Motion-Aware Video Quality Assessment

Marina Georgia Arvanitidou and Thomas Sikora

Communication Systems Group
Technische Universität Berlin
Berlin Germany

***Abstract*—This work focuses on considering motion towards improving video quality assessment algorithms. The improvement refers to improving computational video quality assessment algorithms in order to be in closer agreement with the subjective evaluation of video quality. We propose a motion saliency model that exploits motion features on spatial level and also an approach for consideration of global motion in the temporal dimension, leading to further improvements in the accuracy of video quality assessment. We perform evaluation by integrating our approaches in existing objective quality models and also by comparing them to existing related state-of-the-art video quality assessment methods.**

I. Introduction

In video quality assessment (*VQA*), motion plays a critical role. By applying image quality assessment metrics (*IQA*) on frame level and subsequently fusing these local assessments using average, motion is ignored. The consideration of only spatial correlations (e.g. PSNR, SSIM [12]) is satisfactory for IQA, more sophisticated considerations are required though in the case of VQA [3], [5] [6], where temporal correspondences constitute a determining factor.

In the literature there are several approaches in this direction. The *video quality model* (VQM) [5], which is adopted by the american national standards institute (ANSI), analyses 3D spatio-temporal blocks to extract features for estimating the video quality map. Moorthy *et al.* [3] propose the motion-compensated structural similarity index (MC-SSIM) that combines block-based motion estimation with SSIM [12]. Each 8×8 block of the reference and the distorted frames is motion compensated using the corresponding preceding frame and the results are used to evaluate temporal quality. The *motion-based video integrity evaluation* (MOVIE) metric [6] utilises properties of the visual cortex neurones to track perceptually relevant distortions both spatially and temporally and evaluates motion quality along computed motion trajectories. Relying on 3D optical flow estimation, the latter is a rather computationally complex metric.

Motivation and proposed approach

Objective quality assessment models for image and video quality assessment often compute quality scores based on the assumption that content over space and time is of equal interest to the observer. It is assumed thus that distortions in different regions in space and time contribute equally to the overall quality perception of the video. Nevertheless, humans do not see in a way that resembles linear scanning. Rather, it is claimed to sample and process the physical world in a way that is space and temporally variant, which has led to considerable interest in visual quality assessment approaches [1] in recent years.

Towards understanding how traditional image quality assessment metrics can benefit from perceptual knowledge and motion, we illustrate an example indicating the shortcomings of the traditionally used PSNR with respect to the way visual content is in general assessed by humans. Figure 1 depicts a fish swimming in the seabed. The viewer will typically focus his attention mainly on the fish and secondly on the seabed. Consequently, the blurring blemish on the sea region (bottom left corner) in Figure 1(a) will be probably perceived only under thorough examination. On the contrary, the blurring which takes place on the region depicting the fish, in Figure 1(b), will be more pronouncedly perceived compared to the former case. Thus the location of the second blurring seems to play an important role on the perceived quality, resulting in the impression that Figure 1(b) has worse quality than Figure 1(a). Evaluation of the quality using PSNR is however not that revealing; both images have the same PSNR.

In the case of video sequences temporal dependencies between frames constitute valuable information. This motivates us to take them into account for assessing the quality of video sequences. Based on the established connection between motion and perception and considering that moving regions will likely attract the viewer's attention, in this work we exploit motion for video quality assessment, in spatial and temporal level. The main idea of the proposed method is to

(a) Distortion at the background (b) Distortion at the foreground

Fig. 1: Deviation of objective and subjective quality assessment on the *BBC fish* sequence. The distorted areas are indicated with red boxes.

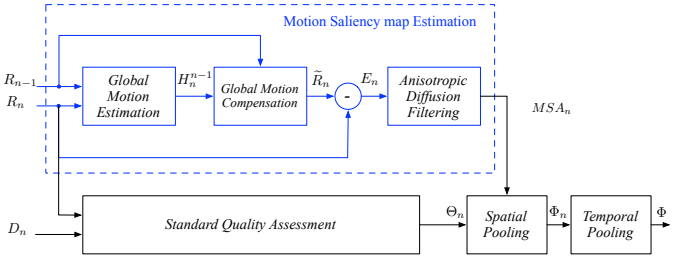


Fig. 2: System overview.

include relative motion information between frames into the calculation of the objective VQA. This is performed by taking into account significant relative motion between frames in the spatial and also in the temporal pooling phase of the objective metric.

The overall system is illustrated in Figure 2. At first stage, the motion model H_n^{n-1} between two successive frames of the reference sequence R_{n-1} and R_n is computed. Based on H_n^{n-1} and R_{n-1}, the estimated frame $\tilde{R}_n$ is computed and subsequently subtracted from R_n. This results in the global motion compensated absolute error frame E_n where high error energy indicates motion of the foreground area. E_n is subsequently filtered using anisotropic diffusion resulting in the MSA_n map that assigns a weight to each pixel location. In the spatial pooling step the standard quality assessment measure, MSA_n is used as a significance map and is combined with Θ_n yielding the local motion saliency-aware model Φ_n. Finally, the local quality metrics are combined in the temporal pooling stage to result in the overall quality measure Φ.

The proposed motion saliency estimation detects regions that contain noticeable motion, in order to emphasize their effect to the image quality index in the spatial pooling stage. If a distortion occurs in a region that contains motion, it is expected to attract the attention of the viewer and to have thus negative impact on the quality assessment in comparison to a distortion that occurs in a region not containing motion. The foreground and background segment regions assumed not to be known.

In the temporal dimension, we propose an approach for consideration of global motion, leading to further improvements in the accuracy of video quality assessment. The proposed global motion indicator considers temporal dependencies between frames in a way that distortions are more profoundly perceived in cases of large global motion.

II. Motion saliency model for spatial pooling

The eight-parameter perspective motion model is used at first stage to describe the background motion between two successive frames of the reference sequence R_{n-1} and R_n. This is realised using the feature-based global motion estimation approach which detects feature points correspondences between two sets of features for successive frames using the *Kanade - Lucas - Tomasi* (KLT) tracking algorithm [9]. Based on the detected features, we use the *random sample consensus* (RANSAC) [2] approach for fast and accurate motion model (H_n^{n-1}) estimation. Considering that these feature correspondences represent motion between this pair of images, the global motion is estimated.

Based on the connection between motion and perception and considering that moving areas will likely attract the viewer's attention our goal is to exploit them for video quality estimation. Furthermore, studies on the human visual system have shown that the human retina is highly space variant in processing and sampling of visual information [1]. The accuracy is highest in the central point of focus, the fovea, and the peripheral visual field is perceived with lower accuracy. Therefore, we consider anisotropic diffusion filtering for the error frame [4] that offers a non-linear and space-variant filtering and is found to be interestingly related to the neural dynamics of brightness perception [13]. 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 $MSA(x, y, n)$, where x, y are the pixel coordinates in the horizontal and vertical direction, and n is the frame number.

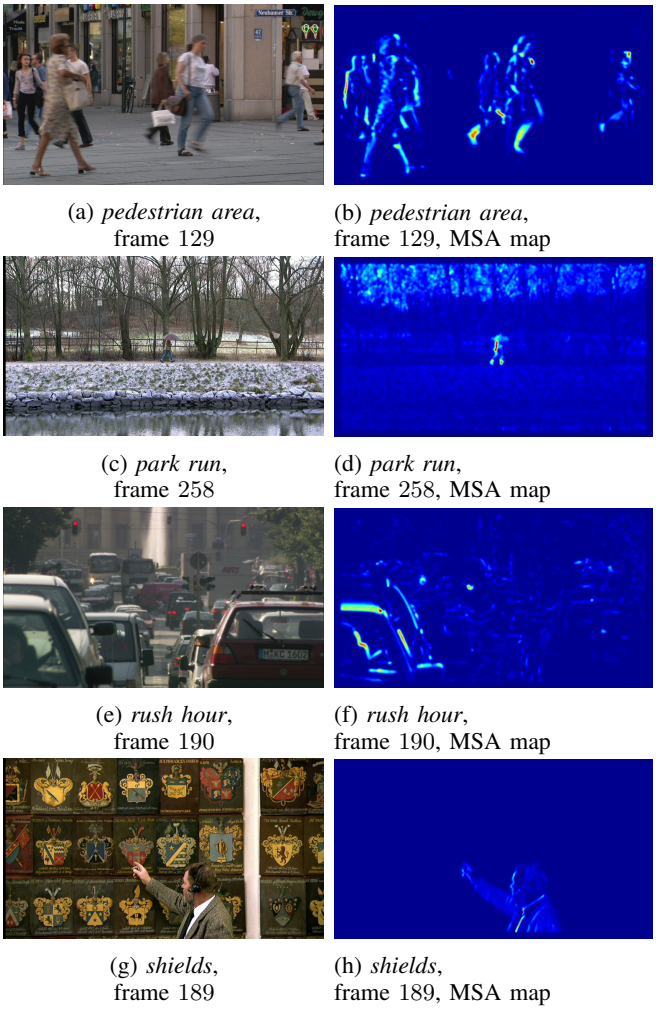
In this way higher weighting is assigned to regions that have moved between two successive frames and we expect that they are more likely to attract visual attention in comparison to other areas that have not moved (or have not moved in relation to the background). Other features such as contrast, colour and structural information will be considered implicitly through the incorporation of standard objective metrics. As shown in the examples in Figure 3 the proposed motion saliency estimation approach indicates moving areas as "warmer" ones in the MSA maps.

In the spatial pooling stage, conventional image quality metrics 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. Towards avoiding uniform spatial pooling, we employ a weighted mean pooling strategy where the estimated motion saliency maps are incorporated in conventional image quality metrics in frame level. For multiscale models, that use M scales, the weighting map is scaled correspondingly

III. Global motion indicator for temporal pooling

Temporal pooling follows the spatial pooling stage, as the local weighted quality scores have to be taken into account to output the overall quality score.

In order to account for the perceived quality degradation due to global motion, we consider the folowing. The perspective motion model H_n^{n-1} that describes motion between two successive frames contains parameters that are closely related to specific transformations. h_1 reveals rotation and/or scaling, h_2 and h_5 indicate translation in the horizontal and vertical direction respectively, h_3 corresponds to rotation, while the



(a) *pedestrian area*, frame 129
(b) *pedestrian area*, frame 129, MSA map
(c) *park run*, frame 258
(d) *park run*, frame 258, MSA map
(e) *rush hour*, frame 190
(f) *rush hour*, frame 190, MSA map
(g) *shields*, frame 189
(h) *shields*, frame 189, MSA map

Fig. 3: The first column depicts example reference frames R_n of the LIVE video database. The second column depicts the corresponding motion saliency maps MSA_n as heat maps, where warmer regions indicate higher motion saliency.

rest of the parameters (h_0, h_4, h_6, h_7) are related to more that one basic transformations.

We propose the *global motion indicator* (gmi) for weighting of the frame-level quality scores across time based on the variation of global motion on the temporal dimension, assuming that large camera motion causes distortions to have a greater impact on perceived video quality and that the perception of distortions is affected mostly by translational motion. The global motion indicator is defined as $gmi(n) = \mathbf{F} \cdot (h_0 \;\; h_1 \;\; h_2 \;\; h_3 \;\; h_4 \;\; h_5 \;\; h_6 \;\; h_7)^T$ where $h_k,\ k = 0, ..., 7$ denote the elements of the eight-parameter homography of the n-th frame derived from global motion estimation using RANSAC and $\mathbf{F}$ is the enhancement matrix defined as $\mathbf{F} = (1 \;\; 1 \;\; f \;\; 1 \;\; 1 \;\; f \;\; 1 \;\; 1)$ where $f = 10$.

IV. Evaluation

For performance evaluation of the proposed approach and towards reproducible research, we employ the *LIVE video quality database* [7] which is publicly available. The LIVE database contains 150 distorted videos obtained from 10 uncompressed reference videos (768×432 pixels) 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 Pearson linear correlation coefficient ρ_p and the Spearman rank order correlation coefficient ρ_s between subjective and objective evaluation of video quality are used as prediction performance indicators according to the Video Quality Experts Group (VQEG) recommendation [10].

We use MSE, SSIM [12], MS-SSIM [11] and VIF [8] as objective image quality assessment metrics. Table I reports for each objective IQA metric the improvement using the proposed spatial pooling using motion saliency, denoted as "MSA", the proposed temporal pooling denoted as "GMI". The temporal pooling method using a temporal pooling function [14] is denoted as "TPF". For each evaluation model we highlight the best results with boldface. The performance of the state-of-the-art VQA models MC-SSIM [3], VQM [5] and MOVIE index [6] is also reported in Table I.

The weighted models using the proposed MSA approach perform better compared to non-weighted models. The improvement ist increased employing additionally the proposed GMI for temporal pooling which also outperforms the previously proposed TPF temporal pooling [14]. The proposed method for the case of MSA-weighted MS-SSIM using the gmi (referred as MS-SSIM-MSA-GMI) outperforms the state-of-the-art motion models, which confirms the validity and the encourages further perspectives of the proposed approach.

Study on each distortion class

To examine the effect of the proposed weighting on different distortion types, we present in Table II the performance improvement, in terms of Spearman rank order correlation coefficient, introduced by the proposed method for each distortion class separately. As expected, our proposed approach contributes on average more in cases of transient distortions (in the presence of packet losses, classes #1 and #2) compared to cases with uniformly distributed distortions (no packet losses, classes #3 and #4). The average improvement in terms of ρ_s for distortion classes #1 and #2 is 0.0881, whereas for classes #3 and #4 is 0.0784, whereas the overall trend of outperformance of motion saliency spatial pooling remains unchanged across the various distortion types.

TABLE I: Performance evaluation of the proposed methods on LIVE video quality database. MC-SSIM, VQM and MOVIE performance as reported in [3].

Algorithm	ρ_p	ρ_s
MSE	0.5614	0.5391
MSE-MSA	0.5669	0.5593
MSE-MSA-TPF [14]	0.5685	0.5609
MSE-MSA-GMI	**0.5748**	**0.5676**
SSIM	0.5411	0.5231
SSIM-MSA	0.6470	0.6334
SSIM-MSA-TPF [14]	0.6386	0.6217
SSIM-MSA-GMI	**0.6678**	**0.6420**
MS-SSIM	0.7556	0.7474
MS-SSIM-MSA	0.8009	0.7964
MS-SSIM-MSA-TPF [14]	0.7892	0.7834
MS-SSIM-MSA-GMI	**0.8155**	**0.8096**
VIF	0.5322	0.5297
VIF-MSA	0.6946	0.6959
VIF-MSA-TPF [14]	0.6846	0.6801
VIF-MSA-GMI	**0.7092**	**0.7121**
MC-SSIM [3]	0.6976	0.6791
VQM [5]	0.7236	0.7026
MOVIE [6]	0.8102	0.7861

TABLE II: Performance improvement in terms of ρ_s of our proposed method using motion saliency (MSA) over standard metrics on the LIVE database for each distortion class.

#	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

V. Conclusion

We proposed a novel motion saliency estimation method for video sequences that exploits motion features on spatial level considering motion between successive frames, and their corresponding parametric camera motion representation. Moreover, we proposed a temporal pooling approach that enables further improvements of objective metrics by exploiting global motion in the temporal dimension.

The proposed models have been incorporated in several objective quality metrics and it has been shown that their performance is improved and existing state-of-the-art VQA approaches are outperformed. It has been shown that the discrepancy between objective metrics and subjective evaluation is reduced, which is an indicator that motion is an important aspect that affects the perception of visual quality assessed by humans and the incorporation of motion and especially global motion is beneficial for VQA.

References

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

[2] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981.

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

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

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

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

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

[8] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, 2005.

[9] Carlo Tomasi and Takeo Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon University, 1991.

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

[11] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *proceedings of the Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1398 – 1402 Vol.2, nov. 2003.

[12] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[13] Joachim Weickert. *Anisotropic diffusion in image processing*, volume 1. Teubner Stuttgart, 1998.

[14] Junyong You, Jari Korhonen, and Andrew Perkis. Attention modeling for video quality assessment: Balancing global quality and local quality. In *proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 914–919, 2010.