot comparison of different VQAs on LIVE VAQ database. (a) PSNR; (b) SSIM; (c) VS-SSIM; (d) SDTW-SSIM (proposed method)

Thus, from both the quantitative evaluation results and the scatter plots, the proposed SDTW-SSIM metric achieves rather good evaluation result. At the same time, the executing procedure is clear and with few parameters dependence and shows a meaningful balance between performance and efficiency. Since many video processing applications utilize a motion compensation scheme, the motion vectors can be reused while applying our saliency detection method, which reduces the consumption of time. In order to meet the requirement of real-time processing, here we only corporate the proposed VA weight with SSIM as a test. Moreover, the proposed model can be easily ported to other VQAs applications, such as FSIM, VIF and IW-SSIM.

5 Conclusions

In this paper, a visual saliency and distortion attention based perceptual VQA metric is proposed. The total algorithm is implemented in three stages. Firstly, each frame is evaluated by an SSIM index map. Secondly, spatial and temporal features of perceptual significance are modeled as visual saliency. Thirdly, in the pooling process, the visual saliency and distortion weight are combined with SSIM index to generate the final quality score. In order to obtain more consistent perception of human visual evaluation, the visual saliency detection and quality pooling stages are executed in both frame level and sequence level. Experimental results show that the proposed method is computationally efficient and performs well on the current widely used VQA database.

Acknowledgement. This work was supported in part by National Basic Research Program of China (973 Program): 2009CB320906, in part by National Natural Science Foundation of China: 61025011, 60833006, 61001177 and 61001108.

References

1. Seshadrinathan, K., Soundararajan, R., et al.: Study of Subjective and Objective Quality Assessment of Video. IEEE Trans. on Image Processing 19, 1427–1441 (2010)
2. Moorthy, A.K., Seshadrinathan, K., et al.: Wireless Video Quality Assessment A Study of Subjective Scores and Objective Algorithms. IEEE Trans. on Circuits Systems for Video Technology 20, 587–599 (2010)
3. Lin, W., Jay Kuo, C.-C.: Perceptual Visual Quality Metrics: A Survey. J. of Visual Communication and Image Representation 22, 297–312 (2011)
4. Wang, Z., Bovik, A.C., et al.: Image Quality Assessment: From Error Measurement to Structural Similarity. IEEE Trans. on Image Processing 13, 600–612 (2004)
5. Sheikh, H.R., Bovik, A.C.: Image Information and Visual Quality. IEEE Trans. on Image Processing 15, 430–444 (2006)
6. Zhang, L., Zhang, L., et al.: FSIM: A Feature Similarity Index for Image Quality Assessment. IEEE Trans. on Image Processing 20, 2378–2386 (2011)
7. Wang, Z., Li, Q.: Information Content Weighting for Perceptual Image Quality Assessment. IEEE Trans. on Image Processing 20, 1185–1198 (2011)
8. Wang, Z., Lu, L., et al.: Video Quality Assessment Based on Structural Distortion Measurement. Signal Processing: Image Communication 19, 121–132 (2004)
9. Seshadrinathan, K., Bovik, A.C.: Motion Tuned Spatio-temporal Quality Assessment of Natural Videos. IEEE Trans. on Image Processing 19, 335–350 (2010)
10. Engelke, U., Kaprykowsky, H., et al.: Visual Attention in Quality Assessment. IEEE Signal Processing Magazine 28, 50–59 (2011)
11. Ma, L., Li, S., et al.: Motion Trajectory Based Visual Saliency for Video Quality Assessment. In: IEEE International Conference on Image Processing, pp. 233–236 (2011)
12. Wang, Z., Shang, X.: Spatial Pooling Strategies for Perceptual Image Quality Assessment. In: IEEE International Conference on Image Processing, pp. 2945–2948 (2006)
13. Guo, C., Ma, Q., et al.: Spatio-temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
14. Ell, T.A., Sangwine, S.J.: Hypercomplex Fourier Transforms for Color Images. IEEE Trans. on Image Processing 16, 22–35 (2007)
15. Gaubatz, M., Hemami, S.S.: MeTriX MuX Visual Quality Assessment Package, `http://foulard.ece.cornell.edu/gaubatz/metrix_mux`
16. Moorthy, A.K., Bovik, A.C.: Efficient Video Quality Assessment Along Temporal Trajectories. IEEE Trans. on Circuits and Systems for Video Technology 20, 1653–1658 (2010)
17. VQEG: Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment (2000), `http://www.vqeg.org/`