SPATIAL AND TEMPORAL POOLING OF IMAGE QUALITY METRICS FOR PERCEPTUAL VIDEO QUALITY ASSESSMENT ON PACKET LOSS STREAMS

Junyong You, Jari Korhonen, and Andrew Perkis

Norwegian University of Science and Technology, NO-7491, Trondheim, Norway
junyong.you@q2s.ntnu.no, jari.korhonen@q2s.ntnu.no, andrew@iet.ntnu.no

ABSTRACT

Video streaming through bandwidth-limited channels often suffer from packet losses. Therefore, perceptual quality assessment on video sequences with packet losses is a critical issue in digital video communications. This paper analyzes several image quality metrics and evaluates their applications using spatial and temporal pooling schemes in perceptual video quality assessment for video streams with packet losses. Several approaches using Minkowski summation and averages over different distorted spatial regions and temporal frames to pool the spatial and temporal qualities are evaluated. The experimental results with respect to the subjective video quality measurements demonstrate that the subjects are more sensitive to the most annoying spatial regions and temporal segments when assessing the video quality of the lossy streams.

Index Terms— Perceptual video quality assessment, image quality metric, packet loss stream

1. INTRODUCTION

Packet losses often occur when compressed video is transmitted through communication networks, even though network providers are dedicated to keep the packet loss rate as low as possible and monitor the traffic in their network to assure acceptable network performance. Although many error concealment methods have been proposed to conceal the effect of packet losses at the receiver, packet losses may still result in annoying video quality degradation. The problem of evaluating video quality with packet loss is challenging, due to the fact that human visual perception system is very complicated. As the first step towards developing an accurate quality metric for videos affected by packet losses, in this work we have investigated the use of (still) image quality metrics (IQMs) for quality assessment on lossy video sequences.

A number of objective video quality metrics have been proposed, and they can be grouped into three categories according to availability of reference video: full reference (FR), reduced reference (RR), and no reference (NR). Based on the methodologies employed in the metrics, they can also be classified into psychophysical approach and engineering approach [1]. Currently, all metrics mainly focus on evaluating the qualities frame by frame or segment by segment, and then the measurement results for individual frames or segments are combined to assess the overall quality of the video. For example, perceptual distortion metric (PDM), proposed by Winkler et al., detects the quality degradation of distorted frames compared against the original frames, and then pools all detected degradations spatially and temporally to calculate the overall quality [2]. Video quality metric (VQM), proposed by Pinson et al., is based on a comparison of several quality features that are extracted from spatial-temporal (S-T) regions and an integration of these features into a quality values using spatial and temporal collapsing and pooling functions [3]. A spatial pooling scheme based on visual importance to improve the correlation with subjective judgment has been studied in [4].

For quality assessment of video sequences with packet losses, some network parameters, such as packet loss rate and packet loss burst length, can be employed. Based on these network parameters and some compression parameters, such as bit rate, encoded frame type, frame rate, Mohamed et al. proposed to use a neural network to quantify the quality of video flows [5]. The impact of several factors to the perceptual quality in packet loss videos, such as error length, loss location, the number of losses, and loss patterns, has been examined [6].

In this paper, we first investigate several well-known image quality metrics and compare their performance in assessing certain image degradation types related to compression and packet loss. Based on an intensive analysis of the spatial and temporal factors on video quality assessment, we evaluate different approaches by applying IQMs and comparing the results against subjective quality assessment results on the respective lossy video sequences. According to the experimental results, some general conclusions can be drawn and our intention in the future is to develop more accurate quality metrics to measure the quality of packet loss videos by taking into account related characteristics, including spatiotemporal masking and quality features related to lost packets, as well as the attributes of human visual system (HVS).

2. IMAGE QUALITY METRICS AND ANALYSIS

Image quality assessment is a mature research issue and many IQMs have been proposed, that have been proven to be credible for certain degradation categories. However, none of these metrics have been found to work well on all types of quality degradation, and the traditionally used metrics, such as MSE and PSNR, are still better than HVS-based methods for certain special degradation types, such as added noise [7]. However, in a general image/video compression and transmission system, pixel-based noise does not always occur. This is why we tested 11 well-known full reference IQMs as listed in Table I on three different distortion types: JPEG compression, JPEG transmission errors, and local block-wise distortions that are extracted from a public image quality database [8]. For the detailed descriptions of these image quality metrics, readers can refer to [7][9].

In this work, the luminance component of color images was employed to compute the quality using the 11 IQMs, and then a nonlinear regression operation between the metric results (*IQ*) and the subjective mean opinion scores (MOS) was performed using the following logistic function:

$$MOS_P = a1/(1+\exp(-a2\cdot(IQ-a3))) \quad (1)$$

The nonlinear regression operation was used to transform the set of metric results to a set of predicted MOS values, MOS_P, that were compared to the actual subjective scores using Pearson correlation coefficient as a criterion. The resulting Pearson correlation coefficients are reported in Table I.

3. IQM BASED VIDEO QUALITY ASSESSMENT ON PACKET LOSS STREAMS

Besides the distortion introduced by lossy compression, video quality is also distorted by packet losses in a noisy channel. Packet losses may occur at any frames and any regions in a frame. In this Section, we study the quality combinations over spatial regions and temporal frames, as well as their influence on the video quality assessment, based on the predicted quality of every frame using the IQMs.

Some existing video quality metrics, such as PDM, employ Minkowski summation, as given in Eq. (2), to pool the qualities over all frames into a temporal combination, where P is an exponent and N is the number of elements in vector $V=\{V_i\}$.

$$Minkowski = \sqrt[P]{\sum V_i^P / N} \quad (2)$$

Other methods, e.g. VQM, use temporal collapsing functions to reflect the influence of unbalanced distortions over time. Therefore, we also studied different temporal pooling approaches, including Minkowski summations with different exponents as well as averaging over a set of frames with different distortion levels.

The measured video quality is also affected by the spatial pooling schemes. Most IQMs compute a distortion map between the reference and distorted images to depict the distribution of quality degradation at image pixels, and the overall quality is usually computed as a mean over all the pixels in the distortion map. However, it has been proven that human perception is more sensitive to certain regions, called attention region. We have proposed a visual attention based video quality metric, which was proven to be accurate for measuring the quality degradation caused by compression errors [10]. As the influence of packet loss on video quality is evidently different from other distortion types, such as compression distortion, we studied different spatial pooling schemes of IQMs and evaluated their performance in quality assessment for packet loss streams. The spatial pooling schemes include: (1) Minkowski summation over all pixels in the distortion map; (2) attention model proposed in [10] detecting the attention regions in every video frame and averaging over the distortion map in the attention regions only; (3) averaging over different regions representing areas with different distortion levels. Because VSNR, IFC and VIF do not make use of the distortion map, we excluded them from our experiments.

Finally, joint temporal and spatial pooling schemes were created to evaluate the performance of studied IQMs for packet loss streams. We computed the Pearson correlation coefficients to compare the metric results using different pooling schemes, fitted using the logistic function in Eq. (1), against the actual subjective scores. The detailed description of the experiments is given in the following Section.

4. EXPERIMENTS AND DISCUSSIONS

4.1. Subjective quality assessment on packet loss videos

To evaluate the performance of the IQMs, we have used the test sequences and the results from the subjective experiments conducted at Ecole Polytechnique Fédérale de Lausanne (EPFL) and Politecnico di Milano (PoliMi) [11]. Six video sequences at CIF resolution, covering a wide range of spatial and temporal activity levels, were used in the study. For each video sequence, a compressed H.264/AVC stream was generated using the reference encoder JM14.2. Twelve corrupted streams were generated by dropping packets from the encoded stream, using given error patterns as explained in [11]. Therefore, 6×13=78 distorted video sequences in total were generated and employed in the subjective measurements. A single stimulus subjective methodology was adopted for the subjective quality assessment. Finally, an analysis of variance (ANOVA), offset correction, and outlier detection and removal were performed to obtain MOS values.

Two sets of subjective scores were obtained at EPFL and PoliMi, respectively. In that study, the correlation between metric values and the PoliMi results was slightly higher than in respect of the EPFL results. We adopted the PoliMi results in our experiments. In addition, even though the single stimulus method is a no reference method, subjects always compare a distorted video to an imaginary "reference" video with the best possible quality which might exist to judge the quality of the distorted video. Therefore, we believe that the performance comparison against the studied 11 full reference IQMs is meaningful.

4.2. Experimental results of IQMs on video quality assessment

In this work, the computation of video quality is based on the image quality values frame by frame; therefore, the quality of each frame was computed first using each of the studied 11 IQMs to compare the distorted sequences and original sequences. The performance of different temporal pooling schemes was then evaluated, including Minkowski summation with different exponents and direct averaging over the distorted frames. In video

TABLE I IQM DESCRIPTIONS AND PEARSON CORRELATION COEFFICIENTS ON DIFFERENT DISTORTION TYPES

IQM	Descriptions	JPEG Comp.	JPEG Trans.	Local Dist.	All
PSNR	Peak signal-to-noise ratio	0.8683	0.7576	0.6384	0.7544
SSIM	Single scale structural similarity	0.9134	0.8233	0.8925	0.5432
MSSIM	Multi-scale structural similarity	0.9590	0.8648	0.7968	0.8469
VSNR	Visual signal-to-noise ratio	0.9345	0.8086	0.2801	0.6906
PHVS	[9] Modified PSNR based on HVS	0.9816	0.8073	0.6606	0.7556
IFC	Information fidelity criterion	0.8561	0.8237	0.7115	0.1151
VIF	Visual information fidelity	0.9538	0.8760	0.8407	0.6381
UQI	Universal quality index	0.8010	0.8464	0.8587	0.5331
NQM	Noise quality measure	0.9360	0.7320	0.2202	0.7000
WSNR	Weighted signal-to-noise ratio	0.9467	0.7359	0.2704	0.7198
JND	Just noticeable distortion model	0.8829	0.7362	0.5554	0.7023

quality assessment, Minkowski summation is widely used in pooling the temporally altering distortion levels. Different exponents are used in different metrics.

In this study, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 8, and 10 were selected for exponents in the Minkowski summation. Figure 1 (a) shows the comparison results between the IQM results and MOS values in terms of Pearson correlation coefficients. Some quality metrics, such as VQM, use temporal collapsing functions to sort the quality values from low to high and then compute the average of all the quality values greater than a given threshold level, because human perception is supposed to be more sensitive to the greater distortions. Therefore, we evaluated the performance of averaging over frames with different quality values, as computed by IQMs. The threshold levels included 0.5, 1, 2, 5, 10, 20, 30, 40, 50, 80, and 100. For example, if the threshold level was 10, we sorted the quality values of all frames from low to high, and then computed the average of the quality values of the frames with quality level greater than 90% threshold level. The threshold level 100 denoted that averaging was performed over the quality values of all frames. Figure 1 (b) shows the obtained results.

According to the performance comparison of IQMs, we observed that the IQMs that perform best with still images usually give the best results in evaluating video quality as well. This observation demonstrates the feasibility to adopt the IQMs in video quality assessment whilst the attributes of temporal activities in video sequences need to be taken into account. Based on the comparison between Minkowski summations and averaging over different frames, we have observed that the latter performs better than the Minkowski summation, regardless of the used IQMs and exponents. Thus, we can draw a conclusion that Minkowski summation is not an appropriate method for pooling the temporal qualities in order to compute the overall video quality from the individual video frames. Although Minkowski summation can be used to emphasize the effect of unbalanced distortion, we have found that the performance decreases when the exponent is increased. On the other hand, when averaging is used, the best performance was achieved when the threshold levels were set as 2, 5, or 10. This indicates that human perception is more sensitive to the frames that have the large distortions between 90% and 98% thresholds regarding the whole ranges of quality distortions. This observation is useful for designing an appropriate temporal pooling scheme for video quality metrics for packet loss streams.

In a similar fashion, we evaluated the performance of different spatial pooling schemes including Minkowski summation over all pixels, averaging over different distorted regions, and averaging over attention regions that were detected from the original sequences by using the attention model described in [10], in each frame. In our evaluations, we have used the same Minkowski exponents and threshold levels as used for evaluation of temporal pooling schemes. Eight of these 11 IQMs, including PSNR, SSIM, MSSIM, PHVS, UQI, NQM, WSNR, JND, use a quality map reflecting the distortion at each pixel either directly or indirectly, to compute the overall quality of an image. Therefore, we can use these image metrics to compute the quality values at each pixel and then pool these values into an overall quality of a frame by using different spatial pooling schemes. To compare the

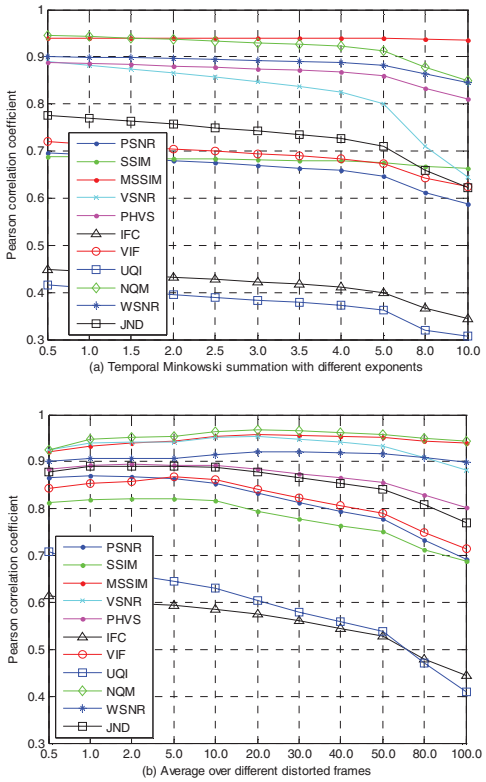

Fig. 1 Evaluation result on temporal pooling schemes

performance of different spatial pooling schemes fairly, we used the averaging over all frames as a temporal pooling scheme in each case.

Figure 2 (a) gives the results with Minkowski summation, and Fig. 2 (b) depicts the results when averaging is used. In general, the performance of Minkowski summation is also in this case worse than averaging over spatial regions, but this conclusion is not as evident as in the temporal pooling. Therefore, we assume that Minkowski summation can be used for spatial pooling, if appropriate exponent is chosen (suggested value is 2). Averaging over different distorted regions has a similar impact on the performance of IQMs as in temporal pooling. The experimental results suggest that human perception is more sensitive to the spatial regions with severe distortions, between 90% and 99.5% or even greater, in respect to the total range of distortions. Furthermore, the impact of unevenly distributed distortions to the overall video quality is stronger in spatial domain than in temporal domain. In our experiments, the regions with large distortions were usually the image blocks impacted by packet loss and not satisfactorily corrected by error concealment. As a comparison, the performance of the spatial pooling scheme based on attention regions is worse than other schemes, even though we obtained promising results when the scheme was used to assess the video quality of sequences with general distortion types, such as compression and noise [10]. A similar observation was reported in

TABLE II EVALUATION RESULTS OF SPATIOTEMPORAL POOLING SCHEMES

Parameters	PSNR	SSIM	MSSIM	PHVS	UQI	NQM	WSNR	JND
Pearson correlation	0.879	0.923	0.978	0.909	0.918	0.983	0.956	0.893
Threshold level (spatial)	10	5	10	20	0.5	2	10	20
Threshold level (temporal)	2	20	5	5	50	10	10	5

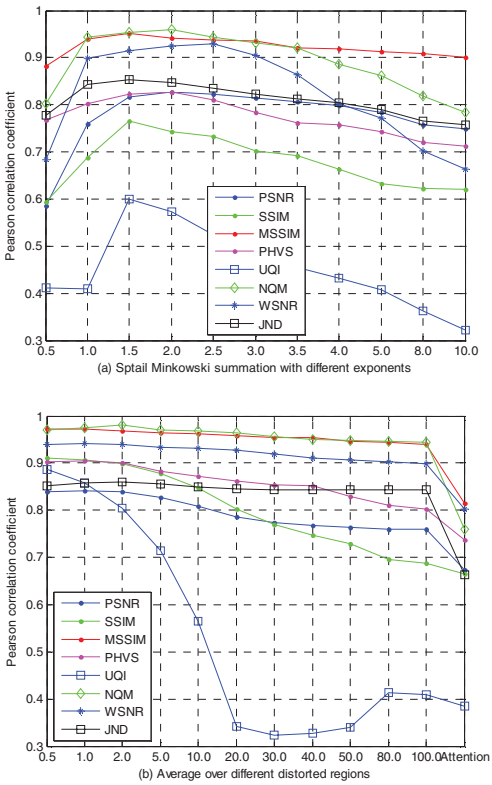

Fig. 2 Evaluation results on spatial pooling schemes

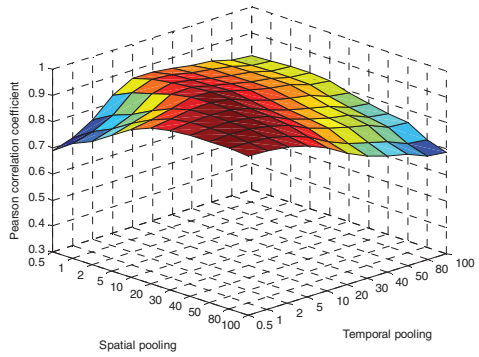

Fig. 3 SSIM evaluation results on spatiotemporal pooling scheme

[12]. The reason might be that subjects usually evaluate the video quality according to those regions with the most severe degradation, even in case these regions are not the attention regions. In the future work, we will develop more appropriate attention models for quality assessment for packet loss videos.

Finally, we integrated the spatial and temporal pooling schemes into a spatiotemporal pooling scheme which was based on averaging over different distorted spatial regions and temporal frames. As an example, Pearson correlation coefficients of SSIM using averaging over different regions and frames are illustrated in Fig. 3. In addition, Table II presents the highest correlation of these IQMs and the corresponding threshold levels in spatial and temporal pooling giving the best performance. According to the statistical results on the spatial, temporal, and spatiotemporal pooling schemes, the most distorted spatial regions (suggested 90%-99.5%) and temporal frames (suggested 90%-98%) should be taken into account. Other regions and frames with smaller distortions can be ignored.

5. CONCLUSIONS

In this paper, we have analyzed and evaluated different spatial and temporal pooling schemes for image quality metrics that are used for video quality assessment of streams impacted by packet losses. Different image quality metrics were used to assess the quality of each video frame or a spatial region of a frame. Different spatial and temporal pooling schemes, including Minkowski summation and averaging, were then used to combine the obtained quality estimates for spatial regions and temporal frames into an overall quality measure of a video sequence. The experimental results show that Minkowski summation is not an appropriate scheme for pooling the temporal quality levels, compared to spatial pooling using Minkowski summation. We have also observed that human perception on video quality judgment is mainly dependent on those regions and frames with the most severe distortion, while the attention regions do not always influence the quality evaluation in the packet loss video streams.

REFERENCES

[1] H. R. Wu and K. R. Rao (Ed.), Digital Video Image Quality and Perceptual Coding. CRC Press, 2006.
[2] S. Winkler, Digital Video Quality: Vision Models and Metrics, John Wiley & Sons, 2005.
[3] 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.
[4] A. K. Moorthy, and A. C. Bovik, "Visual Importance Pooling for Image Quality Qssessment," IEEE J. Selected Topics in Signal Processing, vol. 3, no. 2, pp. 193-201, Apr. 2009.
[5] S. Mohamed, and G. Rubino, "A Study of Real-time Packet Video Quality using Random Neural Networks," IEEE. Trans. Circuits and Systems for Video Technology, vol. 12, no. 12, pp. 1071-1083, Dec. 2002.
[6].T. Liu, Y. Wang, J. M. Boyce, et al. "A Novel Video Quality Metric for Low Bit-rate Video Considering both Coding and Packet-loss Artifacts," IEEE J. Selected Topics in Signal Processing, vol. 3, no. 2, pp. 280-293, Apr. 2009.
[7] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," IEEE Trans. Image Processing, vol. 15, no. 11, pp. 3441-3452, Nov. 2006.
[8] TID2008 page, http://www.ponomarenko.info/tid2008.htm.
[9] N. Ponomarenko, F. Battisti, K. Egiazarian, et al., "On Between-coefficient Contrast Masking of DCT Basis Functions," in Proc. of VPQM'07, Scottsdale, Arizona, USA, Jan. 2007.
[10] J. You, A. Perkis, M. M. Hannuksela, and M. Gabbouj, "Perceptual Quality Assessment Based on Visual Attention Analysis," in Proc. of ACM Multimedia'09, Beijing, China, Oct. 2009.
[11] F. D. Simone, M. Naccari, M. Taglisacchi, et al., "Subjective Assessment of H.264/AVC Video Sequences Transmitted Over A Noise Channel," in Proc. of QoMEX'09, San Diego, California, USA, Jul. 2009.
[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. of IEEE ICIP'07, San Antonio, Texas, USA, Sep. 2007.