Image and Vision Computing 29 (2011) 1–14

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Visual attention guided bit allocation in video compression

Zhicheng Li [a,b], Shiyin Qin [a], Laurent Itti [b,*]

[a] School of Automation Science and Electrical Engineering, Beihang University, Beijing, China
[b] Computer Science Department, University of Southern California, Los Angeles, CA, USA

ARTICLE INFO

Article history:
Received 2 November 2009
Received in revised form 13 May 2010
Accepted 12 July 2010

Keywords:
Visual attention
Video compression
Eye-tracking
Video subjective quality

ABSTRACT

A visual attention-based bit allocation strategy for video compression is proposed. Saliency-based attention prediction is used to detect interesting regions in video. From the top salient locations from the computed saliency map, a guidance map is generated to guide the bit allocation strategy through a new constrained global optimization approach, which can be solved in a closed form and independently of video frame content. Fifty video sequences (300 frames each) and eye-tracking data from 14 subjects were collected to evaluate both the accuracy of the attention prediction model and the subjective quality of the encoded video. Results show that the area under the curve of the guidance map is 0.773 ± 0.002, significantly above chance (0.500). Using a new eye-tracking-weighted PSNR (EWPSNR) measure of subjective quality, more than 90% of the encoded video clips with the proposed method achieve better subjective quality compared to standard encoding with matched bit rate. The improvement in EWPSNR is up to over 2 dB and on average 0.79 dB.

1. Introduction

Significant improvements in video coding efficiency have been achieved with modern hybrid video coding methods such as H.264/AVC [1,2] in the last two decades. Spatial and temporal redundancy in video sequences has been dramatically decreased by introducing intensive spatial–temporal prediction, transform coding, and entropy coding. However, to achieve better compression performance, reducing such kind of so-called objective redundancy is limited and highly complex in computation.

On the other hand, research on human visual characteristics shows that people only perceive clearly a small region of 2–5° of visual angle. The human retina possesses a non-uniform spatial resolution of photoreceptors, with highest density on that part of the retina aligned with the visual axis (the fovea), and the resolution around the fovea decreases logarithmically with eccentricity [3]. What's more, research results show that observers' scanpaths are similar, and predictable to some extent [3]. These research results provide a new pathway to compress images/videos based on human visual characteristics: only encode a small number of well selected interesting regions (attention regions) with high priority to keep a high subjective quality, while treating less interesting regions with low priority to save bits.

Recently, many subjective quality-based video coding methods have been developed. According to the way of obtaining attention regions, they can be coarsely classified into four categories, as follows: (1) In the first approach, considering that human attention prediction is still an open problem, human–machine interaction methods are adopted to obtain the attention regions. One example of online human–machine interactive methods is gaze-contingent video transmission, which uses an eye-tracking device to record eye position from a human observer on the receiving end and applies in real-time a foveation filter to the video contents at the source [4–8]. This approach is particularly effective because observers usually do not notice any degradation of the received frames, since high-quality encoding continuously follows the high-acuity central region of the observers' foveas. However, this application is restricted to specific cases where an eye-tracking apparatus is available at the receiving end. For general-purpose video compression, this approach faces severe limitations if an eye-tracker is not available or several viewers may watch a video stream simultaneously. To address this, offline interactive methods are designed to obtain the interesting regions by asking subjects to manually draw regions which are interesting, and then applying this to the encoding procedure [9]. (2) The second class of approaches uses machine vision algorithms to automatically detect interesting regions. For instance, due to the importance of human faces while people perceive the world [10,11], it is reasonable to consider that human faces may likely constitute interesting regions. In [12–14], face regions are thus defined as the regions-of-interest. Face detection and tracking methods are explored to keep the interesting regions focused onto human faces, and more resources are allocated during encoding to these face regions, to keep these regions in high quality. With the development of face detection algorithms and object tracking methods in machine vision, this kind of video compression is very effective in the occasions where human faces indeed are central to the visual understanding of a video sequence, such as for video

* Corresponding author. University of Southern California - Hedco Neurosciences Building, room HNB-07A - 3641 Watt Way, Loa Angeles, CA 90089-2520 - USA. Tel./fax: +1 (213) 740 3527/5687.
E-mail address: itti@pollux.usc.edu (L. Itti).

0262-8856/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2010.07.001

telephone or video conference. However, this type of approach is obviously only workable when human faces are present. For unconstrained video compression where there may or may not be faces in the streams to be encoded, this method will fail to find interesting regions. (3) The third class of approaches uses knowledge about human psychophysics to guide the encoding process. For example, research results show that the human visual system (HVS) can tolerate certain amounts of noise (distortion) depending on its sensitivity to the source and type of noise for a given region in a given frame. Under certain conditions, the HVS can tolerate more distortion than the objective distortion measurements such as mean square error (MSE) would predict; on the other hand, there are some types of distortions which, despite low MSE, are vividly perceived and impair the viewing experience [15–17]. Based on this theory, many image/video encoding techniques have sought to optimize perceptual rather than objective (MSE) quality: these techniques allocate more bits to the image areas where human can easily see coding distortions, and allocate fewer bits to the areas where coding distortions are less noticeable. Experimental subjective quality assessment results show that visual artifacts can be reduced through this approach; however, there are two problems: one is that the mechanisms of human perceptual sensitivity are still not fully understood, especially as captured by computational models; the other is that perceptual sensitivity may not necessarily explain people's attention. For example, smoothly textured regions and objects with regular motions often belong to the background of a scene and do not necessarily catch people's attention, but these types of regions are highly perceptually sensitive if attended to. (4) The fourth class of approaches exploits recent computational neuroscience models to predict which regions in video streams are more likely to attract human attention and to be gazed at. With the development of brain and human vision science, progress has been made in understanding visual selective attention in a plausible biological way, and several computational attention models have been proposed [18–20]. In these models, low-level features such as orientation, intensity, motion, etc. are first extracted, and then through nonlinear biologically inspired combination of these features, an attention map (usually called saliency map) can be generated. In this map, the interesting locations are highlighted and the intensity value of the map represents the attention importance. Under the guidance of the attention map, resource can be allocated non-uniformly to improve the subjective quality or save the bandwidth [21–24]. Although such research shows promising results, it is still not a completely resolved problem.

Once interesting regions are extracted, a number of strategies have been proposed to modulate compression and encoding quality of interesting vs. uninteresting regions [21,25–29]. One straightforward approach is to reduce the information in the input frames. In [4,21,22], the frames to be encoded are first blurred (foveated) according to the attention map. The foveated image only keeps the attention regions in high quality while the other regions are all blurred. Through the blurring, redundancy is reduced significantly, and the compression ratio can be several times higher than the normal encoding method. However, blurring yields obvious degradation of subjective quality in the low saliency regions. In [23], a bit allocation scheme through tuning the quantization parameter is proposed with a constrained global optimization approach. Results show that 60% of the test video sequences encoded by this approach have better subjective visual quality compared to the video encoded by the normal method under the same bandwidth. In rate-distortion optimization, different mode may get different video quality and bit rate. The mode decision is usually determined by minimize the cost function which is the sum of encode error and bit rate multiple by a parameter (called Lagrange multiplier). Considering that the Lagrange multiplier will affect the mode decision in rate-distortion optimization, a Lagrange multiplier adjustment method is explored in [25]. An optimized rate control algorithm with foveated video is proposed in [26], and foveal peak signal-to-noise ratio (FPSNR) is introduced as subjective quality assessment. In [28], a region-of-interest based resource allocation method is proposed, in which the quantization parameter, mode decision, number of referenced frames, accuracy of motion vectors, and search range of motion estimation are adaptively adjusted at the macroblock (MB) level according to the relative importance (obtained from the attention map) of each MB.

How to evaluate the quality of a compressed image/video is still an open problem. Many quality assessment metrics have been developed to evaluate the objective or subjective quality of video. Among them, MSE and PSNR are two widely adopted objective quality measurements, even though they often are not consistent with human perception. Many additional types of objective (including human vision-based objective) quality assessment methods have been proposed [26,30–32]. However, the research results of the video quality experts group (VQEG) show that there is no objective measurement which can reflect the subjective quality in all conditions [33]. The suggested subjective quality from VQEG was obtained by using the mean opinion score (MOS) from pool of human subjects. Specifically, subjective quality scales ranging between excellent, good, fair, poor and bad (weight values are 5, 4, 3, 2, and 1, respectively) can be obtained from naive observers, and the weighted mean MOS score can be used as the subjective quality.

In this paper, we use a neurobiological model of visual attention, which automatically selects (predicts) high saliency regions in unconstrained input frames to generate a saliency map (SM). Considering the human's foveated retina characteristic, a guidance map (GM) is generated by finding the top salient locations in the saliency map. The GM is then used to guide the bit allocation in video coding through tuning the quantization parameters in a constrained optimization method. The overview of the proposed method can be seen in Fig. 1. For experimental validation, 50 high-definition (1920×1080) video sequences were captured using a raw uncompressed video camera, which include scenes at a library, pool, road traffic, gardens, a dinner hall, lab rooms, etc. Instead of using a subjective rating method, an eye-tracking experiment which records human subjects' eye fixation positions over the video frames was conducted to validate both the attention prediction model and the compressed video subjective quality. The focus of this paper is to combine the attention model with the latest video compression framework, and to validate the result in a quantitative way through an eye-tracking approach. The experiment results show that the proposed method is effective in both predicting human attention regions and improving subjective video quality while keeping the same bit rate.

The present paper complements our previous work [21], in which we showed that a saliency map model can predict human gaze well above chance, and can be used to guide video compression through selective blurring of low-salience image regions. The key innovation in the present work is to replace the selective blurring step, which yields quite obvious distortions in low-salience video regions, with a more sophisticated and more subtle localized modulation of the H.264 encoding parameters. Our new algorithm employs a constrained global optimization approach to derive the encoding parameters at every location in every video frame. We find that the optimization can be solved in closed form, which gives rise to an efficient implementation. This new optimization approach is an important step as it yields encoded videos that subjectively look very natural and are not degraded by blurring. Further, we develop and test a new eye-tracking weighted PSNR (EWPSNR) measure of subjective quality. Using this measure, we find that videos compressed with the proposed technique have better EWPSNR on our test video clips. Because our proposed method is purely algorithmic, requires no human intervention or parameter tuning, is applicable to a wide variety of video scenes, and yields improved EWPSNR, we suggest that it could be integrated to future generations of general-purpose video codecs.