Motion saliency for spatial pooling of objective video quality metrics

Marina Georgia Arvanitidou
Communication Systems Group
Technische Universität Berlin
Berlin Germany
arvanitidou@nue.tu-berlin.de

Thomas Sikora
Communication Systems Group
Technische Universität Berlin
Berlin Germany
sikora@nue.tu-berlin.de

ABSTRACT

In this paper we propose a novel motion saliency estimation method for video sequences considering the motion between successive frames and their corresponding parametric camera motion representation. Background motion is compensated for every pair of frames, revealing areas that contain relative motion. Considering that these areas will likely attract the attention of the viewer and in line with properties of the human visual system, regarding spatially invariant focus distribution, we augment their effect on the quality estimation. The generated saliency maps are thus incorporated in the spatial pooling stage of several video quality metrics, and experimental evaluation on the LIVE video database[1] shows that this strategy enhances their performance.

Keywords

video quality assessment, spatial pooling, motion saliency, global motion estimation

1. INTRODUCTION

The broad use of video in communication services, such as IPTV and video conferencing, resulted in the increasing interest of the image processing research community in the topic of objective Video Quality Assessment (VQA). As humans are the final judges of service quality, key issue is the development of algorithms that efficiently assess the quality experienced by users (QoE). The commonly acceptable way for assessing the video quality is to conduct a large scale subjective study where a group of observers are asked to provide their personal opinions of the video. This subjective Mean Opinion Score (MOS) can then be regarded as the ground-truth subjective quality of the video sequences. In practice subjective experiments are time-, effort- and resource- consuming, and therefore objective video quality metrics that can automatically evaluate the video's perceptual quality are appreciated.

[1]Publicly available at
http://live.ece.utexas.edu/research/quality/live_video.html

In this paper we focus on full reference objective quality assessment algorithms that employ both the reference and the distorted video sequences. Among the most widely used metrics in this category are the Mean Square Error (MSE) and the Peak Signal to Noise Ratio (PSNR), which are straight-forward in implementation but do not consider any properties of the Human Visual System (HVS). Towards this direction a large variety of metrics have been proposed in the past. The Structural SIMilarity index (SSIM) [14] captures the structure information loss between the reference and distorted image to estimate the perceptual quality. Initially it was proposed in single scale, and multiscale versions of it [16] followed. The Visual Information Fidelity criterion (VIF) [11] is another powerful metric that employs the mutual information between the original and the distorted image to estimate the perceived quality. The Video Quality Model (VQM) [8], which is adopted by the American national standards institute, analyzes 3D spatio-temporal blocks to extract the salient features for estimating the video quality map, whereas the MOtion- based Video Integrity Evaluation (MOVIE) metric [9] utilizes properties of the visual cortex neurons to track perceptually relevant distortions along motion trajectories. The latter is a computationally complex metric since it relies in 3D optical flow estimation.

Existing works on the exploitation of temporal distortions in video sequences report improvements on the prediction performance of standard quality metrics by taking into account importance maps and appropriate pooling techniques. Moorthy *et al.* [6] propose a computationally efficient VQA algorithm that assesses the quality in block-based level and subsequently employs percentile pooling. Towards VQA enhancement Ma *et al.* [5] propose a visual saliency estimation algorithm, based on the quaternion Fourier transform of the motion vectors after block matching. Recently, the authors in [13] have proposed a video quality metric that employs the structural information contained in two descriptors extracted from the 3D structure tensors, and its corresponding eigenvector, whereas in [15] a model of human visual speed perception is incorporated to model visual perception in an information communication framework.

In line with the idea that the performance of VQA metrics can be improved by considering features along the temporal trajectories, in this paper we propose to exploit the motion information between successive video frames for estimation of motion saliency maps and employ them for spatial pooling of video quality metrics. By giving higher weighting to

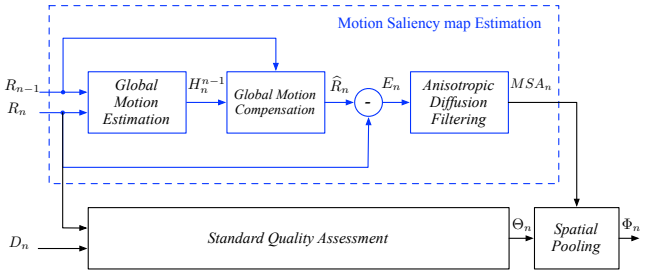

Figure 1: Proposed method overview

regions with detected motion, we expect these regions to attract humans' attention, and thus make distortion on these areas more perceivable. Experimental results on various video quality metrics evaluate the efficiency of our method.

The paper is organized as follows. Section 2 provides an overview of the proposed motion saliency estimation and describes the spatial pooling strategy that is employed for incorporation in objective video quality prediction algorithms. Section 3 presents and discusses the experimental evaluation of our algorithm and Ssection 4 summarizes and closes this paper.

2. PROPOSED MOTION SALIENCY MAPS AND VQA ENHANCEMENT

The main idea of our proposed method is to detect regions that contain significant relative motion between frames, and augment their effect to the image quality index in the spatial pooling stage that will follow. This means that if a distortion occurs in a region that contains motion, it is expected to attract the attention of the viewer and thus to have a negative impact on the quality assessment in comparison to a distortion that occurs in a region not containing motion.

2.1 Motion Saliency maps Estimation

Motion between successive frames is what differentiates a video sequence from a set of independent still images. Motion can be considered as a mixture of foreground object motion and camera motion. If we assume that the background (i.e. camera) motion is the dominant motion between two frames of a video sequence, then the foreground motion is likely to attract visual attention, according to the properties of the HVS that are explored with respect to this point of view in [15]. Based on this observation we propose the following strategy, which is illustrated in Figure 1.

First we estimate the perspective (eight-parameter) motion model that describes the background motion between two successive frames of the reference sequence $R_n - 1$ and R_n. This is realized as described in [3] where a set of feature-points are detected in each frame, the correspondences between the two sets of points are established for successive frames, and finally a motion model H_n^{n-1} is fitted to the correspondences. The RANdom SAmpled Consensus (RANSAC) method is employed for fast and accurate model fit based on features which are detected using the KLT feature tracker. Based on the estimated motion model H_n^{n-1} and the corresponding frame R_{n-1}, the estimated frame $\widehat{R}_n$ is computed and subsequently subtracted from R_n. This results in the (global motion) compensated absolute error frame E_n where high error energy corresponds mainly to motion of the foreground area.

Studies on the HVS properties have shown that the human retina is highly space variant in processing and sampling of visual information [2]. The accuracy is highest in the central point of focus, the fovea, and the peripheral visual field is perceived with lower accuracy. In our case, we consider the locations of the highest motion compensated error energy as the central points of focus, and to address the gradually decreasing focus, the error maps are low-pass filtered resulting in the motion saliency map MSA.

$$MSA(x,y,n) = \alpha * |\widehat{R}(x,y,n) - R(x,y,n)| \qquad (1)$$

Anisotropic diffusion filtering (α) [7] is employed here, as similarly employed for the scope of object segmentation in [1]. Anisotropic diffusion offers a non-linear and space-variant filtering of the error frame, that while having a low pass character preserves the edges of the image. In this way we give higher weighting to regions that have moved between two successive frames and we expect that they are more likely to attract visual attention in comparison to other areas that have not moved (or have moved with the background). As shown in Figure 2 our motion saliency estimation method can significantly detect the motion of the foreground (brighter areas) in the MSA map. Of course motion is not the only feature that attracts visual attention. Other features such as contrast, color and structural information will be considered implicitly through the incorporation in standard objective metrics that is following described.

2.2 Spatial Pooling

Standard image quality metrics tend to generate a quality index Θ between a reference and a distorted image (R and D respectively) and then consider that every pixel contributes equally to the overall image metric by averaging over all pixel locations. As we want to avoid this uniform spatial pooling, we employ the weighted mean spatial pooling strategy [2]. The estimated motion saliency maps are incorporated in MSE, SSIM [14], MS-SSIM [16] and VIF [11] metrics. The weighted mean for single scale metrics in (x, y) location of the n^{th} frame [15] is formulated as

$$\Phi(x,y,n) = \frac{\sum_{x=1}^{N_x}\sum_{y=1}^{N_y} w(x,y,n)\Theta(x,y,n)}{\sum_{x=1}^{N_x}\sum_{y=1}^{N_y} w(x,y,n)} \qquad (2)$$

where Φ is the weighted metric and N_x, N_y are the frame dimensions. The squared error map serves as quality index in the case of MSE, and the SSIM index map is employed for SSIM. The proposed motion saliency maps $MSA(x,y,n)$ are used as weighting maps w, whereas in the following section, a local saliency map is additionally considered for comparison. For multiscale metrics, that use M scales, the weighting map is scaled correspondingly and the overall metric is calculated as following

$$\Phi(x,y,n) = \prod_{j=1}^{M} \frac{\sum_{x=1}^{N_x}\sum_{y=1}^{N_y} w(x,y,n)\Theta(x,y,n)}{\sum_{x=1}^{N_x}\sum_{y=1}^{N_y} w(x,y,n)} \qquad (3)$$

MS-SSIM incorporates SSIM evaluations in different scales. To this end, SSIM index maps are, weighted according to Eq. 2 in each scale and then the various weighted scaled indexes

are combined as described in [16]. For VIF [11], the mutual information (between the input and the output of the HVS channel) for the reference image and the mutual information (between the input and the output of the HVS channel) for the distorted image are separately weighted using Eq. 2, scaled to the corresponding scales and finally combined over multiple scales.

After the local weighted quality score of every frame $\Phi(n)$ is generated, by considering the motion saliency map, temporal pooling follows. The local scores are averaged over the T frames of the video sequence to yield the overall weighted quality score Φ.

3. EXPERIMENTAL EVALUATION

We evaluate the performance of our proposed algorithm on the above mentioned metrics on the LIVE video database [10], developed at the University of Texas at Austin. The LIVE database contains 150 distorted videos obtained from 10 uncompressed reference videos (768×432 pixels, 3206 frames totally) of natural scenes. The distorted videos are created using four commonly encountered distortion types. These include MPEG-2 compression, H.264 compression, simulated transmission of H.264 compressed bitstreams through error-prone IP networks, and through error-prone wireless networks. Each video was assessed by 38 human subjects in a single stimulus study with hidden reference removal, where the subjects scored the video quality on a continuous quality scale. The difference scores of a given subject are computed by subtracting the score assigned by the subject to the distorted video sequence from the score assigned by the same subject to the corresponding reference video sequence. Following the Difference Mean Opinion Score (DMOS) of each video is computed as the mean of the rescaled standardized difference scores (Z-scores) of statistically reliable subjects.

The quality prediction performance of the weighted metrics is evaluated using three performance indicators. As the Video Quality Experts Group (VQEG) recommends [12], we use the Pearson Linear Correlation Coefficient (LCC) and the Spearman Rank Order Correlation Coefficient (SROCC) between DMOS and the objective score after nonlinear regression. Additionally, we employ the Root Mean Square Error (RMSE) in the same manner. Non linear regression is applied in order to align each video quality metric output to the subjective rating scale using the logistic function described in [12]. Figure 3 illustrates an example scatter plot of predicted DMOS using standard MS-SSIM and weighted MS-SSIM using the proposed method (green and blue marks respectively) versus DMOS.

Table 1 shows the performance evaluation of various objective VQA algorithms. VS denotes the Visual Saliency model proposed in [5]. MSA is our proposed method, whereas local saliency denotes the employment of local saliency maps proposed by Itti & Koch [4] for weighting in the same manner as described in the previous section. For each evaluation metric we highlight the best results with boldface. Larger LCC and SROCC indicate better correlation between objective and subjective scores, while smaller RMSE is indicator of better performance. Saliency weighted metrics perform better compared to non-weighted metrics. Motion saliency spatial pooling proves to be more beneficial for objective VQA compared to local saliency pooling. This shows that motion saliency is beneficial for VQA metrics providing a better agreement with the subjective ground-truth scores. However, saliency weighted methods still cannot outperform MOVIE [9], which employs a complex HVS based model for exploiting temporal and distortions. Our proposed method does not seek to explicitly model properties of the HVS, however it is competitive with MOVIE, especially for the case of MSA-weighted MS-SSIM, even though it avoids computationally complex optical flow estimation or multi-scale filtering over large temporal trajectories. To examine the effect of our proposed weighting on different distortion types, we show in Table 2 the performance enhancement, in terms of Spearman Rank Order Correlation Coefficient, for our proposed method for each distortion class separately. As expected, our proposed method contributes on average more in cases of transient distortions (in the presence of packet losses, classes 1&2) compared to cases with uniformly distributed distortions (no packet losses, classes 3 & 4).

Table 1: VQA metrics performance on LIVE database. Data for VS from [5]

Algorithm	LCC	SROCC	RMSE
MSE	0.5614	0.5391	9.0839
MSE VS [5]	**0.6295**	**0.6268**	**8.5310**
MSE w=local saliency	0.5410	0.5267	9.2319
MSE w=proposed MSA	0.5669	0.5593	9.0427
SSIM	0.5411	0.5231	9.2315
SSIM VS [5]	0.6308	0.6187	8.5310
SSIM w=local saliency	0.6064	0.5825	8.7284
SSIM w=proposed MSA	**0.6470**	**0.6334**	**8.3698**
MS-SSIM	0.7556	0.7474	7.1911
MSSSIM VS [5]	0.7583	0.7468	7.1570
MS-SSIM w=local saliency	0.7623	0.7589	7.1042
MS-SSIM w=proposed MSA	**0.8009**	**0.7964**	**6.5726**
VIF	0.5322	0.5297	9.2936
VIF w=local saliency	0.6734	0.6566	8.1153
VIF w=proposed MSA	**0.6946**	**0.6959**	**7.8968**

4. CONCLUSIONS

In this paper we have provided a novel motion saliency estimation method for video sequences considering the motion between successive frames, and their corresponding parametric camera motion representation, for VQA. The proposed saliency maps are incorporated in the spatial pooling stage of several video quality metrics. Experimental eval-

Table 2: Performance enhancement in terms of SROCC for our proposed method on LIVE database

#	Distortion class	MSE	SSIM	MS-SSIM	VIF
1	H264 + wireless	-0.0291	0.1328	0.0638	0.1538
2	H264 + IP	0.1139	0.1166	0.0206	0.1326
	average (#1,#2)	0.0424	0.1247	0.0422	0.1432
3	H264	0.0251	0.1099	0.0901	0.1546
4	MPEG2	0.0238	0.1110	0.0662	0.0463
	average (#3,#4)	0.0245	0.1105	0.0782	0.1005
	All data	0.0202	0.1103	0.0490	0.1662

Figure 2: Example frames of four sequences of the LIVE database. Reference frames R_n and the corresponding motion saliency maps MSA_n (first and second rows respectivelly).

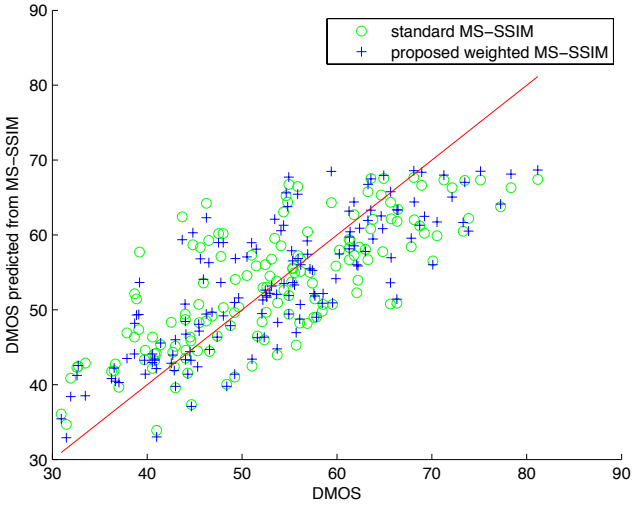

Figure 3: Scatter plot of predicted DMOS using standard MS-SSIM and weighted MS-SSIM using the proposed method (blue and green marks respectively) versus DMOS. Evaluation on LIVE video database.

uation on the LIVE video database has shown that thus, objective metrics are more closely in accordance with the subjective ground-truth scores, which is an indicator that motion saliency can be beneficial for VQA. Motion saliency aware temporal pooling and consideration of more neighboring frames remain interesting topics for future exploration.t

5. REFERENCES

[1] M. G. Arvanitidou, M. Tok, A. Krutz, and T. Sikora. Short-term motion-based object segmentation. In *Proc. of the IEEE Internl. Conf. on Multimedia & Expo*, Barcelona, Spain, Jul. 2011.

[2] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya. Visual attention in quality assessment. *IEEE Signal Proces. Mag.*, 28(6):50–59, Nov. 2011.

[3] D. Farin and P. H. de Witha. Evaluation of a feature-based global-motion estimation system. In *SPIE Visual Comm. and Image Proces.*, volume 5960, 2005.

[4] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[5] L. Ma, S. Li, and K. Ngan. Motion trajectory based visual saliency for video quality assessment. In *Proc. of the IEEE Internl. Conf. on Image Proces.*, pages 233–236, sept. 2011.

[6] A. Moorthy and A. Bovik. Efficient video quality assessment along temporal trajectories. *IEEE Trans. on Circuits and Systems for Video Technology,*, 20(11):1653–1658, Nov. 2010.

[7] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(7):629 –639, Jul. 1990.

[8] M. H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Trans. on Broadcasting*, 50(3):312–322, Sep. 2004.

[9] K. Seshadrinathan and A. C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Trans. on Image Proces.*, 19(2):335–350, 2010.

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

[11] H. Sheikh and A. Bovik. Image information and visual quality. *IEEE Trans. on Image Proces.*, 15(2):430–444, Feb. 2006.

[12] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, ph.II. http://www.vqeg.org, 2003.

[13] Y. Wang, T. Jiang, S. Ma, and W. Gao. Novel spatio-temporal structural information based video quality metric. *IEEE Trans. on Circuits and Systems for Video Technology*, PP(99):1, 2012.

[14] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Proces.*, 13(4):600–612, Apr. 2004.

[15] Z. Wang and Q. Li. Video quality assessment using a statistical model of human visual speed perception. *Journal of the Optical Society of America*, 24(12):B61–B69, Dec 2007.

[16] Z. Wang, E. Simoncelli, and A. Bovik. Multiscale structural similarity for image quality assessment. In *Proc. of the Asilomar Conf. on Signals, Systems and Computers*, volume 2, pages 1398– 1402 Vol.2, Nov. 2003.