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.

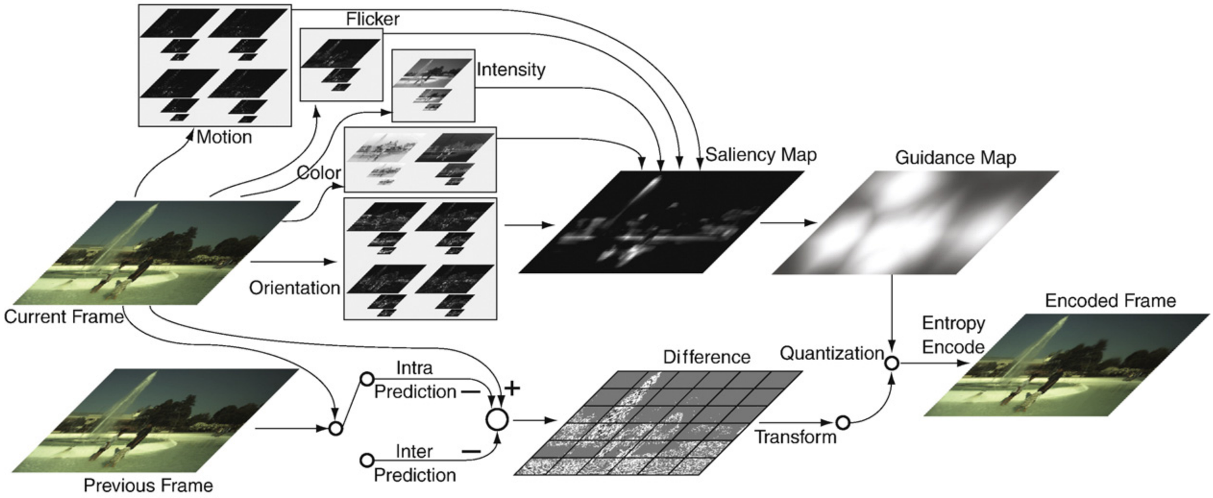


Fig. 1. Overview of the model. For the attention model path, the current input frame is first decomposed into multi-scale analysis with channels sensitive to low-level visual features (two color contrasts: blue–yellow, red–green; temporal flicker; intensity contrast; four orientations, 0°, 45°, 90°, and 135°; and for directional motion energies, up, right, down, and left). The saliency map is obtained after within-channel within-scales and cross-scales nonlinear competition. Assuming that the top salient locations in the saliency map are likely to attract attention and gaze of viewers, a guidance map is generated by foveating these positions. On the compression path, the current macroblocks (MBs) are predicted by previous encoded frame MBs through intra (which means the prediction result is generated from the current frame) or inter (which means the prediction result is generated from the previous frame) mode. The prediction error (difference) is then passed through transform and quantization; here the generated guidance map is used to adjust the quantization parameters to realize the non-uniform bit allocation. An encoded frame is complete after quantization and entropy encoding.

2. Method

2.1. Attention model

The model computes a topographic saliency map which indicates how conspicuous every location in the input image is. We used the freely available implementation of the Itti–Koch saliency model [34]. In this model, an image is analyzed along multiple low-level feature channels to give rise to multi-scale feature maps, which detect potentially interesting local spatial discontinuities using simulated center-surround neurons. Twelve feature channels are used to simulate the neural features which is sensitive to color contrasts (red/green and blue/yellow), temporal intensity flicker, intensity contrast, four orientations (0°, 45°, 90°, and 135°) and four oriented motion energies (up, down, left, and right). The particular low-level features extracted here have been shown to attract attention in humans and monkeys, as had been previously investigated in details [19,35,36]. Center-surround scales are obtained from dyadic pyramids with 9 scales, from scale 0 (the original image) to scale 8 (the image reduced by factor to $2^8 = 256$ in both the horizontal and vertical dimensions). Six center-surround difference maps are then computed as point-to-point difference across pyramid scales, for combination of three center scales ($c = \{2, 3, 4\}$) and two center-surround scale differences ($\delta = \{3, 4\}$). Thus, six feature maps are computed for each of the 12 features, yielding a total of 72 feature maps. Each feature map is additionally endowed with internal dynamics that provide a strong spatial within-feature and within-scale competition for activity, followed by within-feature, across-scale competition [37]. In this way, initially possibly very noisy feature maps are reduced to sparse representations of only those locations which strongly stand out from their surroundings. All feature maps finally contribute to the unique scalar saliency map, which represents visual conspicuity of each location in the visual field.

After the saliency map is computed, a small number of discrete virtual foveas endowed with mass/spring/friction dynamics attempt to track a collection of most salient objects, using proximity as well as feature similarity to establish association between n salient locations and p fovea centers (similar to the approach described in our previous work [21,22]). The association is established through an exhaustive scoring of all $n \times p$ possible parings between a new salient location $X_t^A(i) = (x_t^A(i),\ y_t^A(i))$, $i \in \{1...n\}$ and an old foveation center $X_t(j) = (x_t(j),\ y_t(j))$, $j \in \{1...p\}$ at time t. (Typically, p is fixed and $n = p + 4$ to ensure robustness against varying saliency ordering from frame to frame, $p = 10$ in the present implementation). Four criteria are included to determine the correspondence: (1) Euclidean spatial distance between the locations of i and j; (2) Euclidean distance between feature vectors extracted at the locations of i and j which coarsely capture the visual appearance of each of the two locations; (3) a penalty term $|i-j|$ that discourages permuting previous pairings by encouraging a fixed ordered pairing; and (4) a tracking priority that increases with salience, enforcing strong tracking of only very salient objects. Combining these criteria tends to assign the most salient object to the first fovea, the second most salient object to the second fovea, etc. Video compression priority at every location is then related to the distance to the closest fovea center (using a 2D chamfer distance transform). For implementation details, please refer to [21] and [34]. It is important to note that the dynamics of the virtual foveas do not attempt to emulate human saccadic eye movement but track salient objects in a smooth and damped manner. The adopted correspondence and tracking algorithm compromises between reliably tracking the few most salient objects, and time-sharing remaining foveas among a larger set of less salient objects.

2.2. Bit allocation strategy

Assume that the rate-distortion (*R-D*) function is as follows [23,38] for a given region (typically, macroblock) i in an image:

$$D_i(R_i) = \sigma_i^2 * e^{-\gamma R_i} \tag{1}$$

in which D_i denotes the mean square error, R_i stands for the bit rate, and σ_i^2 is a measurement of the variance of the encoding signal (both spatial and temporal) and describes the complexity of the video content, γ is a constant coefficient. This approach assumes that the distortion is to be computed in a uniform manner, i.e., distortion in different areas of an image is equally important. However, if we take the human's visual spatial resolution into consideration, then the encoding distortion in area i can be written as follows:

$$D_i^{'} = w_i * D_i \tag{2}$$

here w_i is the weight coefficient, it stands for the human's spatial resolution in area i. In the area around the center of gaze (the fovea), w_i should be higher than in areas far from the gaze position, because even small distortions in the foveal region can cause people's awareness while in peripheral regions relatively larger distortions may not catch people's attention.

According to this non-uniform distribution of human eye resolution, there are two ways to optimize the bit allocation in encoding. One is as proposed in [23]: keep the sum of bit rate as a constant value and maximize the subjective quality. Under the hypothesis that σ_i^2 are equal at every location, the conclusion is then that the quantization parameter QP_i is inverse exponentially with the optimized bit rate:

$$R_i = R + \frac{1}{\gamma NS} \sum_{j \neq i} S_j \cdot \log \frac{w_i}{w_j} (i = 1, 2 \ldots N) \tag{3}$$

$$QP_i = e^{\frac{\alpha - R_i}{\beta}} \tag{4}$$

where S is the area size of the entire frame, S_i is the area of region i, N is the region number in one frame. α accounts for overhead bits and β is the adjustment parameter. However, in reality, $\sigma_i^2$are quite different over space within one frame and it is hard to determine the parameters correctly. In [23], the parameters are calculated from training videos that are similar to those to be encoded by the system, which is time consuming and may be unreliable if the test set differs substantially from the training set. To avoid this, we take a different approach: preserve the subjective quality while minimizing the bit rate. With this method, we find that the optimized distortion distribution is independent of the video frame contents and only depends on the weight coefficients. The details of this new method are described as follows:

To minimize the bit rate while keeping the subjective quality the same, we can write this global optimization problem as follows:

$$\begin{cases} Min \sum_i S_i * R_i / S \\ s.t. \sum_i w_i * D_i / W = D \end{cases} \tag{5}$$

here W is the sum of all of the weight coefficients in different areas, D is the target distortion. With the Lagrange multiplier method we can solve this equation in closed form:

$$\begin{cases} f(D_1, D_2, \ldots D_N) = \sum_i s_i * R_i / S + \lambda(\sum_i w_i * D_i / W - D) \\ R_i = \frac{1}{\gamma} (\log \sigma_i^2 - \log D_i) \end{cases} \tag{6}$$

in which N is the number of areas in the encoded image. To obtain the minimum value, we pose that, at the minimum:

$$\frac{\partial f}{\partial D_1} = \frac{\partial f}{\partial D_2} = \ldots = \frac{\partial f}{\partial D_N} = 0 \tag{7}$$

Solving these equations above, we obtain:

$$D_i = \frac{W * s_i}{w_i * S} * D \ (i = 1, 2, \ldots, N) \tag{8}$$

We can see from this equation that the optimum D_i is independent of the video characteristic related parameters γ and σ_i. Hence the optimum process can be applied to any video no matter what the content of the video is, and we need not train on any sample videos to compute the optimum parameters. Furthermore, from the Eq. (8) we can see that the optimized distortion should be inversely proportional to the weight coefficient. One special condition is when all weights are equal over all locations, in which case the distortion should be equally distributed.

Now we can determine the bit allocation strategy with the calculated distortion distribution. In mainstream video compression schemes developed so far, the distortion stems only from quantization. The basic quantization operation is as follows:

$$Y = \text{round}\left(X / Q_{step}\right) \tag{9}$$

where X usually is the coefficient after the transform (DCT, DWT, etc.), Q_{step} is the quantization step and Y is the quantized result. The Q_{step}-distortion mapping is linear and we can simply write the Q_{step}-distortion (Q-D) model as follows:

$$D = k * Q_{step} \tag{10}$$

where k is a constant coefficient related to the video content. Then the optimized distortion for each area transformed to the optimized quantization step is, at every location i:

$$Q_{istep} = \frac{W * s_i}{w_i * S} Q_{step} \tag{11}$$

The formula above shows that in the human visual characteristic based video coding, the quantization step should be inversely proportional to the subjective quality weight coefficient.

According to the analysis above, we can apply the GM to guide the quantization parameter adjustment to conduct the optimized bit allocation. The GM values can be seen as the subjective weight coefficients and the quantization parameters then can be computed from above formula.

3. Video acquisition

Fifty video clips (1920×1080) were collected for this experiment and each of them was cut to 300 frames (Fig. 2). All these clips were captured by a Silicon Imaging SI-1920HD camera with an EPIX E4 frame grabber card at frame rate of 30±0.2 fps. The original frames were captured as Bayer format without any compression and saved in round-robin onto 4 separate hard drives to avoid limitations in frame rate due to limited disk bandwidth. The clips were captured around the USC campus and include both outdoor and indoor scenes at daytime. The outdoor scenes include library, pool, traffic road, gardens, museum, park, gates, fountains, square, lawn, track & field, and the indoor scenes include dinner hall, lab rooms, etc. After video capture, all the frames were converted to RGB color images through linear interpolation [39] and enhanced by gamma correction. Finally the frames were assembled into video clips in the YUV (4:2:0) format for further processing.

To facilitate future research, this raw uncompressed video dataset, as well as all the eye-tracking data described below, are made freely available on the Internet (http://iLab.usc.edu/vagba/). We hope that this comprehensive dataset and the associated human eye movements can benefit a number of research projects aiming at improving video compression quality, and beyond.

4. Human eye-tracking

The collected 50 uncompressed YUV format video clips were presented to 14 subjects and their eye fixation points were recorded over frames from each clip by an eye-tracker machine. The recorded eye traces represent the subjects' shifting overt attention, thus the

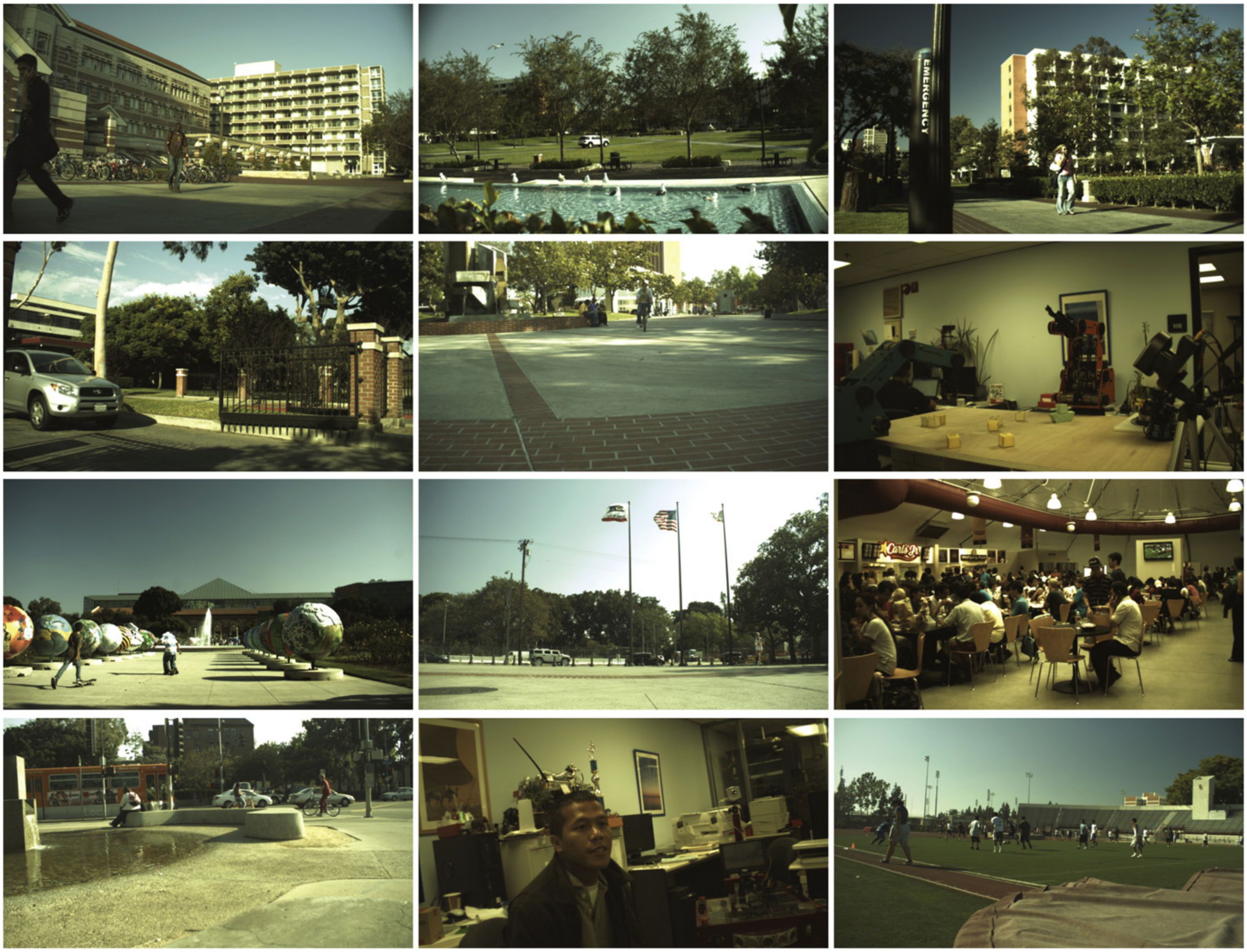


Fig. 2. Example of captured frames, first row: *dance01*, *seagull01*, and *garden09*, second row: *road02*, *fountain01* and *robarm01*, third row: *park01*, *gate03* and *lot01*, fourth row: *foutain05*, *room02* and *field06*.

eye-tracking data are qualified to validate the performance of the attention prediction model and the visual subjective quality.

Subjects were naïve to the purpose of the experiment and had never seen these video clips before. They were also naïve to attention theory, saliency theory, and video compression theory. They were USC students and staff (7 males, 7 females, mixed ethnicities, ages 22–32, normal or corrected-to-normal vision). They were instructed to watch the video clips without any specific task, and to attempt to follow whatever interesting things they might like. Later they were asked some general questions about what they had watched. The motivation of these instructions was to try to make the experiment similar to ordinary video watching. We believe that our instructions did not bias subjects toward low-level salient frame locations as defined by the Itti–Koch Model [18].

Stimuli were presented on a Sony Bravia XBR-III 46″ 60 Hz 1080 p LCD-TV display connected to a Linux computer. Subjects were seated

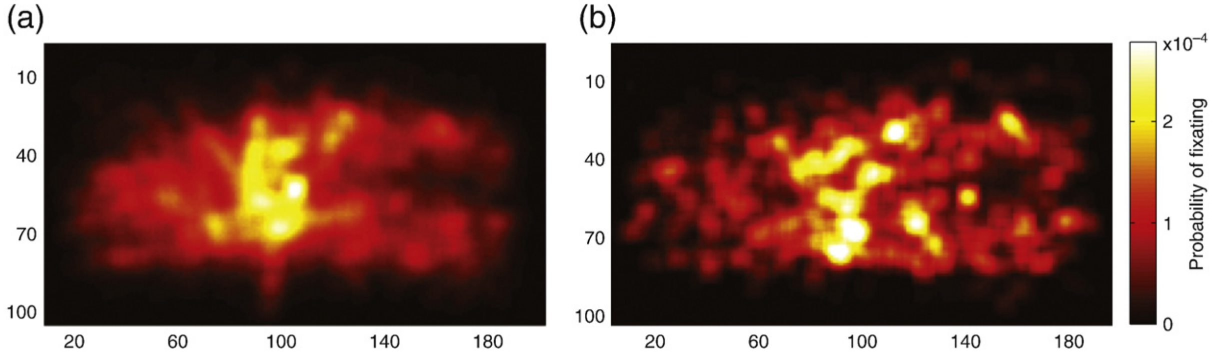


Fig. 3. Eye fixation distribution examples. The maps are histogrammed over 10 × 10 image tiles and normalized to 1. (a) The overall distribution for all subjects and clips shows a bias toward the center of the display. (b) Eye fixation distribution from one of the subjects over all the clips.

on an adjustable chair at a viewing distance of 97.8 cm, which responded to a field of view of 54.8 × 32.7°, and rested on a chin-rest. A nine-point eye-tracker calibration was performed every ten clips. Each calibration point consisted of fixating first a central cross, then a blinking dot at a random point on a 3 × 3 matrix covering the screen area. For each clip, subjects first fixated a central cross, pressed a key to start, at which point the eye-tracker was triggered, the cross blinked for 1066 ms, and the clip started. After each clip, the display became grey and the eye-tracker was disabled. The experiment was self-paced: the next clip was shown when the subject pressed the space button. Every ten clips, subjects could stretch before the nine-point calibration. Stimuli were presented on a Linux computer, under SCHED_FIFO scheduling (process would keep 100% of the CPU as long as needed) to guarantee timing. Each uncompressed clip (1920 × 1080, YUV 4:2:0 format) was entirely preloaded into memory. Frame displays were hardware-locked to the vertical retrace of the monitor (one movie frame was shown for two screen retraces, yielding a playback rate of 30.00 fps). Microsecond-accurate timestamps were stored in memory as each frame was presented, and later saved to disk to check for dropped frames. No frame drop ever occurred and all timestamps were spaced by 33.333 ± 0.001 ms. Eye position was tracked using a 240-Hz infrared-video-based eye-tracker (ISCAN, Inc., model RK-464). Methods were similar to previously described [21]. In brief, this machine estimates point of regard (POR) in real-time from comparative tracking of both the center of the pupil and the specular reflection of the infrared light source on the cornea. This technique renders POR measurements immune to small head translations (tested up to 10 mm in our laboratory). All analysis was performed offline. Linearity of the machine's POR-to-stimulus coordinate mapping was excellent, as previously tested using a 7 × 5 calibration matrix in our laboratory, justifying a 3 × 3 here. The eye-tracker calibration traces were filtered for blinks and segmented into two fixation periods (the central cross, then the flashing point), or discarded if that segmentation failed a number of quality control criteria. An affine POR-to-stimulus transform was computed in the least-square sense, outlier calibration points were eliminated, and the affine transform was recomputed. If fewer than six points remained after outlier elimination, recordings were discarded until the next calibration. Otherwise, a thin-plate-spline nonlinear warping was then applied to account for any small residual nonlinearity. Data was discarded until the next calibration if residual errors greater than 34 pixels (about 1° field of view) on any calibration point or 17 pixels (about 0.5° field of view) overall remained. Eye traces for the ten clips following a calibration were remapped to screen coordinates, or discarded if they failed some quality control criteria (excessive eye-blinks, motion, eye wetting, or squinting). Calibrated eye traces were visually inspected when superimposed with the clips.

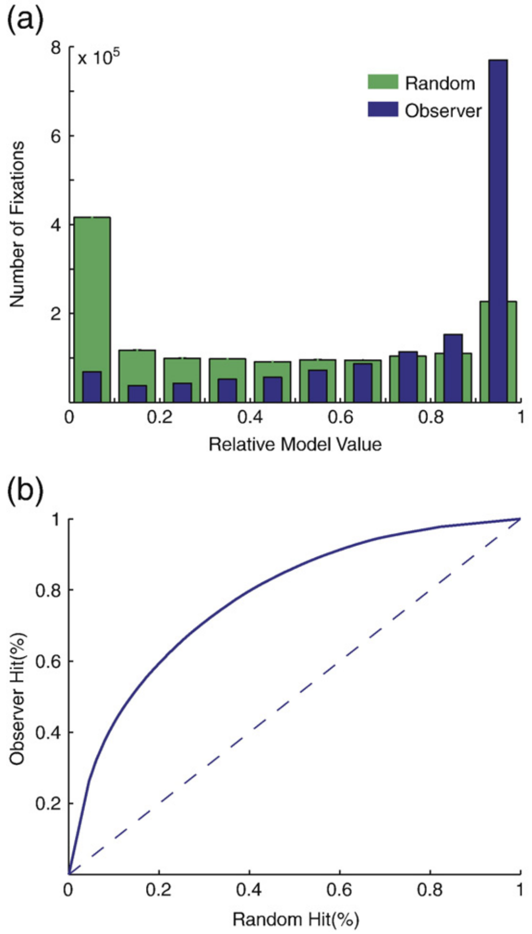


Fig. 5. Ordinal dominance analysis, there are 1,455,279 fixation points in total. (a) Histogram of guidance map values at eye positions and random locations. (b) Ordinal dominance curve, the dashed line is the chance level.

Eye fixation distribution examples can be seen in Figs. 3 and 4. We can see from Fig. 3 that the overall distribution of eye fixations is quite strongly center-biased on average, however, for different content clips, the eye fixation distributions are totally different and not necessarily center-biased (Fig. 4). This is important as it suggests that a simplistic saliency map – which would simply mark central screen regions as more salient – may work well on average but not necessarily for individual video clips (see [21] for

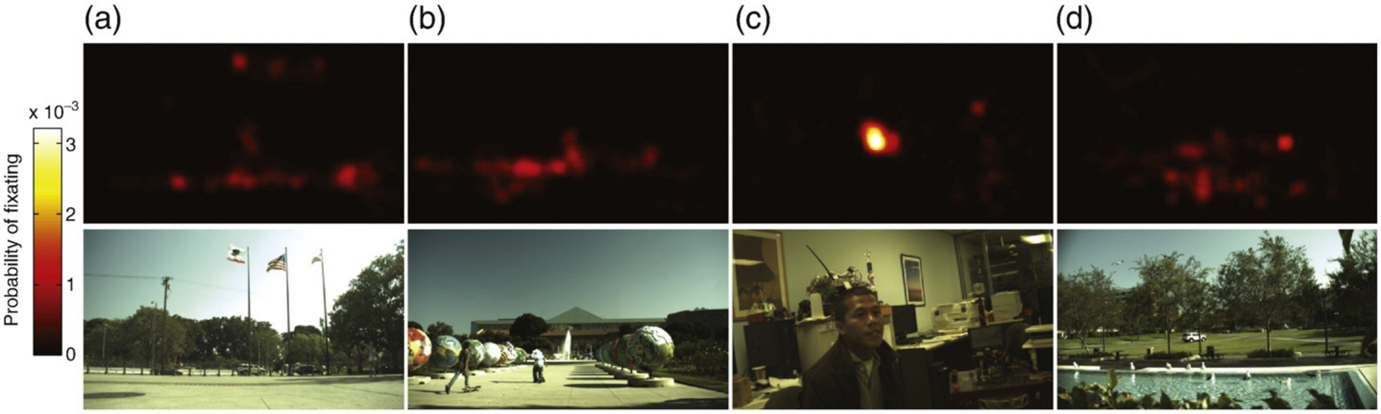


Fig. 4. Examples of eye fixation distribution map at different clips, the maps are histogrammed over 16 × 16 image tiles and normalized to 1. (a) *gate03*, (b) *park01*, (c) *room02*, (d) *seagull01*.

further discussion). In our results below we not only report average performance but also performance on individual clips and worst-case performance.

5. Experiment result

Uncompressed videos were shown to the subjects and the eye-tracking data was used to validate both the attention prediction model and the attention-based bit allocation scheme.

Considering that a good attention prediction model should output a model map which highlights the eye's fixation point, differences between guidance map values at subjects' gaze targets and at randomly selected locations were quantified, to evaluate performance, using ordinal dominance analysis [40]. Model map values at subjects' fixation points and at randomly selected locations were first normalized by the maximum value in the map when the eye fixation occurred (100 random locations are selected in this paper). Then, histograms of values at eye positions and random locations were created. Fig. 5(a) shows the results of the histograms over all video clip types and subjects. It is easy to see from the figure that many more human fixations were to high model salience values than expected by chance, which means the proposed model performs better than the random model at predicting human gaze. The mean observer value at the guidance map is 0.778 while the median value is 0.922, compared to mean 0.435 and median 0.400 obtained with random fixations (both model values are significantly higher than random value, $p<10^{-20}$, considering both t-tests for mean value comparisons and sign tests for median value comparisons). Fig. 6 shows the example of frames and their corresponding guidance maps. The subjects' eye fixation points are marked as small color patch in the frames.

To further measure the difference between the observer and random histograms, a threshold was decremented from 1 to 0, and at each threshold the percentage of eye positions and of random positions that were to a map value larger than the threshold ("hits") were computed. An ordinal dominance curve (similar to a receiver operating characteristic curve) was created with "observer-hits" versus "random-hits". The curve summarizes how well a binary decision rule based on thresholding the map values could discriminate signal (map values at observer eye positions) from noise (map values at random locations). The overall performance can be summarized by the area under the curve (AUC). An AUC area of 0.5 stands for a model which is at chance at predicting human gaze, while larger AUC values indicate better prediction performance. In our experiment, the ordinal dominance curve is plotted in Fig. 5(b), and the AUC value is 0.773 ± 0.002. As an upper bound, inter-observer correlations among humans yield an AUC of 0.854 ± 0.001.

As to the attention-based bit allocation in video compression, the latest video compression standard H.264/AVC and its reference software JM9.8 are adopted to implement the experiment. In H.264, a total of 52 different values of Q_{step} are supported and they are indexed by a Quantization Parameter (QP). Q_{step} increases by 12.5% for each increment of 1 in QP. In this paper, QPs are adjusted at the MB level, which means different QPs are computed for each MB. There are two reasons for this: first, in H.264 frames are encoded at the MB level, second, the generated guidance map has the same size as the frame size in MBs. QPs are computed according to Eq. (11) where w_i are replaced by the corresponding GM values and Q_{step} are taken from the baseline QP value. Furthermore, in order to keep the smoothness of perceptual quality, the biggest Q_{step} is constrained to equal or less than 2 times of the smallest Q_{step}, this means the difference between QPs in one frame is constrained into 6. In the implementation, the smallest QP is set to $\text{QP}_{\text{baseline}} - 2$ while the biggest QP is set to $\text{QP}_{\text{baseline}} + 3$.

To measure the subjective quality of encoded frames, eye-tracking data are applied to compute the weighted distortion. Here we propose to use a new eye-tracking weighted mean square error (EWMSE) and eye-tracking weighted peak signal-to-noise ratio (EWPSNR) metrics

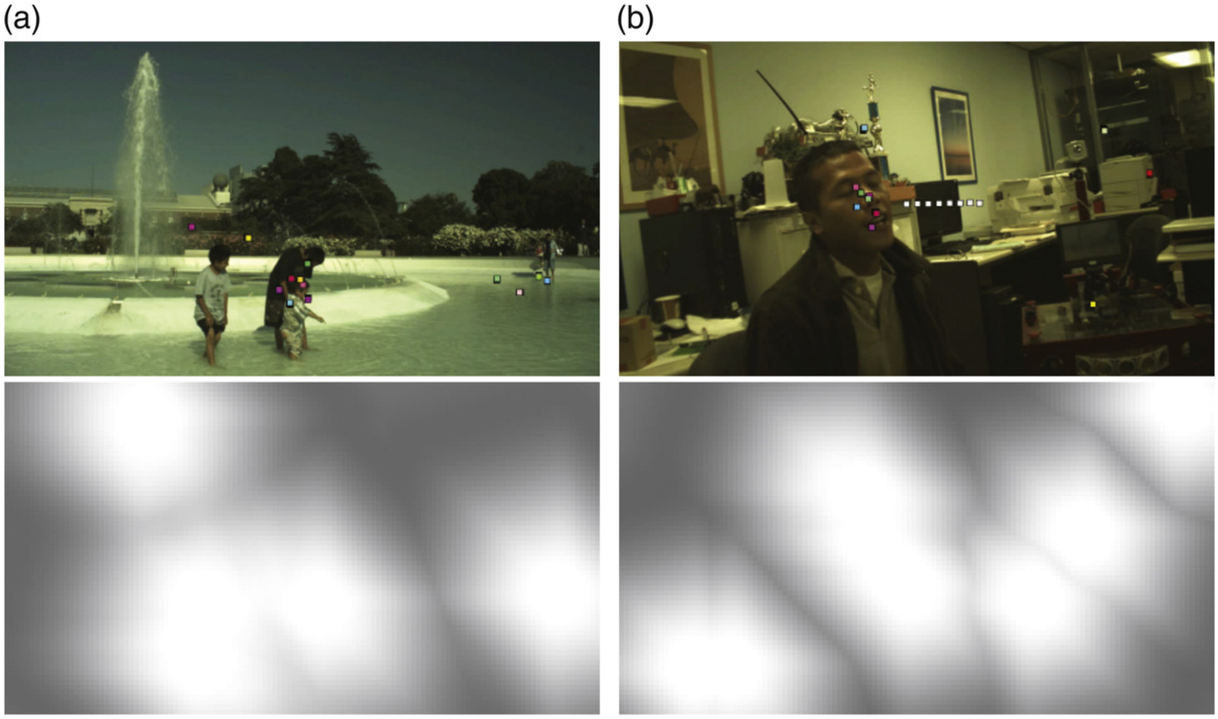


Fig. 6. Example of video frames (top) and the corresponding guidance maps (bottom). Subjects' eye fixations are marked as small square color patches in the frame, the white patches in (b) means a saccade. (a) Frame from *park03*, (b) frame from *room02*.

to measure subjective quality. The corresponding computation formulas are as follows:

$$EWMSE = \frac{1}{MN\sum\limits_{x=1}^{M}\sum\limits_{y=1}^{N} w_{x,y}} \sum_{x=1}^{M}\sum_{y=1}^{N}\left(w_{x,y}\bullet\left(I'_{x,y}-I_{x,y}\right)^{2}\right) \tag{12}$$

$$EWPSNR = 10 * \log\left(\frac{\left(2^{n}-1\right)^{2}}{EWMSE}\right) \tag{13}$$

$$w_{x,y} = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\left(\frac{(x-x_e)^2}{2\sigma_x^2}+\frac{(y-y_e)^2}{2\sigma_y^2}\right)} \tag{14}$$

where I and I' are the original frame and the encoded frame, respectively, M and N are the frame's height and width in pixels, n is the bit depth of the color component. $w_{x,y}$ is the weight for distortion at position (x, y) and normalized to $\sum_{x,y} w_{x,y} = MN$. $w_{x,y}$ is computed based on the subjects' eye fixation position (x_e, y_e) from eye-tracking experiment, $\sigma_x$and $\sigma_y$are two parameters related to the distance and view angle, usually taken from fovea size. Here we use 2° (64 pixels) of view field as σ_x and σ_y. The rationale for this weighting formula is that the photoreceptors in the human retina are in a highly non-uniform distribution: only a small region of 2–5° of visual angle (the fovea) around the center of gaze is captured at high resolution and the resolution falls off quickly around the fovea [3]. In our experiment, the eye fixations are recorded with a 240 Hz eye-tracker, considering that the video frame rate is 30 Hz, for each subject, 8 eye

Table 1

Comparison of EWPSNR results between JM9.8 and the proposed model (VAGBA — visual attention guided bit allocation) for different clips. Units in the table are in dB. QP means baseline quantization parameter. Gain (in dB) is the improvement of EWPSNR compared with the standard JM9.8 method. Gain>0 means that, for the current clip, the video subjective quality encoded by the proposed method is better than the one encoded by JM9.8.

Clip index	QP = 24			QP = 28			QP = 32			QP = 36		
	JM	VAGBA	Gain	JM	VAGBA	Gain	JM	VAGBA	Gain	JM	VAGBA	Gain
1	41.15	42.89	1.75	38.66	40.16	1.50	35.67	37.49	1.82	32.77	34.74	1.97
2	41.36	42.02	0.66	38.48	39.27	0.79	35.41	36.59	1.18	32.93	33.93	1.00
3	43.18	42.99	−0.19	40.56	40.56	0.00	38.08	38.17	0.09	35.77	35.85	0.08
4	42.67	42.83	0.16	40.47	40.52	0.05	38.03	38.29	0.27	34.92	36.08	1.16
5	41.77	43.08	1.31	38.93	40.34	1.41	36.03	37.58	1.56	33.21	34.84	1.64
6	41.95	42.47	0.52	39.25	39.74	0.49	36.54	37.06	0.51	33.78	34.46	0.69
7	42.76	42.69	−0.07	40.06	40.08	0.02	37.14	37.49	0.35	34.32	34.96	0.65
8	43.35	43.74	0.39	40.71	41.20	0.49	38.35	38.69	0.34	35.67	36.18	0.51
9	42.07	42.58	0.50	39.29	39.91	0.61	36.56	37.30	0.74	33.97	34.59	0.62
10	41.51	42.00	0.49	38.79	39.24	0.45	36.06	36.52	0.46	33.38	33.84	0.45
11	41.89	42.50	0.61	39.34	39.82	0.48	36.74	37.23	0.49	34.04	34.59	0.56
12	41.30	41.66	0.36	38.57	38.88	0.31	35.73	36.18	0.45	33.00	33.51	0.51
13	42.15	42.76	0.61	38.98	40.12	1.14	36.83	37.50	0.67	33.68	34.84	1.16
14	40.22	41.14	0.92	37.38	38.23	0.85	34.57	35.39	0.82	31.92	32.66	0.73
15	41.84	42.57	0.73	39.14	40.01	0.87	36.54	37.42	0.88	33.92	34.83	0.91
16	41.22	42.39	1.17	38.15	39.56	1.41	35.48	36.83	1.36	32.70	34.11	1.41
17	43.07	43.45	0.38	40.44	40.93	0.49	37.64	38.45	0.82	35.24	35.95	0.71
18	41.29	42.13	0.84	38.32	39.48	1.16	35.64	36.91	1.27	33.15	34.40	1.25
19	41.68	42.67	0.98	38.72	39.98	1.26	36.27	37.32	1.05	33.49	34.66	1.17
20	42.73	42.64	−0.09	40.20	39.99	−0.20	37.55	37.37	−0.18	34.98	34.75	−0.24
21	42.87	43.39	0.52	40.25	40.81	0.56	37.51	38.23	0.72	34.91	35.66	0.75
22	42.99	42.50	−0.48	40.48	39.92	−0.55	37.90	37.34	−0.56	35.28	34.81	−0.46
23	44.76	45.70	0.94	42.57	43.49	0.92	40.09	41.09	1.01	37.65	38.55	0.90
24	41.69	42.48	0.79	39.09	39.83	0.73	36.44	37.17	0.72	33.77	34.51	0.73
25	43.12	43.55	0.43	40.61	41.02	0.41	38.10	38.47	0.37	35.49	35.85	0.36
26	42.68	42.84	0.16	39.99	40.12	0.13	37.18	37.36	0.18	34.32	34.62	0.31
27	44.82	45.50	0.67	42.75	43.47	0.72	40.43	41.20	0.77	37.94	38.73	0.79
28	42.45	43.25	0.81	39.55	40.76	1.204	37.17	38.29	1.12	35.05	35.79	0.75
29	42.05	43.10	1.05	39.05	40.38	1.33	36.16	37.68	1.52	33.41	34.98	1.57
30	42.18	42.34	0.16	39.52	39.58	0.06	36.93	36.96	0.03	34.18	34.38	0.21
31	43.66	45.95	2.29	41.72	43.94	2.22	38.84	41.61	2.77	36.34	39.12	2.77
32	41.15	42.55	1.40	38.38	39.77	1.39	35.28	37.02	1.74	32.25	34.28	2.03
33	42.22	43.04	0.83	39.48	40.39	0.91	36.40	37.79	1.39	33.86	35.23	1.37
34	42.69	43.36	0.67	40.21	40.79	0.58	37.72	38.27	0.55	34.95	35.84	0.88
35	44.65	45.38	0.74	43.08	43.61	0.54	41.01	41.48	0.47	38.75	39.21	0.46
36	42.15	42.68	0.53	39.17	40.01	0.84	36.35	37.38	1.03	33.33	34.79	1.46
37	42.49	42.84	0.35	39.73	40.13	0.40	36.97	37.50	0.53	34.26	34.90	0.64
38	42.56	43.37	0.81	40.08	40.82	0.74	37.54	38.27	0.73	34.85	35.71	0.87
39	42.00	42.04	0.03	39.49	39.36	− 0.13	36.92	36.77	− 0.15	34.42	34.22	− 0.20
40	41.77	42.79	1.02	38.74	40.08	1.35	35.74	37.46	1.72	32.95	34.84	1.89
41	42.73	43.17	0.44	39.79	40.48	0.69	37.00	37.88	0.88	34.43	35.30	0.87
42	41.86	42.62	0.76	39.29	39.98	0.69	36.74	37.40	0.66	34.23	34.91	0.68
43	43.61	44.05	0.44	40.82	41.54	0.72	38.22	39.02	0.79	35.64	36.56	0.93
44	42.79	43.67	0.88	40.46	41.12	0.65	38.59	38.53	− 0.06	35.80	36.03	0.23
45	42.79	42.84	0.05	40.22	40.24	0.02	37.41	37.63	0.223	34.51	35.08	0.56
46	41.63	42.74	1.10	38.97	39.97	1.00	36.26	37.26	1.01	33.37	34.59	1.22
47	42.14	43.64	1.50	39.23	41.05	1.81	36.74	38.47	1.73	34.17	35.95	1.79
48	42.51	42.74	0.23	39.89	40.14	0.25	37.15	37.50	0.36	34.27	34.90	0.62
49	41.54	43.07	1.53	38.65	40.31	1.66	35.58	37.58	2.01	32.57	34.83	2.26
50	42.48	43.81	1.33	40.16	41.31	1.15	37.02	38.77	1.75	34.42	36.22	1.79
Average	42.36	43.04	0.68	39.71	40.44	0.73	37.03	37.85	0.82	34.35	35.27	0.92

fixation points need to be taken into account for each frame. Therefore, the weight $w_{x,y}$ in reality is a combination of all 8 different eye fixation points. Furthermore, the saccade data are not considered in computing the EWPNSR and only the fixation points are taken into account. We did this because human take saccade very quickly and do not pay much attention to the saccade regions. The mean EWPSNR from all the subjects is adopted as the measurement to evaluate the video subjective quality: the higher EWPNSR value, the better subjective quality.

To show the effectiveness of the proposed visual attention guided bit allocation method in improving the video subjective quality, we compare the encoded video EWPSNR from the proposed method and the standard method in JM9.8 with matched bit rate through the frame-level rate control algorithm. The configuration of the encoder is as follows: intra period = 30, Hadamard transform, UVLC, no fast motion estimation, no B frame, high complexity RDO mode, no restriction in search range. We test 4 baseline QPs (initial QP): 24, 28, 32 and 36, and the bit rates range from 260 Kbps to 10 Mbps. The rate-controlled bit rates with the standard encoder precisely match the bit rate with the proposed new encoder (within 1% difference). Table 1 lists all the EWPSNR results from proposed method, the results from the JM9.8 standard method, and the subjective quality improvement (gain). Better EWMSE and EWPSNR is expected to be obtained for our method vs. JM only if, on average, the predictions of the saliency model agree with where humans look, and, when they disagree, the higher distortions that our system will introduce at the locations looked at by humans do not outweigh the lower distortions obtained when the model is correct. In addition, Fig. 7 plots the results at different baseline QP, and sorts the results according to improvement. It is easy to see that only for a few (3–4) clips the subjective quality is worse with our proposed method than with the standard method, while most of the clips achieve a better subjective quality, with improvement for some of them up to about 2 dB EWPSNR. Thus, the proposed scheme can significantly ($p<0.002$, t-test) achieve better subjective quality (as defined by our EWPSNR measure) while keeping the same bit rate. Furthermore, the comparison of histograms of PSNR results at eye fixation regions (2° of visual field) with different methods is plotted in Fig. 8, from the figure it is easy to see that more encoded frames have higher PSNR in the fixation region with the proposed method compared with the standard JM9.8 method. Fig. 9 shows two examples of EWPSNR over the clip frames. We can see

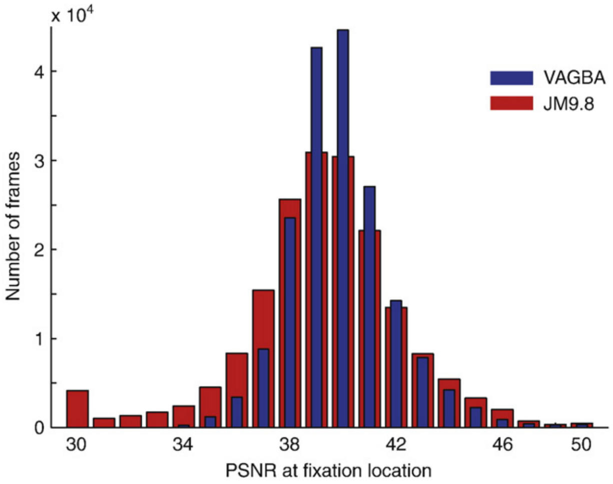


Fig. 8. Comparison of histograms of PSNR results at eye fixation regions with standard JM9.8 method and the proposed VAGBA method (initial QP = 28). The fixation regions are used 2° of view.

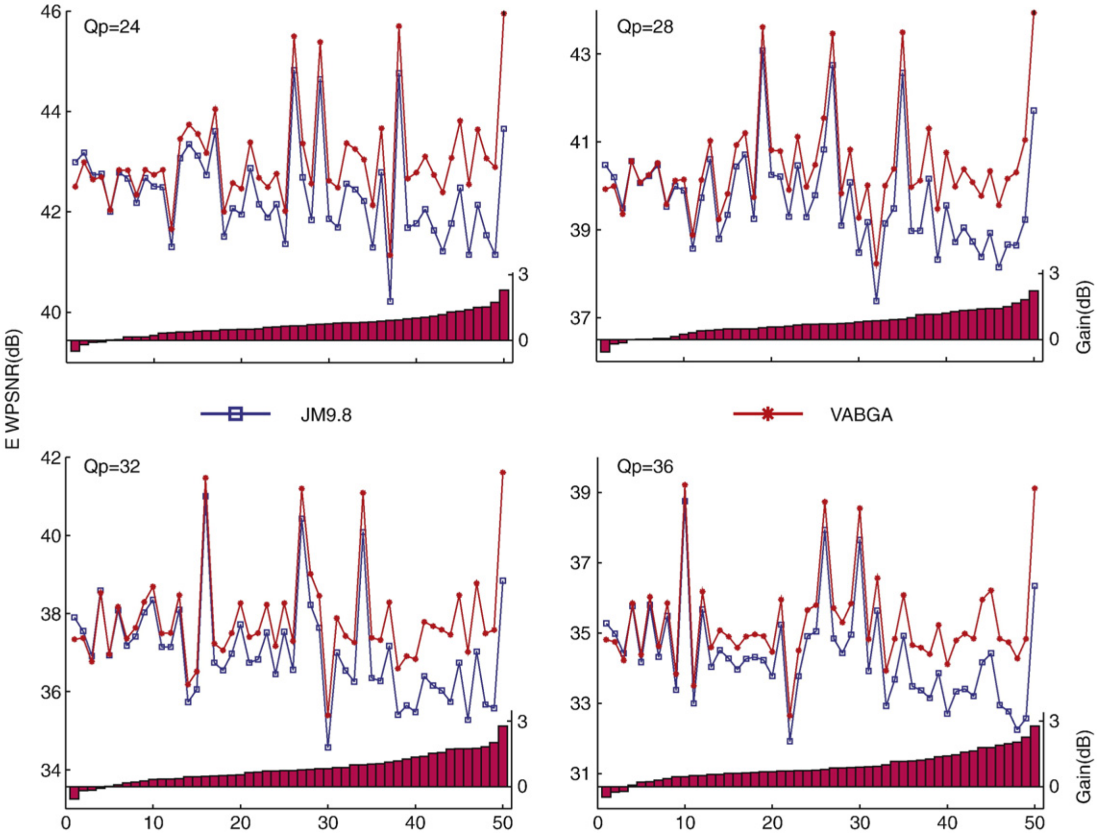


Fig. 7. EWPSNR results comparison between our proposed VAGBA model and JM9.8 at the same bit rate, the horizontal axis represent the clip index, after sorting by the subjective quality improvement (the red bars shown in each plot).

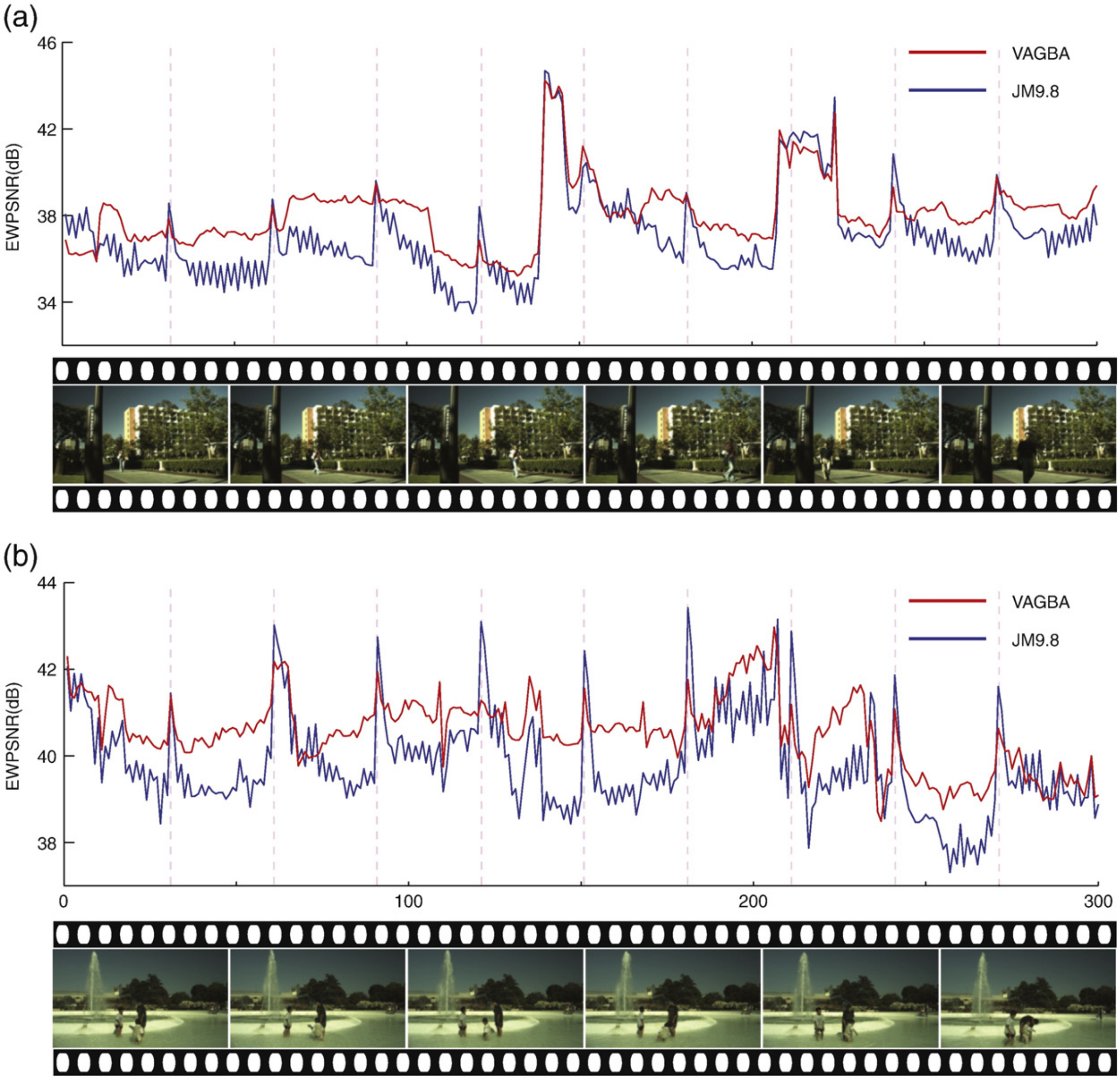


Fig. 9. Two examples of EWPSNR over the clip frames. (a) *garden09* from subject KC, initial QP = 28; (b) *park03* from subject SH, initial QP = 28. Dashed grey lines indicate intra-coded frames (every 30 frames with our codec settings).

from the figure that for most frames, the subjective quality of the proposed method is better than the standard JM9.8 encoded result. However, for some frames, the standard method achieves better subjective quality. There are mainly two reasons for this: first, if the current frame is an intra refresh frame, the rate control algorithm in H.264 usually assigns a relatively smaller QP to such intra frame to achieve better prediction results for later P frames. In these cases, the current frame's quality could be better than with the proposed method. Second, the attention prediction model is not guaranteed to always accurately predict human's attention regions for all the frames, such that if the prediction failed for the current frame, then due to the bit allocation strategy, the true attention region will receive fewer bits to encode thus will make the subjective quality worse.

Furthermore, the comparison between the proposed method and our previous method proposed in [21] which guide video compression through selective blurring (foveation) of low-salience image regions is conducted. The foveated clips are encoded by JM9.8 with matched bit rate through the frame-level rate control algorithm (within 1% difference). Fig. 10 shows the comparison results at different baseline QP, and sorts the results according to improvement of EWPSNR. From the figure we can see that for all clips the subjective quality is significantly better with the proposed method than with the foveation method ($p<10^{-10}$, *t*-test). The average improvement in EWPNSR is 2.533 dB. Also we can see from the figure that the improvement is higher when the baseline QP is lower, this is because the foveation degrade the video quality more than the encode error when the quantization step is small. As another example, Fig. 11 compares the visual qualities of three partially reconstructed frames among the standard rate control method, the foveation method and the proposed bit allocation method. The difference frames (encoding error) are also plotted to make the comparison clearer. As shown in the figure, the encoded frame by the proposed method has better visual quality than the frame encoded by standard method in the interesting region.

6. Discussion

The proposed attention model predicts human attention accurately in most cases, and based on this, the bit allocation algorithm can improve the subjective visual quality significantly while keeping the same bit rate. The contributions in this paper include several aspects:

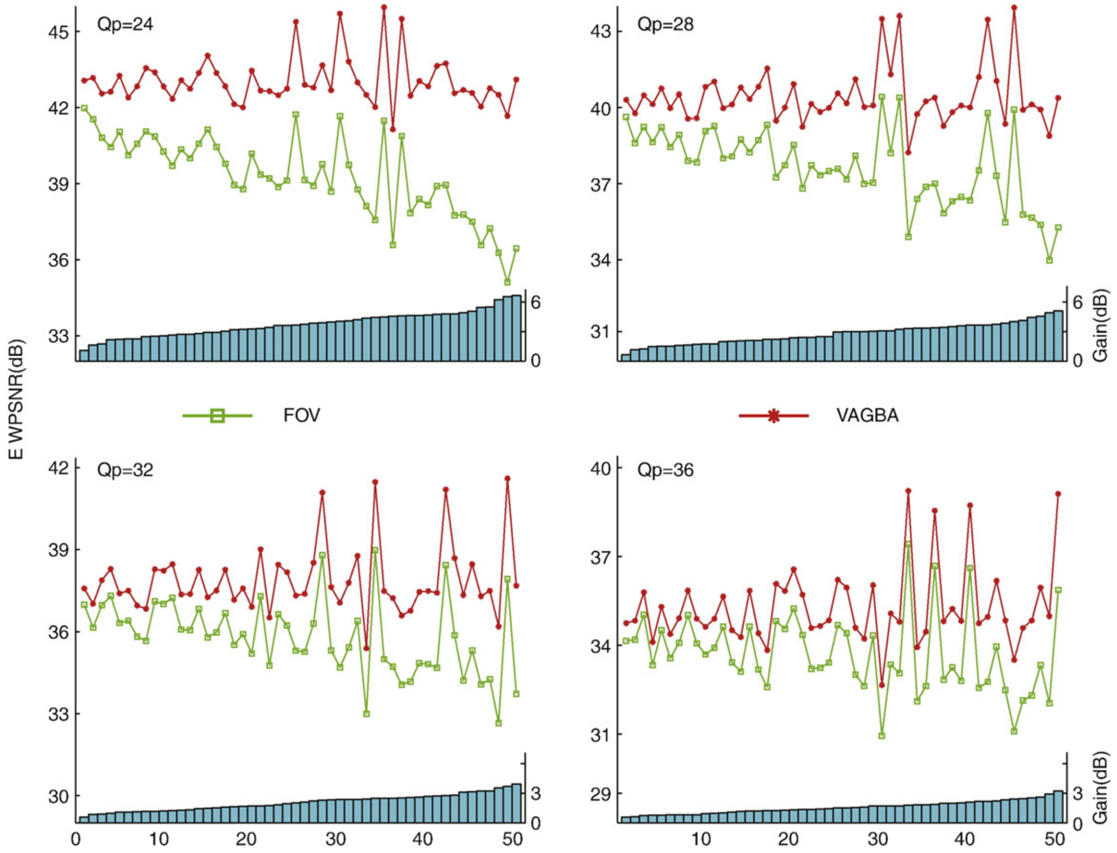
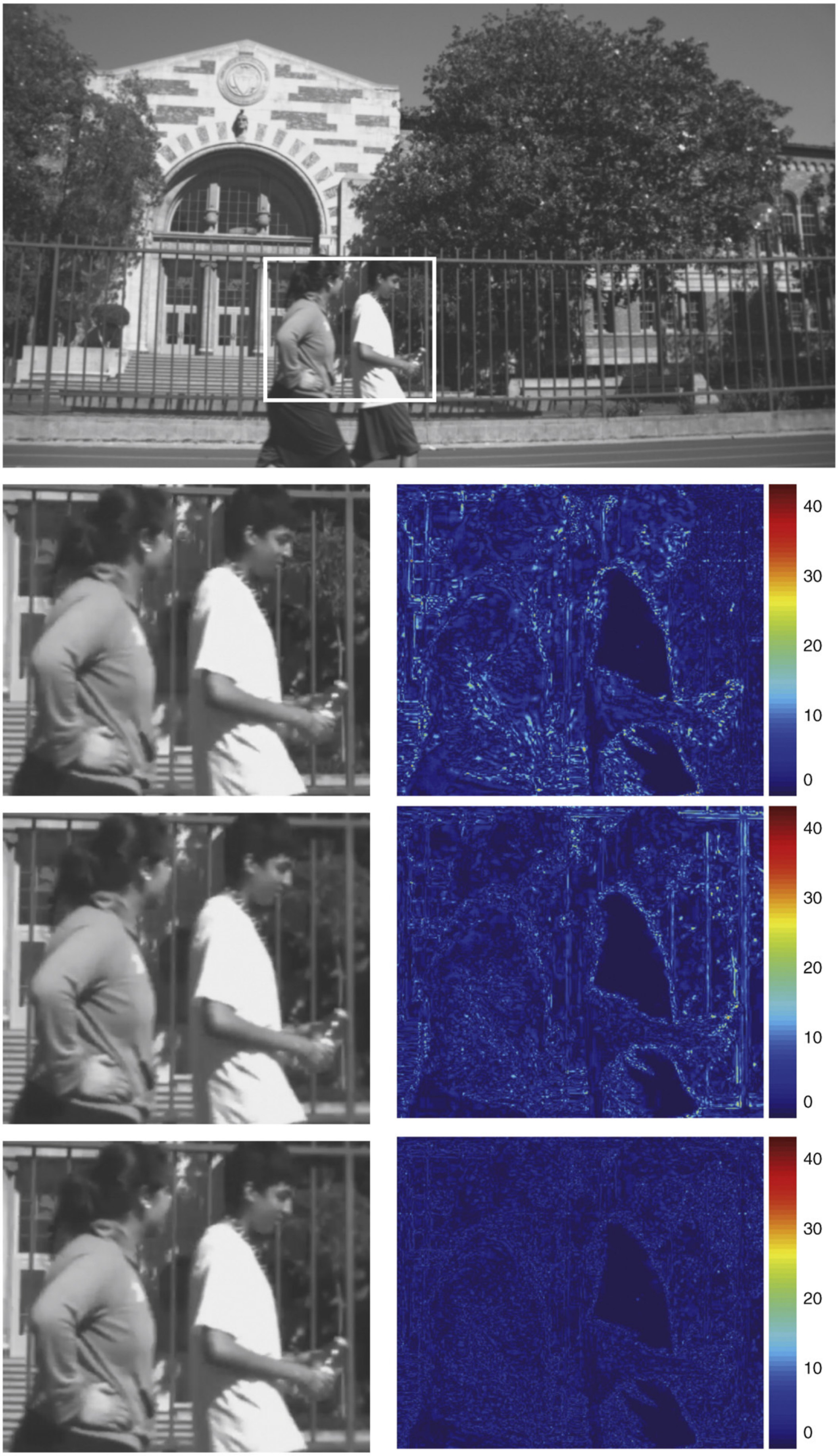


Fig. 10. EWPSNR results comparison between our proposed VAGBA model and the foveation method (mark as FOV in the figure) at the same bit rate, the horizontal axis represent the clip index, after sorting by the subjective quality improvement (the green bars shown in each plot).

(1) Combining the attention prediction model with state-of-the-art video compression method, the proposed method is fully automatic and can be applied to any kind of video categories without any restriction. In addition to combining with the H.264 codec, the proposed method can be applied to any kind of codec proposed so far. (2) A new bit allocation strategy was proposed through solving in closed form the constrained global optimization problem. (3) Using eye-tracking data to evaluate compressed video frame quality in a quantitative way (instead of subjective rating or binary selection). The new proposed EWPSNR subjective quality measurement is based on the human vision's characteristic and can represent the non-uniform subjective distortion in a reasonable way. (4) Collecting a high-definition video sequence database with eye-tracking data and distributing them on the Internet. This dataset can be used for video compression purposes as well as attention prediction purposes. Also, the raw captured frames are in Bayer format, which means this dataset can be used for Bayer format related image processing research.

The target in this paper is to validate the effectiveness in improving the subjective quality while keeping the same bit rate and employing a purely algorithmic method which does not require manual parameter tuning. Thus far, the bit allocation strategy is an open-loop algorithm, which means that it only adjusts the bit allocation according to the guidance map and takes no constrain to keep any presumed bit rate. In our implementation, we first compress the video sequences with the proposed method, after that, we use the available rate control algorithm in JM9.8 to match the bit rate and use this result to compare with our proposed method. The comparison is reasonable because there is no scene change in the test sequences and thus both bit rate and visual quality should not fluctuate too much over video frames. Although the proposed method is not bit rate constrained, it can be used in many applications such as video storage, band-free video stream transmission, etc. However, it remains a great task to develop a bit rate constrained bit allocation strategy based on the visual attention model. Another point which is not addressed by an open-loop algorithm but could be studied more closely in the future is how the algorithm itself may introduce artifacts in the low-bit rate regions of the compressed video, which may themselves be salient and attract human attention (see [21] for more detailed discussion). This could be addressed in future versions of our algorithm where the saliency map may be computed on the compressed video clips as well, to check for the introduction of possibly salient artifacts during compression.

Eye-tracking data recorded from subjects viewing the uncompressed video clips are applied in evaluating the subjective quality. The rationale for viewing uncompressed video lies in two aspects: first, the eye-tracking traces from the uncompressed videos show the real attention regions of the original clips. Second, the ideal subjective quality measurement should use eye-tracking data from both the proposed method encoded video and standard rate-controlled video. However, it is impossible to obtain these two kinds of eye-tracking data from the same subject without affecting a priori knowledge: no matter which kind of encoded video is presented to subjects first, the eye-tracking data from the second presentation would likely be affected by the fact that subjects have already seen essentially the same clips before. Considering that artifacts might attract attention, two steps are adopted to reduce the quality fluctuation: first, both temporal and spatial smooth operation are conducted in computing the guidance map; Second, the biggest Q_{step} is constrained to equal or less than 2 times of the smallest Q_{step} in one frame to keep the perceptual quality, in the implementation, the smallest QP is set to $\mathrm{QP}_{\mathrm{baseline}}-2$ while the biggest QP is set to


40
30
20
10
0
40
30
20
10
0
40
30
20
10
0

$QP_{\text{baseline}}+3$. In our experiment, we check the compressed videos and found there is no big quality fluctuation in both spatial and temporal.

EWPSNR is proposed in computing the subjective quality. Here, we wanted to investigate whether we could test our algorithm in a more objective and more informative way than using subjective quality ratings, where observers may not always be able to rationalize or explain their ratings. Many existing computational subjective quality measurements [26,41,42] often tend to rely on some knowledge about the human visual system to decide what may be more visible or important to a human observer. For example, see the measures of FPNSR (Foveated PSNR, [26]), SSIM (structural similarity, [41]), and DVQ (Digital Video Quality, [42]). Thus we have been concerned that using these measures may be somewhat circular: (1) process video images with some saliency algorithm and allocate more bits (lower distortion) to more salient regions; (2) measure subjective quality with an algorithm that is very similar to our saliency computation and hence may be quite strongly correlated with it. We would quite naturally expect good subjective quality. This is what prompted us to develop the EWPSNR metric. We believe that it is an objective way to measure subjective quality, and it has the advantage of not relying on any algorithmic assumptions regarding how subjective quality may be defined. Here we just assume that the locations which people look at are the ones which will matter in terms of subjective quality. Note that this assumption is itself an imperfect one (subjective quality seems to be influenced not only by foveal vision but also by peripheral vision). But we believe that it at least avoids circularity in our testing, and it also provides a very informative assessment of where the algorithm is working (good agreement between the algorithm and human gaze) or failing.

Over the 50 tested video clips, there are 3–4 cases in which the subjective quality of clips encoded by our proposed method is worse than the clips encoded by the standard H.264 method. The worst two clips are *gate03* and *seagull01*, example frames from these two clips can be seen in Figs. 2 and 4. The reason of the failure mainly is that, for these clips, the attention prediction model results do not match well the subjects' attention. In the proposed attention prediction model, high motion regions take higher saliency value, however, in the *gate03* clip, the high speed cars were less interesting to our human observers than the jogging girl and the flags. In *seagull01* clip, the seagulls fly everywhere and the video is less meaningful in content, the subjects' eye-tracking traces are highly divergent, thus the proposed attention prediction model cannot predict the attention for all the subjects accurately.

The proposed attention prediction model in this paper purely depends on the bottom-up low-level features. These features are independent of the video contents and can be applied to any kind of conditions. However, in many specific cases, top-down influences can be taken into consideration to improve the attention prediction performance. For example, in teleconference videos or face-oriented conditions, face information (using, e.g., face detection algorithms) can be added as an important factor in predicting the attention [11]. In a specific search task (person detection), the saliency, target features, and scene context combined models can predict 94% of human agreement [43]. Also, considering the layout information, the “gist” of scenes can be applied to improve the prediction by learning broad scene categories [44,45]. All these top-down factors may combine with the bottom-up model to improve the attention prediction performance and thus improve the subjective quality of compression in specific corresponding conditions.

Acknowledgments

This study was supported by NSF, ARO, DARPA, Natural Science Foundation of China (NSFC) and China Scholarship Council (CSC). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

[1] T. Weigand, G.J. Sullivan, A. Luthra, Draft ITU-T recommendation H.264 and final draft international standard 14496-10 advanced video coding, Joint Video Teams of ISO/IECJTCI/SC29/WG11 and ITU-T SG/16/Q.6Doc.JVT-G050r, Geneva, Switzerland, 2003, May.

[2] T. Wiegand, G.J. Sullivan, G. Bjntegaard, A. Luthra, Overview of the H.264/AVC video coding standard, IEEE Trans. Circuits and Syst. Video Technol, vol. 13, July 2003, pp. 560–576.

[3] B. Wandell, Foundations of Vision, Sinauer, Sunderland, MA, 1995.

[4] P.T. Kortum, W.S. Geisler, Implementation of a foveated image coding system for bandwidth reduction of video images, Proc. SPIE, vol. 2657, 1996, pp. 350–360.

[5] N. Doulamis, A. Doulamis, D. Kalogera, S. Kollias, Improving the performance of MPEG coders using adaptive regions of interest, IEEE Trans. On Circuits syst. Video Technol, vol. 8, 1998, pp. 928–934.

[6] S. Lee, A.C. Bovik, Y.Y. Kim, Low delay foveated visual communications over wireless channels, Proc. IEEE Int. Conf. Image Processing, 1999, pp. 90–94.

[7] U. Rauschenbach, H. Schumann, Demand-driven image transmission with levels of detail and region s of interest, Comput. Graph, vol. 23, no. 6, 1999, pp. 857–866.

[8] D.J. Parkhurst, E. Niebur, Variable-resolution displays: a theoretical, practical, and behavioral evalutation, Human Factors, vol. 44, no. 4, 2002, pp. 611–629.

[9] M. Chi, M. Chen, C. Yeh, J. Jhu, Region-of-interest video coding based on rate and distortion variations for H.263+, Image Communication, vol. 23, no. 2, 2008, pp. 127–142.

[10] O. Hershler, S. Hochstein, At first sight: a high-level pop out effects for faces, Vision Research, vol. 45, no. 13, 2005, pp. 1707–1724.

[11] M. Cerf, J. Harel, W. Einhauser, C. Koch, Predicting human gaze using low-level saliency combined with face detection, Advances in neural information processing systems, vol. 20, 2008, pp. 241–248.

[12] K.C. Lai, S.C. Wong, K. Lun, A rate control algorithm using human visual system for video conferencing systems, Proc. Int. Conf. Signal Processing, vol. 1, Aug. 2002, pp. 656–659.

[13] L. Tong, K.R. Rao, Region-of-interest based rate control for low-bit-rate video conferencing, Journal of Electronic Imaging, vol. 15, no. 3, July, 2006.

[14] L. S. Karlsson, “Spatio-temporal pre-processing methods for region-of-interest video coding,” PhD dissertation, Sundsvall, Sweden, 2007.

[15] C.-W. Tang, C.-H. Chen, Y.-H. Yu, C.-J. Tsai, Visual sensitivity guided bit allocation for video coding, IEEE Trans. Multimedia, vol. 8, Feb. 2006, pp. 11–18.

[16] K. Minoo, T.Q. Nguyen, Perceptual video coding with H.264, Proc. 39th Asilomar Conf. Signals, Systems, and Computers, Nov. 2005.

[17] C. Huang, C. Lin, A novel 4-D perceptual quantization modeling for H.264 bit-rate control, IEEE Trans. On Multimedia, vol. 9, no. 6, 2007, pp. 1113–1124.

[18] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, Nov. 1998, pp. 1254–1259.

[19] L. Itti, C. Koch, Computational modeling of visual attention, Nature Reviews, Neuroscience, vol. 2, Mar, 2001, pp. 194–203.

[20] T. Liu, N. Zheng, W. Ding, Z. Yuan, Video attention: learning to detect a salient object sequence, Proc. ICPR, 2008, pp. 1–4.

[21] L. Itti, Automatic foveation for video compression using a neurobiological model of visual attention, IEEE Trans. on Image Processing, vol. 13, no. 10, 2004, pp. 1304–1318.

[22] L. Itti, Automatic attention-based prioritization of unconstrained video for compression, Proc. SPIE Human Vision and Electronic Imaging, vol. 5292, 2004, pp. 272–283.

[23] W. Lai, X. Gu, R. Wang, W. Ma, H. Zhang, A content-based bit allocation model for video streaming, Proc. IEEE international Conference on Multimedia and Expo, 2004, ICME).

[24] Z. Chen, G. Qiu, Y. Lu, L. Zhu, Q. Chen, X. Gu, W. Charles, Improving video coding at scene cuts using attention based adaptive bit allocation, Proc. ISCAS, 2007, pp. 3634–3638.

[25] M. Jiang, N. Ling, On Lagrange multiplier and quantizer adjustment for H.264 frame layer video rate control, IEEE Trans. Circuits and Syst. Video Technol, vol. 16, no. 5, 2006, pp. 663–669.

[26] S. Lee, M.S. Pattechis, A.C. Bovik, Foveated video compression with optimal rate control, IEEE Trans. on Image Processing, vol. 10, no. 7, 2001, pp. 977–992.

[27] Y. Sun, I. Ahmad, D. Li, Y. Zhang, Region-based rate control and bit allocation for wireless video transmission, IEEE Trans. on Multimedia, vol. 8, no. 1, 2006, pp. 1–10.

Fig. 11. Example of visual quality comparison between encoded frame (partially) by proposed method, foveation method and standard method. Top part is the Y component of the original 230th frame from *field03* clip. From row 2 to row 4 are the results from JM9.8 rate control method (PSNR = 38.37 dB), foveation method (PSNR = 30.76 dB), and proposed method (PSNR = 39.37 dB), respectively. For each row from row 2 to row 4, left side shows the encoded frame while right side shows the encode error. All the encoded frames are shown only partially to the interesting region.

[28] Y. Liu, Z. Li, Y.C. Soh, Region-of-interest based resource allocation for conversational video communications of H.264/AVC, IEEE Trans. Circuits and Syst. Video Technol, vol. 18, no. 1, 2008, pp. 134–139.

[29] Z. Li, L. Itti, Visual attention guided video compression, Proc. Vision Science Society Annual Meeting (VSS08), May, 2008.

[30] N. Wang, Q. Zhang, G. Li, Objective quality evaluation of digital video, Proc. PCCAS, 2000, pp. 791–794.

[31] A.A. Webster, C.T. Jones, M.H. Pinson, S.D. Voran, S. Wolf, A.A. Webster, C.T. Jones, M.H. Pinson, S.D. Voran, S. Wolf, Objective video quality assessment system based on human perception, Proc. SPIE, vol. 1913, 1993, pp. 15–26.

[32] A.B. Watson, J. Hu, J.F. McGowan, Digital video quality metric based on human vision, Journal of Electronic Imaging, vol. 10, no. 1, 2001, pp. 20–29.

[33] "Tutorial: objective perceptual assessment of video quality: full reference television,", ITU-T Technical tutorials, 2005.

[34] http://ilab.usc.edu/toolkit.

[35] J.M. Wolfe, Visual memory: what do you know about what you saw? Current Biology, vol. 8, 1998, pp. 303–304.

[36] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cognition Psychology, vol. 12, 1980, pp. 97–136.

[37] L. Itti, C. Koch, Feature combination strategies for saliency-based visual attention systems, Journal of Electronic Imaging, vol. 10, Jan. 2001, pp. 161–169.

[38] Y. Wang, J. Ostermann, Y. Zhang, Video Processing and Communication, Pearson Education, 2002.

[39] H.S. Malvar, L.W. He, R. Cutler, High-quality linear interporlation for demosaicing of bayer-patterned color images, Proc. ICASSP, 2004, pp. 485–488.

[40] D. Bamber, The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, Journal of Mathematical Psychology, vol. 12, 1975, pp. 375–387.

[41] M. Vranjes, S. Rimac-Drlje, D. Zagar, Subjective and objective quality evaluation of the H.264/AVC coded video, Proc. Systems, Signals and Image Processing, 2008.

[42] A.B. Watson, Toward a perceptual video quality metric, Human Vision, Visual Processing, and Digital Display, vol. 3299, 1998, pp. 139–147.

[43] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, A. Oliva, Modeling search for people in 900 scenes: a combined source model of eye guidance, Visual Cognition, vol. 17, no. 6&7, 2009, pp. 945–978.

[44] Z. Li, L. Itti, Gist based top-down templates for gaze prediction, Proc. Vision Science Society Annual Meeting, 2009, VSS09.

[45] J. Peters, L. Itti, Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention, Proc. CVPR, 2007.