Efficient Video Quality Assessment Based on Spacetime Texture Representation

Peng Peng
Simon Fraser University
8888 University Drive
Burnaby, BC, Canada
pengp@sfu.ca

Kevin Cannons
Simon Fraser University
8888 University Drive
Burnaby, BC, Canada
kcannons@sfu.ca

Ze-Nian Li
Simon Fraser University
8888 University Drive
Burnaby, BC, Canada
li@cs.sfu.ca

ABSTRACT

Most existing video quality metrics measure temporal distortions based on optical-flow estimation, which typically has limited descriptive power of visual dynamics and low efficiency. This paper presents a unified and efficient framework to measure temporal distortions based on a spacetime texture representation of motion. We first propose an effective motion-tuning scheme to capture temporal distortions along motion trajectories by exploiting the distributive characteristic of the spacetime texture. Then we reuse the motion descriptors to build a self-information based spatiotemporal saliency model to guide the spatial pooling. At last, a comprehensive quality metric is developed by combining the temporal distortion measure with spatial distortion measure. Our method demonstrates high efficiency and excellent correlation with the human perception of video quality.

Categories and Subject Descriptors

I.4.9 [**Computing Methodologies**]: IMAGE PROCESSING AND COMPUTER VISION Applications

General Terms

Algorithms, Experimentation, Measurement

Keywords

Video quality assessment, spatiotemporal oriented energy (SOE), spacetime texture representation, visual attention

1. INTRODUCTION

Digital videos typically pass through several processing stages (e.g., lossy source encoding and transmission over error prone channels) that may result in impairment of quality before they reach the end users. Methods for evaluating video quality have received growing interest from content providers and network operators, as they play a crucial role in Quality-of-Service (QoS) monitoring, and perceptually optimal design and/or performance evaluation of video processing systems. Objective video quality assessment (VQA) methods are based on automated algorithms that attempt to predict video quality in a way consistent with the human perception. As a result, their performance is evaluated by how well the predicted quality correlates with the human-supplied subjective quality. Depending on the presence of a reference video, the VQA methods can be divided into three classes: full-reference, reduced-reference, and no-reference. In this study, we focus on the full-reference VQA algorithms.

Many factors affect the quality of digital videos. Distortions in video that arise primarily from the occurrence of motion are referred to as "temporal distortions" (e.g., ghosting, jerkiness, and mosquito effect), as opposed to the "spatial distortions" (e.g., blocking, ringing, and noise) [8]. Along with the progress of image quality assessment (IQA), there has been extensive research for measuring spatial distortions in videos. On the contrary, how to measure temporal distortions is much less studied.

In the context of video coding and quality assessment, motion has often been related to the notion of "optical flow", a motion vector at each spatiotemporal point with its length representing the magnitude of motion and its direction the direction of motion. It is no surprise that almost all existing VQA methods that explicitly incorporate motion estimation are based on the optical-flow representation. In early studies [12, 13], motion vectors are employed to estimate the weights for pooling the local quality measures into a single quality score. More recently, Moorthy et al. [6] proposed a motion compensated SSIM index for video quality assessment, in which the SSIM index [11] is performed on motion-compensated image blocks. Lately, Seshadrinathan and Bovik [8] propose the MOVIE index that captures the temporal distortions along motion trajectories by using Gabor filters tuned to the motion directions based on an optical flow method. It achieves the best quality prediction performance among a set of state-of-the-art VQA methods [9].

The optical-flow methods that attempt to precisely estimate motion vectors usually suffer from the high computational overhead caused by iterative optimization. A more critical issue is that the conventional optical-flow representation is inadequate in capturing general visual dynamics, such as pure temporal variations (e.g., campfire), and semi-transparency scene with more than one motion present at a single point (e.g., rising smoke) [3]. In this paper we introduce a recently proposed spacetime texture representation of motion [3] to the field of VQA. Besides its high efficiency,

MM'13, October 21–25, 2013, Barcelona, Spain.

ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502168.

it provides a distributed representation of motion and thus better descriptive power of general visual dynamics.

We first propose a local motion-tuned temporal distortion measure by exploiting the distributive characteristic of the spacetime-texture descriptors. The descriptors are then reused to build a spatiotemporal attention model for spatial pooling. In the end, the temporal distortion measure is combined with a simple spatial distortion measure to give a comprehensive evaluation of video quality. Our method achieves remarkable performance in both quality prediction and computational efficiency.

2. PROPOSED METHOD

2.1 Motion-tuned temporal distortion measure

To compute the spacetime texture representation [3], a video sequence is first filtered spatiotemporally using a bank of broadly tuned Gaussian third derivative filters, $G_{3_{\hat{\theta}}}$, where $\hat{\theta}$ is a unit vector that captures the spatiotemporal direction of the filter symmetry axis. Each filter responds best to a stimulus moving in a specific direction in the spatiotemporal space. As in [3], the filter responses are point-wise rectified and summed over a spatiotemporal neighborhood (a spatiotemporal region Π) to yield a measurement of signal energy for this region at each orientation $\hat{\theta}$:

$$E_{\hat{\theta}}(x,y,t) = \sum_{(x,y,t)\in\Pi} (G_{3_{\hat{\theta}}} * V)^2 \tag{1}$$

where $V = V(x,y,t)$ denotes the input spatiotemporal volume (i.e., the input video sequence), and the symbol "*" denotes convolution. The bandpass nature of the G_3 filters leads to the invariance of the energies to additive intensity variations. However, the local energy estimates still increase monotonically with contrast. In order to capture the temporal distortions irrespective of both the additive intensity variations and contrast change, a pixel-wise divisive normalization is performed. Specifically, the local energy measures are normalized by the summation of energy responses from all filters considered at each location.

Let $\hat{E}^r_{\hat{\theta}_k}(x,y,t)$ and $\hat{E}^d_{\hat{\theta}_k}(x,y,t)$ denote the normalized local energy measure along direction $\hat{\theta}_k$ in the reference video and distorted video, respectively. Now, we can obtain a local temporal distortion measure at each location (x,y,t) by calculating the similarity between the two corresponding energy distributions in the reference and distorted videos. In this paper, we use the efficient $\mathcal{L}_2$ distance:

$$TD(x,y,t) = [\sum_{k=1}^{K} (\hat{E}^r_{\hat{\theta}_k}(x,y,t) - \hat{E}^d_{\hat{\theta}_k}(x,y,t))^2]^{\frac{1}{2}}. \tag{2}$$

In this measure, each filter in the selected filter bank plays an equally important part. However, it has been shown that assigning biased weights to the filters according to the local motion pattern can better capture the temporal distortions [8]. Inspired by this, we propose a local motion-tuned temporal distortion measure. Following the method in [3], a distributed motion representation can be efficiently computed by "appearance marginalization" of the oriented energies in Eq. (1). The goal of this marginalization is to capture the purely dynamic properties of a scene, i.e., the motion-related properties independent from the spatial appearance. As a pattern with a specific velocity manifests itself as a plane through the origin in the frequency domain, the purely spatial orientation component in Eq. (1) can be discounted by summation across a set of spatiotemporal oriented energy measurements consistent with the corresponding frequency plane. Let a frequency plane be parameterized by its unit normal $\hat{n}_k$, and N the order of the Gaussian filters (here, $N = 3$). On each plane, $N+1$ equally spaced directions $\{\hat{\theta}_j, j = 1,\ldots,N+1\}$ are sampled for summation,

$$\tilde{E}_{\hat{n}_k} = \sum_{j=1}^{N+1} E_{\hat{\theta}_{k,j}}, k = 1,\ldots,K \tag{3}$$

with each $E_{\hat{\theta}_{k,j}}$ being the spatiotemporal energy given in Eq. (1). In this study, we selected 13 planes (i.e., $K = 13$) corresponding to the following motion directions: static (no motion), motion in eight directions (leftward, rightward, upward, downward and the four diagonals), and flicker in four directions (horizontal, vertical and two diagonals). We find that tuning the directions more finely did not lead to noticeably better performance but incurs greater computation. To attain insensitivity to contrast change, each $\tilde{E}_{\hat{n}_k}$ is normalized by their summation over all K directions. This results in a K-bin histogram $\{\hat{E}_{\hat{n}_k}, k = 1,\ldots,K\}$, which encapsulates a relative indication of the motion strength corresponding to each plane. Let $\hat{E}^r_{\hat{n}_k}(x,y,t)$ denote the histogram corresponding to a plane $\hat{n}_k$ at a spatiotemporal location (x,y,t) in the reference video. To obtain a motion-tuned temporal distortion measure, we integrate $\hat{E}^r_{\hat{n}_k}(x,y,t)$ into the temporal distortion measure in Eq. (2):

$$\begin{aligned} MT\text{-}TD(x,y,t) = [\sum_{k=1}^{K} (\hat{E}^r_{\hat{n}_k}(x,y,t) \times \\ \sum_{j=1}^{N+1} (\hat{E}^r_{\hat{\theta}_{k,j}}(x,y,t) - \hat{E}^d_{\hat{\theta}_{k,j}}(x,y,t))^2)]^{\frac{1}{2}}, \end{aligned} \tag{4}$$

where the $N+1$ filters $\{\hat{\theta}_{k,j}, j = 1,\ldots,N+1\}$ consistent with a certain frequency plane $\hat{n}_k$ are weighted by $\hat{E}^r_{\hat{n}_k}(x,y,t)$.

2.2 Attention-guided spatial pooling

In recent years, visual attention has received increasing interest in the area of visual quality assessment. While most attention-guided VQA methods employ attention models purely based on spatial cues (e.g., color, intensity and spatial orientation), some recent VQA methods take into consideration the motion-driven attention, in which the motion estimates are usually based on the optical-flow representation (e.g. [15]) or simply described by the adjacent frame difference (e.g., [5]). In this study, we reuse the local motion descriptors computed in Section 2.1 for visual attention modeling. Following the Attention by Information Maximization (AIM) principle proposed by Bruce and Tsotsos [2], a spatiotemporal saliency model is built based on the self-information of the local features, $\hat{E}_{\hat{n}_k}$ and $\tilde{E}_{\hat{n}_k}$. At each spatiotemporal location, we compute two self-information measures:

$$SI_M(x,y,t) = \sum_{k=1}^{K} -log(p(\hat{E}_{\hat{n}_k}(x,y,t))), \tag{5}$$

$$SI_{MC}(x,y,t) = \sum_{k=1}^{K} -log(p(\tilde{E}_{\hat{n}_k}(x,y,t))), \tag{6}$$

where $p(\hat{E}_{\hat{n}_k}(x,y,t))$ and $p(\tilde{E}_{\hat{n}_k}(x,y,t))$ are the probabilities of seeing certain values of the motion descriptors given their surround, which are estimated by histogram density estimation over all the pixels in the current frame. Note that, SI_M is based on a motion descriptor that is invariant to local luminance contrast. Hence it highlights the regions with motion patterns that are very different from their surround. On the other hand, SI_{MC} is based on a motion descriptor that is confounded by luminance contrast. As a result, it also highlights the regions of high luminance contrast. The overall saliency is a combination of SI_M and SI_{MC}. How to optimally combine saliency maps driven by multiple cues remains an open problem. A widely accepted principle is that salient motion plays a much more significant role than static features (including luminance contrast) in attracting visual attention [4]. Following this principle, we compute the combined saliency as

$$SI_{COM}(x,y,t) = \gamma \cdot SI_M(x,y,t) + (1-\gamma) \cdot SI_{MC}(x,y,t) \cdot SI_M(x,y,t), \quad (7)$$

where SI_M and SI_{MC} are normalized to the range of $[0,1]$, and γ is a free parameter in the range of $[0,1]$. In our implementation, we set $\gamma = 0.5$ empirically. With this combination scheme, if motion saliency exists, SI_M will dominate the overall saliency. Otherwise, if SI_M is largely smooth, SI_{MC} (in this case, primarily driven by luminance contrast) will play a significant role. In our implementation, we also take into account the center bias by combining the saliency map SI_{COM} with a center-bias map CB:

$$A(x,y,t) = SI_{COM}(x,y,t) \cdot CB(x,y), \quad (8)$$

where $CB(x,y) = 1 - d/D$ is a decreasing function of the distance d between the image center and the spatial position (x,y). D is the distance between the center and a corner. Based on this saliency model, an attention-guided motion-tuned temporal distortion measure at each frame t is computed as

$$AG\text{-}MT\text{-}TD(t) = \frac{\sum_{x,y} MT\text{-}TD(x,y,t) \cdot A(x,y,t)}{\sum_{x,y} A(x,y,t)}. \quad (9)$$

With this spatial pooling scheme, the temporal distortions in the highly attentional regions are heavily penalized.

2.3 Overall video quality prediction

It is clear that the human perception of video quality is affected by both temporal and spatial distortions. Therefore, we combine the proposed temporal distortion measure with a spatial distortion measure to give a comprehensive judgment of video quality:

$$D_{overall} = D_{AG\text{-}MT\text{-}TD} \cdot (1 - Q_{MS\text{-}SSIM}), \quad (10)$$

where $D_{AG\text{-}MT\text{-}TD}$ is a temporal distortion measure by temporally pooling the frame-level $AG\text{-}MT\text{-}TD(t)$ scores (see Eq. (9)), and $Q_{MS\text{-}SSIM}$ is a spatial quality measure by temporally pooling the frame-level MS-SSIM [14] scores. Note that a higher MS-SSIM score (always in the range of $[0,1]$) indicates high quality. The MS-SSIM index is selected because it has shown good effectiveness in measuring a variety of spatial distortions as well as high computational efficiency. For temporal pooling, we employ a method similar to the one used in [7] to penalize high temporal variations, which adjusts the mean value of the frame-level quality indices with an additive term that increases with the temporal variations of quality.

3. EXPERIMENTAL RESULTS

There are two commonly used public databases for VQA, namely, the VQEG FRTV Phase 1 database [1] and the LIVE video quality database [10]. The former was published in 2000, and the distortions in its test videos (e.g., MPEG-2 and H.263 compression) are considered to be outdated. The later was published in 2010, which was designed to be a replacement of the VQEG FRTV Phase 1 database. It contains videos compressed by H.264 and MPEG-2, as well as videos obtained by simulated transmission of H.264 compressed streams through error prone IP and wireless networks [10]. Consequently, most of the recent work uses the LIVE database for performance evaluation. We also follow this trend in this paper.

We employ two commonly used metrics, the Spearman's Rank Correlation Coefficient (SRCC) and the Pearson Linear Correlation Coefficient (PLCC), to measure the correlation between the algorithm-supplied and human-supplied quality scores – they measure the prediction monotonicity and prediction accuracy, respectively. The intermediate results are presented in Table 1, where the suffix "(tv)" indicates that the temporal variation of quality is taken into account in temporal pooling. Columns without the suffix use the mean of frame-level quality scores. By comparing the performance of the "TD", "MT-TD" and "AG-MT-TD" methods, we can see that both the motion-tuning scheme and the attention model play very beneficial roles in the proposed method. In addition, the "Proposed (tv)" method achieves considerable performance gain over "AG-MT-TD (tv)" and "MS-SSIM (tv)", which indicates that the two components compliment each other.

We have compared our method with a set of VQA algorithms studied in a recent survey paper [9] and two recent visual attention-guided VQA methods ("VA-You" [15] and "VA-Ma" [5]) in Table 2. In comparison, our method demonstrates remarkable quality-prediction performance. The optical flow-based MOVIE index [8] also achieves impressive results. It is worth mentioning that the MOVIE index does not incorporate a visual attention model and hence its performance may also be improved by exploring visual attention adequately.

We have also analyzed the computation time of our method. It shows that our method implemented in MATLAB took 178 seconds (26 seconds for the $D_{AG\text{-}MT\text{-}TD}$ component and 152 seconds for the $Q_{MS\text{-}SSIM}$ component) to evaluate a 10-second 25 fps 432×768 test video on a Linux machine with an Intel Core i2 CPU (2.33 GHz) and 8 GB memory. In our implementation, the $Q_{MS\text{-}SSIM}$ component was performed at five scales starting from the original scale to four gradually coarser scales, whereas the $D_{AG\text{-}MT\text{-}SD}$ component was performed at a coarse scale (144×256) for higher efficiency (if it was performed at the original scale, it would take about 254 seconds to run and give close prediction performance). Clearly, our method is highly efficient even when compared with the simple frame-by-frame MS-SSIM method. Under the same test condition, the MOVIE index took about 5.88 hours using the C++ code [8] provided by the authors, where a considerable portion of the running time (more than 55%) was purely devoted to the computation of optical flow.

Table 1: Intermediate results: SRCC and PLCC on the LIVE Video Quaity database

Metrics	TD	MT-TD	AG-MT-TD	AG-MT-TD (tv)	MS-SSIM	MS-SSIM (tv)	Proposed (tv)
SRCC	0.7011	0.7432	0.7736	0.7914	0.7367	0.7536	0.8215
PLCC	0.7116	0.7478	0.7777	0.7933	0.7441	0.7647	0.8274

Table 2: Comparative results: SRCC and PLCC on the LIVE Video Quality database.

Metrics	PSNR	SSIM	MS-SSIM	VSNR	SpeedSSIM	VQM	V-VIF	MOVIE	VA-You	VA-Ma	Proposed
SRCC	0.3684	0.5257	0.7367	0.6755	0.5849	0.7026	0.5710	**0.7890**	–	0.7484	**0.8215**
PLCC	0.4035	0.5444	0.7441	0.6896	0.5962	0.7236	0.5756	**0.8116**	0.776	0.7768	**0.8274**

4. CONCLUSION

In this paper, we present a unified and efficient framework for motion-tuned and attention-guided VQA based on a spacetime texture representation of motion. When evaluated on the LIVE video quality database, our method achieves excellent correlation with human perception of video quality. With the exciting success in this initial attempt, we believe spacetime texture can serve as an excellent alternative to the optical-flow methods in the field of VQA, especially in real-time and/or mobile device-based applications that require high computational efficiency. Despite this, spacetime texture as formulated is not a replacement of optical flow for all application domains, e.g., in cases that motion vectors need to (and can) be precisely estimated and that computation overhead is not a major concern. In the future, we will employ the reliable and efficient spacetime texture representation to conduct further analysis of the location (spatial and temporal), intensity and pattern of visual dynamics, which we believe will facilitate the modeling of visual attention and the measurement of several challenging temporal distortions, including stalling, frame freezing, and frame skipping.

5. REFERENCES

[1] VQEG FRTV Phase 1 Database. http://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx.

[2] N. Bruce and J. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3), 2009.

[3] K. Derpanis and R. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1193–1205, 2012.

[4] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 631–637, 2005.

[5] L. Ma, S. Li, and K. N. Ngan. Motion trajectory based visual saliency for video quality assessment. In *IEEE International Conference on Image Processing (ICIP)*, pages 233–236. IEEE, 2011.

[6] A. Moorthy and A. Bovik. A motion compensated approach to video quality assessment. In *Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, pages 872–875. IEEE, 2009.

[7] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. Considering temporal variations of spatial visual distortions in video quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):253–265, 2009.

[8] K. Seshadrinathan and A. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, 19(2):335–350, 2010. Code: http://live.ece.utexas.edu/research/quality/movie.html

[9] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6):1427–1441, 2010.

[10] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack. A subjective study to evaluate video quality assessment algorithms. In *SPIE Proceedings Human Vision and Electronic Imaging*, volume 7527, 2010.

[11] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[12] Z. Wang, L. Lu, and A. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, 19(2):121–132, 2004.

[13] Z. Wang and Q. Li. Video quality assessment using a statistical model of human visual speed perception. *J. Opt. Soc. Amer. A*, 24(12):B61–B69, 2007.

[14] Z. Wang, E. Simoncelli, and A. Bovik. Multiscale structural similarity for image quality assessment. In *IEEE Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1398–1402, 2003.

[15] J. You, J. Korhonen, and A. Perkis. Attention modeling for video quality assessment: balancing global quality and local quality. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 914–919. IEEE, 2010.