A NON-REFERENCE PERCEPTUAL QUALITY METRIC BASED ON VISUAL ATTENTION MODEL FOR VIDEOS

Fahad Fazal Elahi Guraya[1], Ali Shariq Imran[1], Yubing Tong[2], Faouzi Alaya Cheikh[1]

Gjovik University College, Gjovik, Norway[1], Universite de Saint Etienne, France[2]

ABSTRACT

Human Visual System (HVS) tends to focus on certain regions of an images or video frames. These regions form the saliency map, which could be used to better estimate the perceived quality of an image/video. In this paper, we propose a novel objective video quality metric based on these salient regions. This metric estimates the degree of blur and blockiness in each video frame from the impaired video only, and uses it with the saliency map to derive a weighting function. The latter is used to modulate the contribution of the pixel differences to the final quality score. The salient regions of the videos are automatically computed using our video saliency model. A psychophysical experiment is conducted to estimate the perceived quality of the impaired videos. The results of this subjective test are compared to the scores obtained with the proposed metric. The objective and subjective scores are found to be highly correlated, which shows that our metric correctly estimates the perceived quality of the video.

1. INTRODUCTION

Recent improvements in imaging and video technology allowed us to capture and record large collections of videos. When we need to share these videos on internet it is hard to share because of their big size. Therefore, it is always preferred to compress video for transmission. During the compression the estimation of the resulting video quality is an important factor. For several applications one may want to estimate the perceptual quality of the compressed video. For instance for video communication one would need a model to estimate video quality score to tune the parameters of the encoder. Similarly, in real time video surveillance system, a number of cameras may need to be controlled for proper functioning to ensure a certain level of quality of the recorded videos. This maybe useful to account for camera malfunction or to adjust it to the changes in the visual scene, such as changes in the illumination or weather conditions etc.. It may be of crucial importance to the surveillance application to have a certain quality of the recorded video for person identification or license plate reading for example. Transmission or storage of these surveillance videos also consume high bandwidth or take big storage space. Therefore, there is always a tradeoff between the video quality and the storage space or bandwidth usage. This is why the estimation of the perceived quality of visual media has become an active field of research lately. Especially of interest are the metrics that can give perceptual quality score based on human attention models [2-4].

There are three types of quality matrices i-e, a full-reference quality metric that takes original and degraded video and computes the difference or quality degradation; the second type of metric is non-reference quality metric, it computes the quality degradation based only on the impaired video and third type is called reduced reference quality metric, that computes certain features from the original and de-graded images, and finds correlation/match between them.

Under normal viewing conditions hum an eye movements are tightly couple to human visual attention [1]. It is known that humans direct attention to the important objects in a scene (image/video frame) using bottom-up and top-down cues [2-4]. Bottom-up cues use low-level features such as color, orientation, and intensity to compute the conspicuity maps. However top-down models use high level features such as face-detection, object/people detection etc. An attention model such as saliency maps could be computed automatically using top-down and bottom-up approaches [4]. In [4], authors used high level feature such as face detection with low level features such as color, intensity, orientation to compute the saliency maps for images and improved the results by 33%. Similarly saliency maps for videos can be computed by considering the temporal changes such as flickering and object motion along with stationary saliency maps. A full referenced quality metrics based on visual attention modeling for images and videos has been proposed in [9]. The authors have used saliency detection to improve PSNR and SSIM.

In the rest of this paper we will first discuss our perceptual model for saliency detection, in section 2. Our no-reference video quality metric is proposed in section 3. Section 4 presents the subjective psychophysical test and its results; we also compare the experimental results with the proposed quality metric and PSNR. The last section concludes the paper with some future directions.

2. MULTI-FEATURE PERCEPTION MODEL FOR SALIENCY DETECTION

Stationary saliency is computed by using multi-feature conspicuities including face and some low level features such as color intensity and orientation; motion saliency is calculated based on motion analysis and distance effect

on visual perception. Finally, stationary saliency map and motion saliency map are fused. Algorithms are included in the following three sections:

(1) Stationary saliency model with face as a high level feature and intensity, color, orientation as low-level features;
(2) Motion saliency model with motion vector field measurements and distance weights in Gaussian model;
(3) Fuse method for stationary saliency map and motion saliency map.

2.1. Multi-feature stationary saliency

Low level features such as intensity, color and orientation contribute much to our attention in Itti's bottom-up attention framework [2, 3]. Every feature is analyzed using Gaussian pyramids and multi-scales. Seven feature maps are generated including one intensity, four orientations (at 0, 45, 90,135 degree) and two color components (red/green and blue/yellow) conspicuity maps. After a normalization step, all those feature maps are summed to three conspicuity maps: intensity conspicuous map C_i, color conspicuity map C_c and orientation conspicuity map C_o.

One drawback of Itti's visual attention mechanism model is that its saliency map model is not well adapted for images with faces or other familiar objects that may attract attention. Psychological tests proved that face, head or hands can be perceived prior to any other details [5]. Several studies in face recognition have shown that skin hue features could be used to detect faces. To detect heads and hands in images, we have used the face recognition and location algorithm used by Walther et al [6]. Then stationary saliency based on multi-features conspicuities can be described as following,

$$S_S = f(S_{Itti}, S_{Face}) \tag{1}$$

Here we choose empirically the weights model as follows,

$$S_S = \frac{1}{8}\left(2C_i + 2C_c + C_o + 3C_F\right) \tag{2}$$

2.2. Motion saliency map and Gaussian weights model

Motion feature must be involved in video saliency map as it plays a very important role in video. With motion perception, we could know what is happening in the current scene. And some regions or objects are not so salient in video although they might be salient in images, for example, the rich texture of object in images will be omitted in videos with fast motion. In this paper, motion information of a video is analyzed with its motion vector field which can be calculated using motion estimation with more than one reference image. Here we used full searching and block matching to find the best motion vectors which are normally used in video compressing.

Based on the motion vector field, the intensity of motion vector, spatial coherence and temporal coherence of the motion are used to describe motion saliency map [7]. The intensity of motion vector is computed with the following equation,

$$I = \sqrt{(mvx)^2 + (mvy)^2} \tag{3}$$

Besides the intensity, the phase θ of motion vector will also be analyzed.

$$\theta = \left|\arctan(\frac{mvy}{mvx})\right| \tag{4}$$

θ distributes in [0, 360] after normalization.

Within the motion vector field, motion vector of the blocks will also be analyzed. The distribution probability density ρ_i is computed using the histogram distribution of θ value within the neighbourhood with size of k x k, here 7x7 is used.

The spatial motion saliency *Cs* is calculated as following,

$$Cs = -\sum_{i=1}^{N} \rho_i \cdot \lg \rho_i \tag{5}$$

N is the number of histogram bins of θ value in *k*x*k* field.

Temporal saliency map *Ct* can be defined in the same way. Then motion saliency map is computed in the following equation

$$S_M = I \cdot Ct(1 - I \cdot Cs) \tag{6}$$

I is motion intensity by computing the magnitude of motion vector, *Ct* is the temporal coherence inductor and *Cs* is the spatial coherency based on spatial phase histogram and temporal phase histogram statistic and analysis.

2.3. Fusion model of stationary and motion saliency map

The stationary saliency map S_S and motion saliency map S_M can be fussed to obtain the final saliency map of every frame in a video. Since we usually are more susceptible to those objects in the centre of the frame than that is far away from the center (x_c, y_c), we propose a distance weighting fusing model as following,

$$S_{V\_G} = \alpha \cdot S_M \cdot w_{i\_mb} + (1-\alpha)\cdot S_s \tag{7}$$

$$w_{i\_mb} = e^{-d/8} \tag{8}$$

$$d = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \tag{9}$$

$$x_c = \frac{width\_mb}{2}; y_c = \frac{height\_mb}{2} \tag{10}$$

where *width_mb* and *height_mb* is the height and width in mb (i.e.16x16).

2.4. Saliency detection for videos

The proposed saliency detection model is used to detect the salient regions in the video frames. The salient regions are used for two purposes. One is to add various kinds of artifacts, as explained in the next section, to the salient regions/non-salient regions of the video frame. Second purpose is to use saliency maps to compute the quality score for each video, where the salient regions are assigned higher weights than the non-salient regions. The original and impaired video frames of video 1 are shown in figure 1.

Figure 1. (a) Original video frame (b) Blur in full frame (c) Blur in non-salient regions (d) Blur in salient regions

3. THE PROPOSED QUALITY METRIC

Most of today's compression standard such as MPEG use 8x8 block size for DCT compression. This causes various artifacts to appear in the compressed videos. Two of the most common artifacts are blocking and blurring. They degrade the video quality drastically. We have proposed a quality metric that detects these blocking and blurring artifacts of video frame and computes the quality score for the overall video.

The saliency map computed in section 2 is used as a mask on the impaired video. Using these saliency maps, weighted video frames are created for the distorted video sequence. The pixels which are more salient are given higher weights than the less salient neighboring pixels, while those regions which are non-salient do not contribute to the weighted frame.

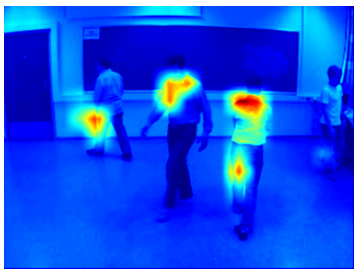

Figure 2. Salient Region in a video frame

The blocking and blurring artifacts are then estimated on the weighted video. The extrapolated difference between the adjacent blocks constituting the salient regions is used to calculate the blockness effect on the perceived image quality.

Let $I_b[row, col, t] = \big[I[row * 8 + i, col * 9 + j, t]\big]$, $(i, j) \in [0 \dots 7]$ be the luminance block sequence situated at time t. The value of the blocking artifact across two adjacent blocks $I_b[row, col, t]$ and $I_{b+1}[row + 1, col, t]$is the discontinuity at the frontier. The value is evaluated for all lines within a block with extrapolated values of the neighboring pixels. The left and right extrapolated values across the boundaries of two adjacent blocks can be calculated as

$$El_b(j) = \tfrac{3}{2}\,[row * 8 + 7, col * 8 + j, t] - \tfrac{1}{2}\,[row * 8 + 6, col * 8 + j, t] \quad (11)$$

$$Er_b(j) = \tfrac{3}{2}\,[row * 8 + 0, col * 8 + j, t] - \tfrac{1}{2}\,[row * 8 + 1, col * 8 + j, t] \quad (12)$$

While the vertical artifact value is the mean of the eight discontinuities within a single block (b).

$$Bv_{(r+1,c+1)} = \tfrac{1}{8}\textstyle\sum_{j=0}^{7} |(El_b(j)) - (Er_b(j))| \quad (13)$$

where Bv give us the blocking artifact value across vertical blocks. The values for the horizontal artifacts can be calculated in similar fashion. A blocking score can then be computed by summing up the vertical and horizontal blocking features.

$$B_k = Bv + Bh \quad (14)$$

Blur on the other hand is hard to compute. It is usually caused by the quantization process and often by the de-blocking filter. The high frequency information is associated with the detail of an image. Quantization process removes such high frequency information detail from an image that results in a blurred image.

Blur can be calculated across the horizontal and vertical boundaries of the 8x8 adjacent blocks. Local variance [8] is used to estimate the blurriness in an image constituting the salient blocks. We first compute the local variance across the vertical blocks and then the horizontal. The local variance is given by:

$$Lvar = \frac{\sqrt{\sum_{i=1}^{X}(pi - pi')}}{X-1} \quad (15)$$

where $(pi - pi')$ is the difference between the pixels across the boundary edges of the blocks. The local blur is the average of the local variance across two blocks

$$B_l = mean\,|Lvar_b - Lvar_{b+1}| \quad (16)$$

The sum of total blur across vertical blocks is:

$$B_v = \textstyle\sum_{i=i}^{b} Bl_i \quad (17)$$

where $\boldsymbol{b}$ is total number of blocks

The values for the horizontal blur can be calculated in similar fashion. An overall image quality value is obtained by combining the features extracted from the dataset. First the average blocking and blurring values are obtained by combining the vertical and horizontal artifacts.

$$B_k = (\frac{Bv+Bh}{2}),\ B_r = (\frac{Bv+Bh}{2}) \quad (18)$$

Then the following prediction model is used to combine the artifacts

$$QpM = 10 * (\propto + \delta * {B_k}^{p1} * {B_r}^{p2}) T^{p3} \quad (19)$$

where $\propto$,δ, *P1*, *P2* and *P3* are adjusted based on the opinion score obtained from the subjective tests results.

4. EXPERIMENTAL TESTS AND RESULTS

Subjective tests were conducted on a 21 inch flat panel, with monitor white point set to D65. The screen light intensity was set to 376.8 lux. The ambient light intensity was set to 200 lux. A test application is created in Matlab which first shows the reference videos followed by the impaired videos. Single stimuli quality rating scale was used. Users rate the video once it is played. And then the program loads the new video with different level of impairments. Figure 3 shows a frame from a sample video used in the subjective experiment, (a) shows the original frame and (b) shows the impaired video frame with blur of size 7x7 kernel.

Figure 3. Sample video sequence-2 (a)original video frame (b) Blur in non-salient region

A total of 90 impaired videos are created by adding blur, compression, blur + compression artifacts in salient, non salient regions and in full frame of the two video sequences. These impaired videos along with two original were shown to the subjects. The original videos were shown at the start while the impaired videos were shown in random fashion. Sixteen non-expert subjects participated in the subjective experiment. At the end of each impaired video they are asked to rate the quality on a scale of 1-5. Where 1 corresponds to really annoying, 2 to annoying, 3 to slightly annoying, 4 to perceptible but not annoying and 5 corresponds to the imperceptible video quality. Mean opinion score (MOS) was then obtained which was used to correlate with the objective score obtained from the quality prediction metric. The results are then compared with the PSNR.

Table 1: Correlation results for PSNR and NR-QpM(Proposed method)

	Video Sequence 1	**Video Sequence 2**	**Video Sequence 1&2**
PSNR	**0.709**	**0.739**	**0.714**
NR-QpM	**0.81**	**0.781**	**0.8**

The results show very good correlation with our proposed method as can be seen from table 1.

5. CONCLUSIONS AND FUTURE DIRECTIONS

We have proposed an objective perceptual quality evaluation metric for video that estimates blockness and blur. Saliency has been introduced to incorporate the HVS. The results obtained with our quality metric show high correlation with the subjective MOS obtained from the psychophysical experiment. The results are also compared with those of PSNR. Our proposed metric is non-reference, which makes it suitable for applications such as video streaming or surveillance videos quality evaluation. More tests with different types of videos and impairments will be performed to make the metric more generic.

6. REFERENCES

[1] T. Jost, N. Ouerhani, R. V. Wartburg, R. Muri, and H. Hugli, Computer Vision and Image Understanding, Elsevier 100,107 (2005).

[2] L. Itti, Ph.D. thesis, California Institute of Technology, Pasadena, California (2000).

[3] L. Itti and C. Koch, Neuroscience 2001 2(3), 194 (2001).

[4] P. Sharma, F. A. Cheikh, and J. Y. Hardeberg, in Sixteenth Color Imaging Conference (The Society for Imaging Science and Technology, 2008), vol. 16, pp. 332-337.

[5] R Desimone, TD Albright, CG Gross and C Bruce. "Stimulus selective properties of inferior temporal neurons in the macaque", Journal of Neuroscience, vol4, 2051-2062, 1984.

[6] Walther, D., Koch, "Modeling Attention to Salient Proto-objects", Neural Networks 19, 1395–1407, 2006

[7] Yufei Ma, Hongjing Zhang. A model of motion attention for video skimming. ICIP 2002.

[8] Liu Debing; Chen Zhibo; Ma Huadong; Xu Feng; Gu Xiaodong, "No Reference Block Based Blur Detection," *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on* , vol., no., pp.75-80, 29-31 July 2009

[9] You, J., Perkis, A., Hannuksela, M. M., and Gabbouj, M. 2009. Perceptual quality assessment based on visual attention analysis. In *Proceedings of the Seventeen ACM international Conference on Multimedia* (Beijing, China, October 19 - 24, 2009). MM '09. ACM, New York, NY, 561-564. DOI= http://doi.acm.org/10.1145/1631272.1631356