SPATIO-TEMPORAL QUALITY POOLING ACCOUNTING FOR TRANSIENT SEVERE IMPAIRMENTS AND EGOMOTION

Jincheol Park†, Kalpana Seshadrinathan‡, Sanghoon Lee†and Alan C. Bovik§

†Wireless Network Lab., Center for IT of Yonsei University, Seoul, Korea, 120-749.
‡Intel Labs, Intel Corporation, Santa Clara, CA.
§Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX.

ABSTRACT

With the increasing popularity of video applications, the reliable measurement of perceived video quality has increased in importance. We study methods for pooling video quality scores over space and time. The method accounts for localized severe impairments of the signal which exhibit significant influence on the subjective impression of the overall signal quality. It also accounts for the effect of camera motion (egomotion) on perceived quality. The method arrived at is tested on the LIVE Video Quality Database and is shown to perform quite well.

1. INTRODUCTION

Video quality assessment (VQA) deals with predicting the perceptual quality of video sequence. An important hypothesis in the field of VQA is that local spatial and temporal regions of very poor quality substantially affect the overall subjective perception of quality [1][2]. This suggests the need for pooling methods that extract these influential poor quality scores and emphasize them when finding the overall quality score. We propose a pooling method, which we term Influential Quality Pooling or IQpooling, based on the hypothesis that such severe impairments have a substantial impact on human quality judgement. In IQpooling, for spatial pooling, influential poor quality scores are classified by a slope criterion applied on the sorted quality scores, whose curve tends to saturate towards good quality scores. Another important factor that affects perceived video quality is the overall motion of a video frame or egomotion. We adaptively apply a slope criterion using computed egomotion to perform temporal pooling and the most significant poor quality scores are captured using a k-means clustering procedure [3]. The performance evaluation of IQpooling on the LIVE Video Quality database [4] shows considerable performance improvement compared to previous VQA algorithms that use simpler spatial or temporal pooling.

2. WEAKNESS OF SAMPLE MEAN AS A POOLING METHOD

An important component that affects the performance of a VQA algorithm is the manner in which local quality scores are combined or pooled to predict an overall quality score for an entire image or video. One simple way of pooling the local quality scores is to use the mean value of the local scores to predict the overall quality. Mean based pooling has been widely used due to its simplicity. Various quality metrics have utilized mean based pooling including mean square error (MSE) and the SSIM index [5][6]. However, mean based pooling may not be consistent with how a human observer evaluates the video quality. Figure 1 illustrates the problem with using mean as the pooling method. A number of distortions caused by compression and lossy transmission of video occurs in specific regions of the video [4]. The severe distortion that occurs in part of the frame provides an observer with a very important cue for quality judgment. However, this cue is largely lost when pooling is performed using the mean, leading to poor prediction of the overall quality. This is illustrated in Fig. 1 and Fig. 2. Although Fig. 2(b) suffers from severe local distortions that are likely to adversely affect subjective quality as compared to Fig. 2(a), the mean SSIM score of Fig. 2(a) is higher than the mean SSIM score of Fig. 2(b).

Percentile pooling using the lowest $p\%$ of quality scores to predict the final score has been proposed as an improvement over using just the mean [2]. Percentile pooling weights the lowest $p\%$ of quality scores higher and has been shown to improve the performance of quality assessment algorithms. However, there is room for improvement in the percentile pooling. The percentile pooling proposed in [2] uses a fixed $p\%$ of scores, whereas the amount of impairment in a video frame can vary considerably which affects human judgement of quality. For example, let us suppose that two videos suffer from similar levels of distortions, but the distorted area is much larger in one video as compared to the other. When a fixed $p\%$ is used, the fixed percentile pooling can fail to distinguish the qualities of the two video frames if p is smaller than the smaller of the two distorted areas. However, the larger the distorted area in a video is, the worse the perceived quality is.

This work was supported by Agency for Defense Development under the contract UD100022ID, Korea

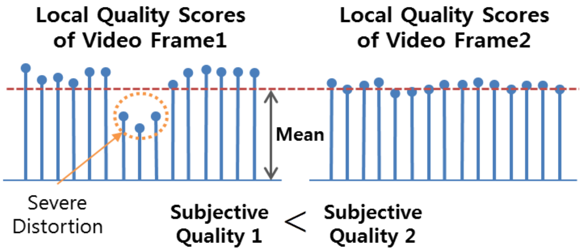

Local Quality Scores of Video Frame1
Local Quality Scores of Video Frame2
Mean
Severe Distortion
Subjective Quality 1
<
Subjective Quality 2

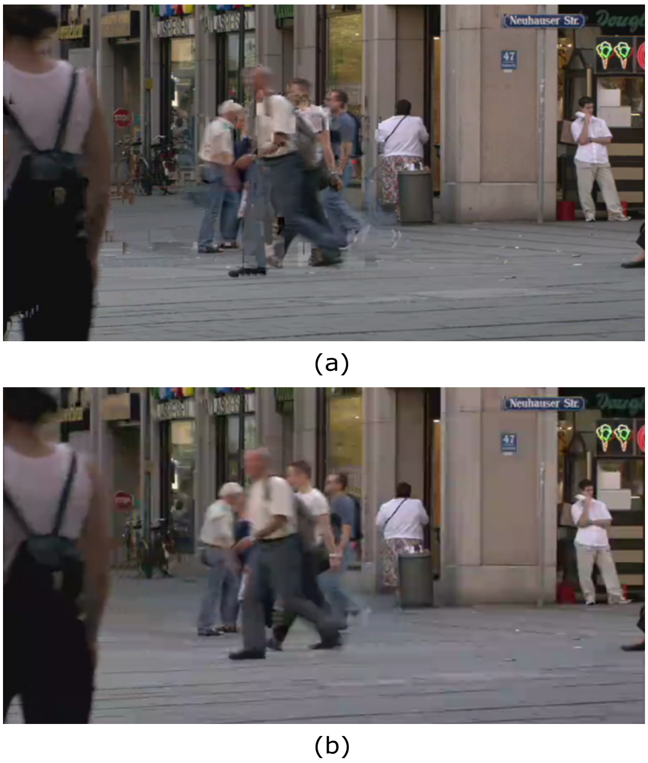
Fig. 1. Illustration of mean quality pooling method.

We hypothesize that if the regions affected by severe distortion are adaptively classified, the performance of quality estimation will be improved. It is hence essential to adaptively compute the overall score by considering the distribution of the quality scores.

3. DISTRIBUTION OF QUALITY SCORES

When local quality scores are obtained for a processed frame of a natural video and sorted in ascending order, it typically shows a saturating tendency in the direction of better quality, as shown in Fig. 3 for scores obtained using the SSIM index. This distribution of quality scores is characteristic of most natural videos that have undergone video compression. This tendency is due to the characteristics of natural videos and the characteristics of the distortion process, namely video compression. A typical natural video frame consists of large areas of smooth variations, with sharp edges and textures occurring between. The smooth variations in the image or video is composed of low frequency signals, and details such as edges and textures are composed of high frequency signals. Typical video coding schemes, such as discrete cosine transform (DCT) based compression schemes, quantize high frequencies in the image more severely than lower frequencies. Hence, encoding distortions are more severe in regions of high spatial activity such as edges, rather than in regions of smooth variation[8]. Consequently, the sorted quality scores tend to follow a saturation curve as depicted in Fig. 3, with relatively little quality degradation in larger regions of smooth variation and more severe quality impairments in smaller regions of high spatial activity. The distribution of quality scores in videos that suffer from distortions introduced due to lossy transmission of video depends on the nature of the lossy channel. However, typical lossy networks such as wireless networks drop packets that affect regions of a video frame, resulting in a distribution of quality scores that is similar to the one depicted in Fig. 3.

The phenomenon described above applies to spatial distribution of quality scores in a single intra-coded frame of the distorted video. A similar reasoning applies to predictively coded video frames due to the characteristics of natural videos and natural video distortions along the temporal dimension.


(a)

(b)

Fig. 2. Example of problem of mean as a pooling using LIVE datbase [4] : (a) The 70th frame of pa7_25fps (mean SSIM : 0.9497), (b) The 70th frame of pa11_25fps (mean SSIM : 0.9043).

Typical natural videos consist of large areas of static regions, interspersed with moving objects in the scene. Typical video compression algorithms utilize motion compensated DPCM technique across frames to achieve compression where static regions are encoded with zero motion vectors. Typical compression schemes produce large prediction errors around the borders of moving objects resulting in small regions of severe distortion [8][9]. Thus, predicted frames also suffer from small areas of severe distortion and larger areas of good quality.

One approach that can be taken is to divide the quality scores depicted in Fig. 3 into two regions using the form of the curve. The first region consists of the higher quality scores in the saturated region of the curve (from areas of the video that do not suffer from severe degradation). The second region consists of the non-saturated region of the curve (quality scores corresponding to regions of the video suffering from severe distortions). One contribution of our paper is that we describe a method to determine the two regions of the curve in an adaptive manner for each video frame, as opposed to using the fixed percentile pooling scheme proposed in [2]. Our classification of the saturated and non saturated regions takes into account the distribution of the quality scores. We perform this classification based on the slope of the curve. Based on a slope threshold, quality scores which are higher than the slope threshold are classified as belonging to the saturated quality region and quality scores below the threshold are classified as belonging to the non-saturated region of the curve. Finally, we hypothesize that the quality scores in the non-saturated region have far higher influence on the overall quality judgment by humans since human observers tend to be critically perceived poor quality regions [2]. Based on this hypothesis, we propose a new adaptive pooling scheme, IQpooling, to improve the performance of objective VQA algorithms.

4. THE PROPOSED SPATIAL AND TEMPORAL POOLINGS ACCOUNTING FOR EGOMOTION

Motion is a very important factor affecting the distribution of quality scores in a video. In particular, existence of egomotion has a significant effect on the distribution of quality scores. The distribution varies according to the existence of egomotion as illustrated in Fig.3. In a frame containing egomotion, the prediction error can occur over the entire frame due to global motion of the frame. This is reflected in the distribution of the quality scores, which contains a lot of intermediate scores between the saturated region and the non-saturated region in an ego-motion frame.

We propose a spatial pooling strategy that applies a different slope threshold t based on the existence of egomotion to classify quality scores into two regions: P_t that contains the set of unsaturated scores and P_t^c that contains the set of saturated scores. Let $f\left(z\right)$ represent a set of sorted quality scores obtained using a VQA algorithm on a given frame, where z indexes over the sorted set. The slope estimate the derivative on $f\left(z\right)$ according to

$$f'\left(z\right) \approx \frac{f\left(z+\Delta\right)-f\left(z\right)}{\Delta}\cdot\lambda, \tag{1}$$

The values of both the quality score and its argument are normalized to the same scale of $[0,1]$ by a normalization parameter, λ, which is the ratio of the number of samples per frame to the largest difference between the maximum and minimum quality scores (eg. for the SSIM index, the largest difference is 1). $f'\left(z\right)$ tends to be monotonically decreasing. Let $\widehat{t}$ be such that $f'\left(\widehat{t}\right) = t$, $f(x) \in P_t$ if $f(x) < f(\widehat{t})$ and $f(x) \in P_t^c$ if $f(x) \geq f(\widehat{t})$. A frame level quality index s_f is then computed.

$$s_f = \frac{\sum_{m\in P_t} Q_m + r\cdot\sum_{m\in {P^c}_t} Q_m}{|P_t| + r\cdot|{P^c}_t|}, \tag{2}$$

where Q_m is the m^{th} local quality score in the f^{th} frame and r is a small multiplier that is used to account for the reduced contribution of the scores in P_t^c to the overall quality of the video.

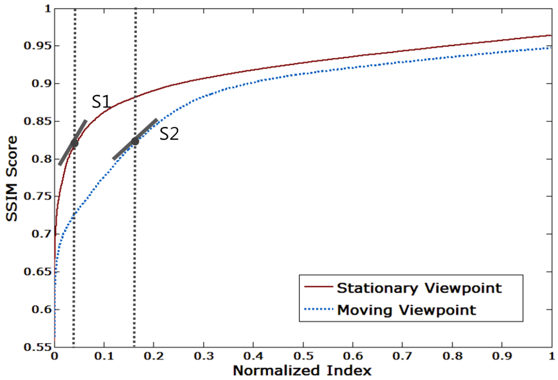


Fig. 3. Sorted quality scores of natural video showing a form of saturation curve.

For ego-motion frames, we simply utilize a slope of $t = 1$ where the x-increment and y-increment are the same. In a typical non-egomotion frame, large prediction errors are likely to be localized to regions of high spatial activity or regions containing moving objects. Hence, the saturated and non-saturated regions are quite distinct in this case and we utilize a slope of $t > 1$. To determine the presence of egomotion in a frame, scene movement is detected using the coefficient of variation (CoV) of motion vectors in a frame [7]. The CoV refers to the ratio of the standard deviation to the mean and a frame is classified as stationary when the CoV is lower than 1 and moving otherwise.

In VQA, to obtain the overall quality score for a video, the frame level quality scores s_f which are obtained by spatial pooling should be pooled along the temporal dimension. Temporal pooling also plays an important role in estimating perceived video quality accurately. We hypothesize that poor quality regions have an increased influence on the overall quality along the temporal dimension of video also. However, the distribution of the frame level quality scores s_f along the temporal dimension varies considerably with video content and distortion type. We classify frame level scores into two regions containing lower quality scores and higher quality scores, similar to spatial pooling, by k-means clustering [3] along the temporal dimension. The scores from the two regions are then combined to obtain the overall quality of the entire video sequence:

$$S = \frac{\sum_{f\in G} s_f + w\cdot\sum_{f\in G^c} s_f}{|G| + w\cdot|G^c|}, \tag{3}$$

where G contains quality scores from the lower quality region and G^c contains quality scores from the higher quality region. A weight w, computed as a function of the gap between the scores in G and G^c, is applied to scores in the higher quality region, where $w = \left|\frac{M_H - M_L}{\widehat{M}}\right|^2$, M_H and M_L are means of scores in the higher and lower quality regions respectively,

Table 1. SROCC results with several VQA algorithms in [4]. P1:Percentile pooling on quality map of SSIM, P2:Percentile pooling on quality map of MOVIE, I1:IQpooling on quality map of SSIM, I2:IQpooling on quality map of MOVIE. (W : Wireless, I : IP, H : H.264, M : MPEG2)

VQA	W	I	H	M	All
PSNR	0.4334	0.3206	0.4296	0.3588	0.3684
SSIM	0.5233	0.4550	0.6514	0.5545	0.5257
MOVIE	0.8019	0.7157	0.7664	0.7733	0.7890
P1	0.7696	0.7428	0.7032	0.6632	0.7659
P2	0.7992	0.7121	0.7386	0.7654	0.7650
I1	0.8141	0.7878	0.8116	0.8320	0.8368
I2	0.8086	0.8060	0.8285	0.8504	0.8441

Table 2. LCC results with several VQA algorithms in [4]. P1:Percentile pooling on quality map of SSIM, P2:Percentile pooling on quality map of MOVIE, I1:IQpooling on quality map of SSIM, I2:IQpooling on quality map of MOVIE. (W : Wireless, I : IP, H : H.264, M : MPEG2)

VQA	W	I	H	M	All
PSNR	0.4675	0.4108	0.4385	0.3856	0.4035
SSIM	0.5401	0.5119	0.6656	0.5491	0.5444
MOVIE	0.8386	0.7622	0.7902	0.7595	0.8116
P1	0.7954	0.7905	0.7339	0.6711	0.7829
P2	0.8174	0.7631	0.7479	0.7702	0.7946
I1	0.8420	0.8382	0.8271	0.8329	0.8516
I2	0.8521	0.7998	0.8438	0.8487	0.8603

and $\widehat{M}$ is the maximum score that is used to normalize w between 0 and 1.

5. PERFORMANCE AND CONCLUSION

We evaluated the IQpooling on the Laboratory for Image and Video Engineering (LIVE) Video Quality (VQ) database [4]. The IQpooling is applied on local quality estimates obtained using the SSIM index [6] and the MOVIE index [7] on the LIVE VQ database. A sampling window of 16×16 that slides by 4 pixels in each increment is utilized to obtain the SSIM quality map. λ is chosen to be the ratio of the number of samples per frame and a normalization factor that depends on the range of scores of the VQA algorithm. The normalization factor is chosen to be 1 for SSIM and 0.2 for temporal MOVIE and 0.35 for spatial MOVIE, respectively. The value of 10^{-4} is utilized for the scaling factor r in (2) although any sufficiently small value is sufficient as in [2]. The slope thresholds that are utilized for the stationary and moving viewpoints are 3 and 1, respectively, in both SSIM and MOVIE.

The Spearman rank order correlation coefficient (SROCC) and the Pearson linear correlation coefficient (LCC) are used as performance evaluation metrics and these are shown in Tables 1 and 2. Tables 1 and 2 clearly show that the IQpooling improves the performances of SSIM and MOVIE considerably.

In conclusion, we proposed a spatial and temporal pooling method, known as IQpooling, that better predicts overall video quality by considering the influence of spatially and temporally localized severe impairments on human judgment of quality. The worst quality scores are classified at the spatial pooling stage using a slope criterion and at the temporal pooling stage using k-means clustering. Furthermore, we explored the impact of egomotion on pooling quality scores. Applying an adaptive strategy based on the existence of egomotion, we obtained noticeable improvement in the performance of the VQA algorithms. In the future, we would like to apply the proposed pooling scheme on quality maps obtained using other VQA algorithms and other databases.

6. REFERENCES

[1] M. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality, *IEEE Trans. on Broad.,* vol. 50, no. 3, pp. 312-322, Sept. 2004.

[2] A. K. Moorthy and A. C. Bovik “Visual importance pooling for image quality assessment, *IEEE J. Special Topics in Signal ,* vol.3, no.2, pp.193-201, Apr. 2009.

[3] J. A. Hartigan, “Clustering Algorithms, Wiley, 1975.

[4] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of subjective and objective quality assessment of video,” *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, 2010.

[5] S. Lee, M. S. Pattichis and A. C. Bovik, “Foveated video quality assessment, *IEEE Trans. on Multimedia,* vol. 4, no. 1, pp. 129-132, Mar. 2002.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image Quality Assessment: From error visivility to structural similarity,” *IEEE Trans. Image Processing ,* vol. 13, no. 4, pp. 600-612, Apr. 2004.

[7] K. Seshadrinathan and A. C. Bovik, “Motion tuned spatio-temporal quality assessment of natural videos,” *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[8] M. Yuen and H. Wu, “A survey of hybrid mc/dpcm/dct video coding distortions,” *Signal Processing,* vol.70, no.3, pp.247-278, 1998.

[9] S. Lee, M.S. Pattichis and A.C. Bovik, “Foveated video compression with optimal rate control, *IEEE Trans. Image Process.,* vol. 10, no. 7, pp. 977-992, July 2001.