Video Processing for Human Perception Oriented Coding

*오형석 **김원하

경희대학교

*ogun5@khu.ac.kr **wonha@khu.ac.kr

Video Processing for Human Perception Oriented Coding

*Oh, Hyung Suk **Kim, Wonha

Kyung Hee University

요약

This paper presents human perception-based video coding method using an online learning framework. In this work, we analyze the relationship between human attention regions and video quality, and also consider human memory. We classify the motion patterns based on the analysis. Then, we devise a motion pattern classification method using Hedge algorithm. Along with the motion patterns, we smooth out the specific regions or sharpen details of the regions using the regional priorities. The preprocessed sequences are applied to the video codec. The performance is excellent on the overall quality as well as the regional quality.

1. Introduction

The currently prevailing video codecs are designed to allocate more bits to areas having high texture and high motion activities for achieving the rate-distortion optimization. So, these video codecs sustain the overall best video quality at a given bandwidth. The video quality produced by the codecs is numerically assessed using PSNR that measures distortions. However, the subjective quality indexes tested in Video Quality Experts Group (VQEG) quite often show that the PSNR value does not always meet the video quality as human actually perceives [1]. This is because visual quality to human is affected by regional priorities determined from regional motion variations, textures, motion history and so on. Therefore, along with the overall video quality, regional priorities for the visual quality is also important to human perception. Thus, it has become necessary to develop video coding method satisfied with overall video quality and also optimized with prioritized regional visual quality.

For deciding regional priorities, visual saliency has been recently developed in the literature. Exemplary researches on motion saliency are conducted by [2] and [3]. Jacobson et al. computes mutual information of the motion feature distribution between a scanning patch and its surroundings [2]. Ma et al. calculated the entropy of motion variations of regions and assign the higher saliency to the regions with the higher entropy[3].

Under the hypothesis that the better texture quality at the more salient regions could provide better visual quality at a given temporal resolution, the saliency-based video coding methods allocate more bandwidth to higher prioritized salient regions by either removing high frequency components in less prioritized regions or adjusting the quantization parameters (QP) for each region priority [4].

It should be observed that the saliency does not always meet the degree of regional texture quality. The capacity of which human or video cameras can acquire space and time resolution for a unit time is inherently limited. Thus, the texture signals of regions with high motion activities are easily motion blurred. Therefore, although the regions with large motion activities often draw high human attention, the regions do not necessarily bear good texture qualities.

Besides the limited capacity of space-time resolution, texture quality that human perceives correlates with motion history memorized in human's memory. At the moment when a region, in which human had kept being interested due to its motion variation, stops motion activities, human tends to sustain momentarily the interest of the same region since

human revives the motion history in human memory. This implies that the region where more bandwidth had been assigned due to its interesting motion behavior should sustain the same bandwidth for a moment even after the motion of the region stops.

In this work, we propose an human perception based video coding method that integrates the effects of the space-time resolution, the motion history and the saliency. We develop the methodology for classifying motion behavior into moving patterns reflecting human's limited capacity and the short history of motion in human memory. Human continuously constructs and updates the moving patterns. So, the patterns should be established in an on-line learning framework. We exploit Hedge algorithm for the on-line learning method. We then apply a saliency model developed in [3] under consideration of moving patterns. At specific moving patterns, we sharpen characteristics of the original image in accordance with the regional priority. The performance is excellent on the overall quality.

2. Human Perception based Moving Pattern

In this section, we classify the moving patterns in accordance with human perception. We consider that human inherently has the limited capacity in space and time resolution correlated with motion blur, and that human consecutively update the memorized moving patterns with motion behavior. Thus, we categorize the moving patterns into three patterns: normal mode, fast mode and potential mode.

Among the regions in motion, we classify the regions into the normal moving regions and the fast moving regions. Since human visual system (HVS) or video cameras have limited space-time resolution, the space resolution of a region is more reduced as the region moves faster. This implies that the texture signals of fast moving regions are blurred. At the region with higher texture energy, the blurring by motion is more easily occurred. The normal moving regions are not blurred by motion. Oppositely, the motion blur occurs in the fast moving regions.

The simple and efficient method for detecting the regions with motion blurs is to measure the ratio between the region's texture energy and the region's temporal residual energy. The temporal residual energy means the difference energy between a region and its motion compensated region. The regions with motion blur are not well recognized since the motion blur smooth out the region's texture energy. Therefore, in the regions with motion blur, the texture energy becomes small compared to the temporal residual energy. The fast mode appears at the regions with motion blur and the normal mode does at the regions without motion blur.

Human perceives a region by reviving the memorized moving patterns. So, in the currently non-moving regions where were previously in motion, human momentarily recognizes that the region is still in motion and conceives the moving pattern of the region as its previous pattern in memory. If the motion in the region is not detected before the moving pattern in memory disappears, the region becomes background. The potential mode is the mode of non-moving regions before the pattern memory disappears. Thus, human recognizes that the region with potential mode continually keeps moving although no motion occurs at the region.

3. Moving Pattern Determination

3.1 Block motion state

Let B_k^t be the k^{th} NxN block in the t^{th} frame. MV_k^t and s_k^t denote the motion vector and the motion state of B_k^t, respectively. We classify the motion states into normal state(NS), fast state(FS) and stationary state(SS). The blocks of normal state are in motion without motion blur. The blocks of fast state are blurred by motion. The blocks of stationary state are static.

Considering that the texture energy of the region with motion blur is smaller than its temporal residual energy, occurrence of the motion blur is detected by the ratio between the texture energy and the temporal residual energy. The energy ratio R_k^t is calculated as

$$R_k^t = \frac{\frac{1}{N\times N}\left(\sum_{q\in B_k^t}\left\{B_k^t(q)-B_k^t(q+MV_k^t)\right\}\right)^2}{\left(\frac{1}{N\times N}\sum_{q\in B_k^t}\left\{B_k^t(q)\right\}^2\right)-\left(\frac{1}{N\times N}\sum_{q\in B_k^t}\left\{B_k^t(q)\right\}\right)^2} \quad (1)$$

where q is the inner pixel of B_k^t.

Using the ratio and the motion vector, the motion state s_k^t is represented as follows.

$$s_k^t = \begin{cases} Normal\ state(NS), & \text{if } MV_k^t \neq 0, R_k^t \leq 1 \\ Fast\ state(FS), & \text{if } MV_k^t \neq 0, R_k^t > 1 \\ Stationary\ state(SS), & MV_k^t = 0 \end{cases} \quad (2)$$

3.2 Hedge based moving pattern construction

By learning the patterns of motion states observed for a certain duration, we can continuously renew the probabilities of each motion state. Thus, we need a learning algorithm. We adopt Hedge algorithm as the learning algorithm. Hedge estimates the probabilities that each motion state occurs.

We define the nonoccurrence vector l_k^t for quantifying the nonoccurrence of the motion states at each time and also define the record vector L_k^t for recording history of motion states observed as

$$l_k^t = \begin{cases} [0,1,1], \text{ if } s_k^t = NS \\ [1,0,1], \text{ if } s_k^t = FS \\ [1,1,0], \text{ if } s_k^t = SS \end{cases}, \quad L_k^t = \sum_{r=t-T}^{t} l_k^r \qquad (3)$$

where T is the size of temporal window. Therefore $L_k^t(0)$, $L_k^t(1)$ and $L_k^t(2)$ are the number of nonoccurrences of the normal state, the fast state and the stationary state, respectively.

Let $w_k^t = [w_{0,k}^t, w_{1,k}^t, w_{2,k}^t]$ be the weight vector of the k^{th} block at time t. Hedge algorithm updates the weights in the following way:

$$\boldsymbol{w}_k^t = \begin{bmatrix} w_{0,k}^t \\ w_{1,k}^t \\ w_{2,k}^t \end{bmatrix} = \begin{bmatrix} w_{0,k}^0 \cdot \exp^{-\eta L_0^t(0)} \\ w_{1,k}^0 \cdot \exp^{-\eta L_0^t(1)} \\ w_{2,k}^0 \cdot \exp^{-\eta L_0^t(2)} \end{bmatrix} \qquad (4)$$

The probability vector p_k^t of the block is

$$\boldsymbol{p}_k^t = \begin{bmatrix} p_{0,k}^t \\ p_{1,k}^t \\ p_{2,k}^t \end{bmatrix} = \frac{1}{|\boldsymbol{w}_k^t|} \cdot \begin{bmatrix} w_{0,k}^t \\ w_{1,k}^t \\ w_{2,k}^t \end{bmatrix} \qquad (5)$$

where $p_{0,k}^t, p_{1,k}^t, p_{2,k}^t$ are the probabilities for the normal state, the fast state and the stationary state, respectively.

From the highest probability among the motion states and the current motion state, we construct the moving pattern BMP_k^t as below.

$$BMP_k^t = \begin{cases} \text{NM if } \max\{p_{i,k}^t\} = p_{0,k}^t, s_k^t \neq SS \\ \text{FM if } \max\{p_{i,k}^t\} = p_{1,k}^t, s_k^t \neq SS \\ \text{SM if } \max\{p_{i,k}^t\} = p_{2,k}^t, s_k^t \neq SS \\ \text{PM if } \max\{p_{i,k}^t\} \neq p_{2,k}^t, s_k^t = SS \end{cases} \qquad (6)$$

4. Application to Video Coding

The texture signals of the fast moving objects are already blurred and so the detailed texture signals of such objects are degenerated to noisy signals even though the regions have high human attentions. Therefore, we apply the strong low-pass filtering to the fast moving objects. For performing the low-pass filtering, we use Gaussian kernel.

The texture quality at objects with normal moving or potential moving pattern would be decided how much such objects draw human attentions. So, the degree of the applied high-pass filtering is more stronger for the object with higher saliency. The saliency of the objects with potential moving pattern momentarily keeps previous saliency since human revives the motion history in short term memory.

For applying the saliency model in the objects with the normal moving and potential moving patterns, we exploit motion saliency developed in [3]. This model counts the motion intensity of a region and the entropy of motion uniformity and variation in the region. Along with the saliency index, we perform stronger unsharp masking for sharpening the details. Therefore, the resultant sequence is processed by the unsharp masking in the regions with higher texture quality and higher human attention, and by the Gaussian kernel in the regions with less texture quality.

5. Experiment

For evaluating the overall video quality, we use the structural similarity index (SSIM) since PSNR does not always reflect the video quality that human recognizes. SSIM index is measured based on how identically human perceives the compared images. As SSIM score draws closer to 1, human regards the compared images as the same images. An SSIM score of 1.0 implies that the compared images are identical.

Table 1 shows the PSNR and the SSIM scores at variable target bit rates. The PSNR values for the video sequences preprocessed by the proposed method are either same or lower compared to the video sequences decoded without preprocessing. On the other hand, the differences of the SSIM scores between the video sequences decoded without preprocessing and the decoded versions of the video sequences preprocessed by the proposed method are very close to 0. It indicates that human identically perceives the compared sequences even though the PSNR values are lower than the video sequences decoded without preprocessing. So, we can know that the proposed method sustains the overall video quality that human recognizes.

6. Conclusion

<Table 1> PSNR and the structural similarity index(SSIM) scores. The presented PSNR and SSIM scores are the average score calculated at each frame of the test video sequences.

Seq.	Measure	Method	Target Bit rates(bps)		
			200K	400K	600K
Paris	PSNR	Original	29.9	33.8	36.1
		Proposed	28.5	31.0	32.0
	SSIM	Original	0.83	0.94	0.96
		Proposed	0.82	0.92	0.95
Ice	PSNR	Original	34.8	38.7	40.9
		Proposed	34.7	38.7	40.8
	SSIM	Original	0.95	0.97	0.98
		Proposed	0.95	0.97	0.98

In this work, a novel method for HVS-oriented motion saliency has been proposed. The proposed method has exhibited excellent performance in the challenging task of determining human's interest regions in videos. Therefore, the proposed algorithm is well-suited as a preprocessing step for applications, such as mobile videos and video conferencing.

Acknowledgement

This work was supported by grant No. 2011-0003804 from National Research Foundation of Korea.

Reference

[1] Final report from the video quality experts group on the validation of objective quality metircs for video quality assessment.

[2] N. Jacobson, Y. Lee, V. Mahadevan, N. Vasconcelos, and T. Q. Nguyen, “A novel approach to FRUC using discriminant saliency and frame segmentation”, IEEE Trans. On Image processing, vol. 19, no. 11, pp. 2924 - 2934, Nov. 2010.

[3] Y. Ma, and H. J. Zhang, “A model of motion attention for video skimming”, IEEE ICIP, pp. 129 - 132, Sept. 2002.

[4] Z. Wang, and L. Lu, and A. C. Bovik, “Foveation scalable video coding with automatic fixation selection”, IEEE Trans. on Image Processing, vol. 12, no. 2, pp. 243-254, Feb. 2003.