Signal Processing: *Image Communication* 28 (2013) 1006–1022

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Visual saliency guided video compression algorithm

Rupesh Gupta [a], Meera Thapar Khanna [b,*], Santanu Chaudhury [b]

[a] *Department of Computer Engineering at Purdue University, West Lafayette, United States*
[b] *Department of Electrical Engineering, IIT Delhi, India*

ARTICLE INFO

Article history:
Received 14 August 2012
Received in revised form
27 April 2013
Accepted 9 July 2013
Available online 16 July 2013

Keywords:
Visual saliency
Feature maps
Mutual information
QP tuning
Machine learning
Motion vector

ABSTRACT

Recently Saliency maps from input images are used to detect interesting regions in images/videos and focus on processing these salient regions. This paper introduces a novel, macroblock level visual saliency guided video compression algorithm. This is modelled as a 2 step process viz. salient region detection and frame foveation. Visual saliency is modelled as a combination of low level, as well as high level features which become important at the higher-level visual cortex. A relevance vector machine is trained over 3 dimensional feature vectors pertaining to global, local and rarity measures of conspicuity, to yield probabilistic values which form the saliency map. These saliency values are used for non-uniform bit-allocation over video frames. To achieve these goals, we also propose a novel video compression architecture, incorporating saliency, to save tremendous amount of computation. This architecture is based on thresholding of mutual information between successive frames for flagging frames requiring re-computation of saliency, and use of motion vectors for propagation of saliency values.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The H.264 video compression standard has a wide range of applications from low bit-rate internet streaming applications to HDTV broadcast. The incipient HEVC seeks to further improve upon H.264, thereby achieving an even higher coding efficiency. This has been possible through improved exploitation of spatio-temporal prediction and entropy coding. The focus has however remained on minimization of the quantifiable 'objective' redundancy. In this paper we propose a video compression scheme which exploits features of human perception of video content for enhancing compression efficiency of H.264 based coding scheme.

It is now well established that the acuity of the human eye is limited to only 1–2° of visual angle [1]. This implies that when viewed from a recommended distance of 1.2 m, the eye can crisply perceive only a 2 cm radial region (computed as $1.2 \times \tan(2^\circ/2)$) on a standard definition 32" LCD. Also, a recent eye-tracking study [2] on inter-observer saliency variations in task-free viewing of natural images has concluded that images known to have salient regions generate highly correlated eye-fixation maps for different viewers. These have reinforced the application of visual saliency to video compression among numerous other applications, like subjective visual quality assessment, progressive transmission, re-targeting for hand-held devices, effective website designs, and thumb nailing. Saliency based compression is achieved by encoding visually salient regions with high priority, while treating less interesting regions with low priority to save bits. In essence, it allows us to meet our goal of compression without significant degradation of viewing experience or 'subjective' quality.

Since 1998 a vast amount of research has gone into modelling of the human visual system (HVS) for the purpose of video processing. The earlier approaches were gaze contingent, dependant on eye trackers to record points of fixation. These were narrowly restricted to the

* Corresponding author. Tel.: +91 112 656 6167.
E-mail addresses: rupesh.iitdelhi@gmail.com (R. Gupta), meerathapars@gmail.com (M. Thapar Khanna), santanuc@ee.iitd.ernet.in (S. Chaudhury).

0923-5965/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.image.2013.07.003

case of a single viewer with eye-tracking apparatus, and hence not very useful. Later approaches exploited the computational neurobiological models to automatically predict the regions likely to attract human attention. However, each model came with its own merits and shortcomings, leaving salient region detection still a challenging and exciting area of research.

Itti et. al. [3] modelled visual attention as a combination of low level features pertaining to the degree of dissimilarity between a region and its surroundings. Novel center-surround approaches like [4] model saliency as the fraction of dissimilar pixels in concentric annular regions around each pixel. Hou and Zhang [5] take a completely different approach, suppressing the response to frequently occurring features while capturing deviances. Other transform domain approaches like [6,7] follow a similar line of thought. Although these approaches work on psychological patterns with high accuracy, they often fail to detect salient objects in real life images. Some failure cases of these approaches are shown in Fig. 1. It is evident that these saliency maps are not quite close to ground truth.

The failure of these approaches can be attributed to Gestalt's grouping principle [8] which concerns the effect produced when the collective presence of a set of elements becomes more meaningful than their presence as separate elements. Thus, in this work we model saliency as a combination of low level, as well as high level features which become important at the higher-level visual cortex. Many authors like [9] resort to a linear combination of features such as contrast and skin color, but do not provide any explanation for the weights chosen. Hence, we propose a learning based feature integration algorithm where we train a Relevance Vector Machine (RVM) [10,11] with 3 dimensional feature vectors to output probabilistic saliency values.

One of the earliest automated (as opposed to gaze contingent), visual saliency based video compression model was proposed by Itti [12] in 2004. In [12] a small number of virtual foveas attempt to track salient objects over video frames; and non-salient regions are Gaussian blurred to achieve compression. Guo and Zhang [6] use their PQFT approach for proto-object detection, and apply a multi-resolution wavelet domain foveation filter suppressing coefficients corresponding to background. The OPTOPOIHSH project which aimed at region of interest (ROI) based video compression to allow acceptable quality video transmission through low bandwidth channels, also employs some form of blurring [13]. Selective blurring can however lead to unpleasant artifacts and generally scores low on subjective evaluation. A novel bit allocation strategy through quantization parameter (QP) tuning, achieving compression while preserving visual quality, is presented in [14] which we adopt here.

A saliency preserving video compression scheme has been presented in [15], which reduces coding artifacts so that saliency of the region of interest is retained. In [16] a bit allocation strategy has been proposed which is based

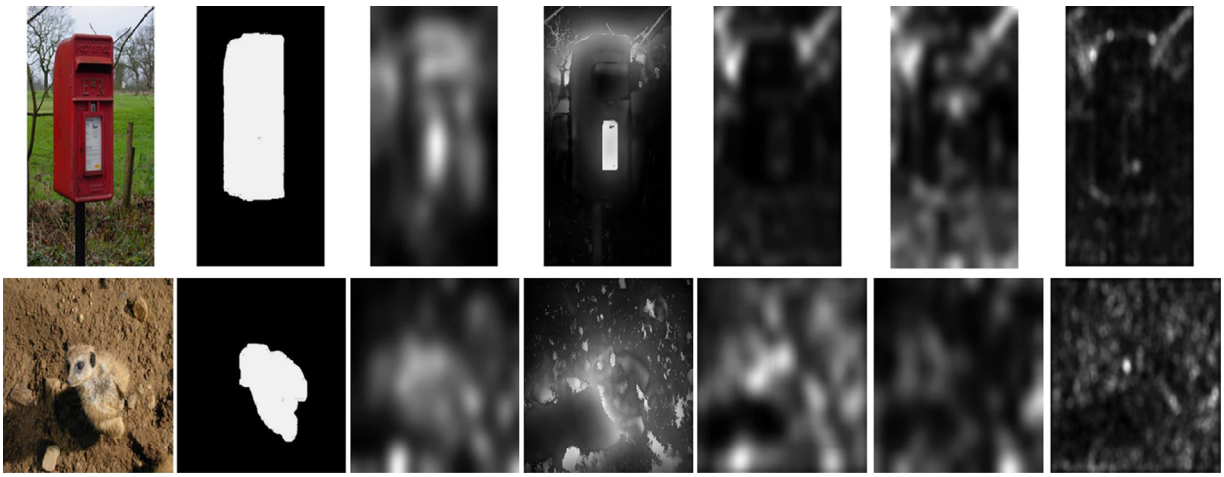

Fig. 1. Some failure examples of existing approaches. Left to right: Original image, ground truth, saliency map obtained from [3–7].

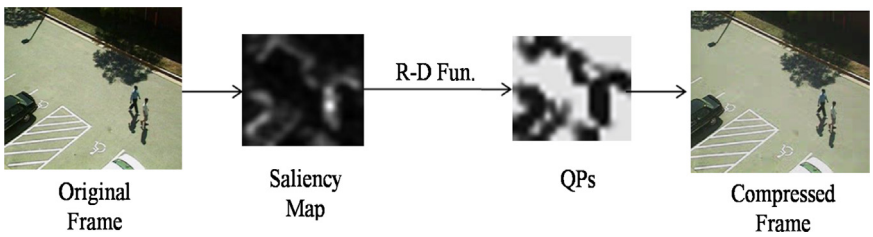

Fig. 2. Flow diagram of our compression approach.

upon evaluation of perceptual distortion sensitivity of each macroblock. In [17] a video coding technique has been proposed which use visual saliency for adjusting image fidelity for compression. They use a saliency computation scheme different from the approach presented in this paper.

A simplified flow diagram of our compression model is shown in Fig. 2. In all the existing compression approaches, the saliency map is computed for each frame which can prove to be computationally very expensive. This is avoidable considering the temporal redundancy inherent in videos. We propose here a video coding architecture, incorporating visual saliency propagation, to save on a large amount of saliency computation, and hence time. This scheme uses thresholding of mutual information (MI) between successive frames for flagging frames which require re-computation of saliency; and use of motion vectors for carrying forward the saliency values.

The contribution of this paper to this field of study is thus twofold. First, a supervised procedure to compute saliency of an image using RVM over 3 dimensional feature vectors, pertaining to global, local and rarity measures of conspicuity is proposed. Second, a video coding architecture aimed at significant decrease in computation, and therefore time, is proposed. To arrive at this architecture, a novel saliency propagation and segmentation scheme based upon MI is implemented.

In this we have used H.264 encoder and decoder for establishing effectiveness of the strategy. However, the same scheme can be used with HEVC encoder and decoder. In HEVC, as in H.264, uniform reconstruction quantization (URQ) is used, with quantization scaling metrices supported for the various block sizes [18]. Accordingly, our QP tuning algorithm can be used with HEVC as well. However, flexibility in block size can permit grouping of MBs based upon saliency values.

The remainder of this paper is organized as follows. In Section 2 we discuss, in detail, the steps involved in our learning based saliency algorithm. Since all video coding operations are MB based, we learn saliency at MB level to save on unnecessary computation. This section also contains some results and comparison of our algorithm with other leading approaches. In Section 3 we describe our complete video coding architecture in which various issues relating to saliency propagation/re-calculation and bit allocation are addressed. Compression result on some varied video sequences and gain over standard H.264 with RDO is presented in Section 4 and conclusions are drawn in Section 5.

2. Our saliency algorithm

We use color spatial variance, center-surround multi scale ratio of dissimilarity and pulse DCT to construct 3 feature maps. Then, a soft, learning based approach is used to arrive at the final saliency map.

2.1. Global conspicuity: color spatial variance

The lesser a particular color is globally present in a frame, the more it is likely to catch the viewer's attention. However, a color sparsely distributed across the entire frame need not be conspicuous owing to Gestalt's principles [8]. Hence, spatial variance of colors can be employed as an appropriate measure of global conspicuity. This has also been employed by [19] and more recently by [20]. We follow the method given in [19]. The steps for generation of saliency map are the following: First, we use k-means clustering to initialize color clusters (starting values of weight, mean and variance) in the image. These clusters are then represented by GMMs and refined using expectation maximization algorithm. If π_c, μ_c, and K_c are the prior, mean and covariance matrix respectively of the cth cluster then the posterior, i.e. the probability that a pixel belongs to a color cluster $p(c|I_{(x,y)})$ is

$$p(c\left|I_{(x,y)}\right.) = \frac{\pi_c N(I_{(x,y)}|\mu_c, K_c)}{\sum_c \pi_c N(I_{(x,y)}|\mu_c, K_c)} \tag{1}$$

Then, the spatial variance $V_x(c)$ of each cluster c along x direction is

$$V_x(c) = \frac{\sum_{(x,y)}[p(c|I_{(x,y)}).|x - M_x(c)|^2]}{\sum_{(x,y)} p(c|I_{(x,y)})}, \tag{2}$$

$$M_x(c) = \frac{\sum_{(x,y)}[p(c|I_{(x,y)}).x]}{\sum_{(x,y)} p(c|I_{(x,y)})} \tag{3}$$

The variance $V_y(c)$ along y direction is similarly calculated. The spatial variance for each cluster c is then $V(c) = V_x(c) + V_y(c)$, which is normalized to the range [0,1]. Finally the feature map is computed as follows and normalize it to the range [0,1]. An example feature map is

Fig. 3. Feature map pertaining to global conspicuity.

shown in Fig. 3.

$$f(x,y)=\sum_{c} p(c|I_{(x,y)}).(1-V(c)) \tag{4}$$

2.2. Local conspicuity: multi-scale ratio of dissimilarity

The pop-out effect has, since long [3] been attributed to the degree of dissimilarity between a stimulus and its surroundings. A simple center-surround method to accurately capture local conspicuity has been recently proposed in [4]. In this method, a multi-scale filter is designed to simulate the visual field. A summation of the fraction of dissimilar pixels in concentric ring-like regions around each pixel gives a measure of conspicuity. We use this method to construct our second feature map. The steps are as follows.

A multi-scale filter contains a series of concentric ring-like regions as shown in Fig. 4. For a pixel p, the ring-like region R_i is defined by

$$R_i(p)=\{q|r_{i-1}<\|p-q\|_2\le r_i,\ q\in\Lambda\},\quad \text{where } i=1,2,..,k \tag{5}$$

As illustrated, r_i is the radius from pixel p to the outer boundary of ring-like region R_i. r_i is set to $r_i=r_{i-1}+\Delta r$ for each of the k ring-like regions where largest radius r_k is taken as $\min(W,H)/4$, W and H being the width and height of image respectively.

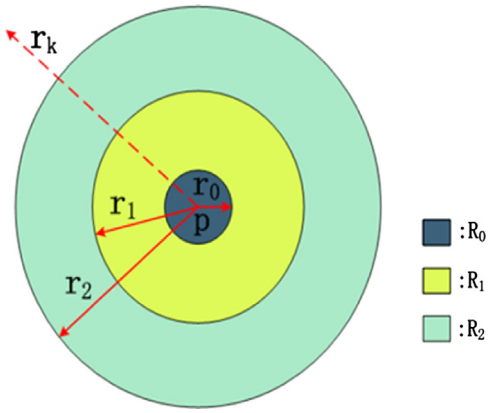

Fig. 4. Illustration of concentric ring-like regions.

Based on these center-surround regions for a pixel, the number of dissimilar pixels are calculated in each of its ring-like regions. In the ith ring-like region, dissimilar pixels are the pixels which satisfy the following equation:

$$D_i(p)=\{q|\sigma_I<\|I_q-M_p^{i-1}\|_2\},\quad q\in R_i(p)\quad \text{where } i=1,2,..,k \tag{6}$$

where, I_q denotes the $L*a*b$ color value of pixel q, $\|.\|$ denotes the Euclidean distance, and M_p^{i-1} denotes the average $L*a*b$ color value of those pixels similar to p, in the ring-like region $R_{i-1}(p)$. M_p^{i-1} is defined by

$$M_p^{i-1}=\left\{\begin{array}{c}\frac{\sum_{q\in R_{i-1}(p)\backslash D_{i-1}(p)} I_q}{|R_{i-1}(p)\backslash D_{i-1}(p)|_{card}}\\ I_p \text{ if } D_{i-1}(p)=R_{i-1}(p)\end{array}\right\} \tag{7}$$

where $|R_{i-1}(p)\backslash D_{i-1}(p)|_{card}$ denotes the number of pixels similar to pixel p in the region $R_{i-1}(p)$. Threshold value σ_I is set as

$$\sigma_I=\sqrt{\sigma_L^2+\sigma_a^2+\sigma_b^2} \tag{8}$$

where σ_L^2, σ_a^2, σ_b^2 are respectively the variances of the three channels of $L*a*b$ color space over the whole input image. Lab color space is used to approximate human vision. Its components closely matches human perception of lightness. Simple Euclidean distance could differentiate color perceptually in Lab color space. Finally, the feature value $f(x,y)$ for pixel p is obtained by summing the ratios of number of dissimilar pixels to the total number of pixels in the corresponding region, over the multiple ring-like regions as follows:

$$f(x,y)=\frac{1}{k}\sum_{i=1}^{k}\frac{|D_i(p)|_{card}}{|R_i(p)|_{card}} \tag{9}$$

The results produced are promising with $k=4$ and $r_0=3.5$. However, this approach is slow, since a large number of computations and comparisons are carried out for every pixel. Noting that background pixels generally have very low values of saliency, computation of feature value for these pixels is superfluous. If these computations are avoided then the time complexity can be significantly reduced. To this end, we first run the SIFT [21] algorithm and locate the key points on the image which are salient not only spatially but also across

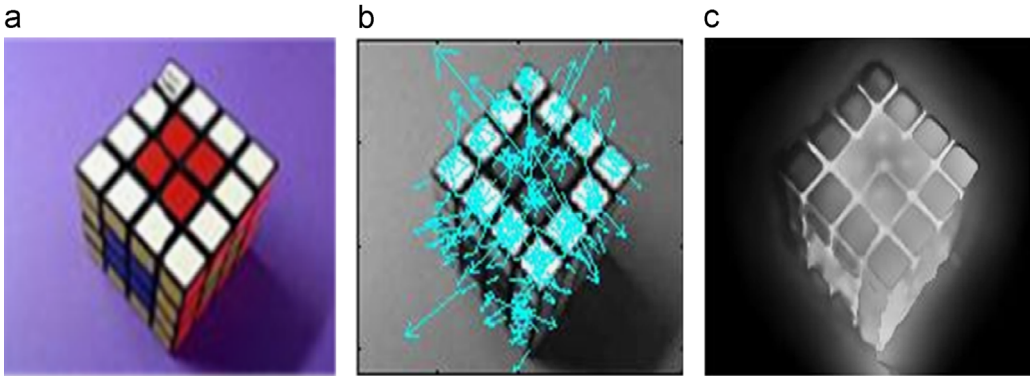

Fig. 5. (a) Image, (b) key points located using SIFT and (c) local conspicuity feature map.

Fig. 6. PCT algorithm applied to original and blurred images. Notice that the feature map corresponding to blurred image is sparser.

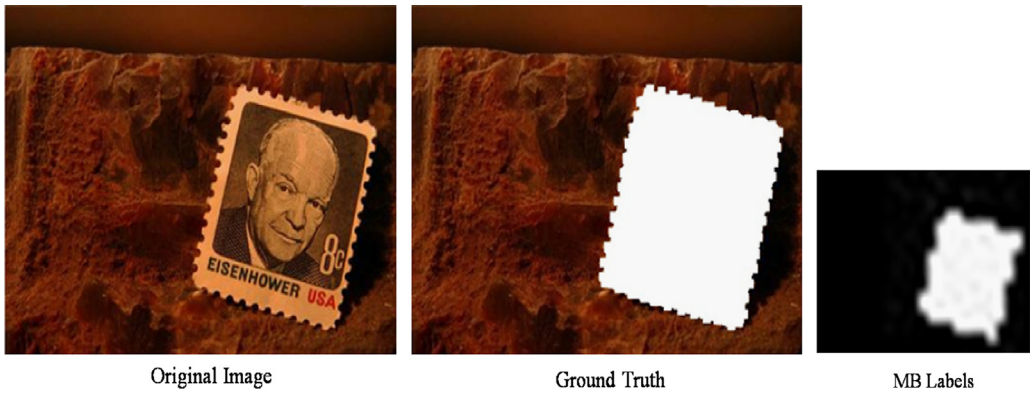

Fig. 7. Macroblock level ground truth preparation.

different scales. We take one key point at a time and compute its feature value using Eq. (9). If feature value of key point $> T_1$ (0.4 here, required since a key point may lie on a cluttered background), we start growing a region from that point. The feature value of neighbouring pixels is used as region membership criterion and all pixels visited are marked so that they are not re-visited when a different seed point is chosen. We stop when the feature distance between the new pixel and region mean exceeds $> T_2$ (0.2 here). The thresholds T1 and T2 are determined experimentally using images containing differing illuminations for the same objects. This feature map is also normalized to [0,1]. An example is shown in Fig. 5.

2.3. Rarity conspicuity: pulse discrete cosine transform

A simple, real time model simulating lateral inhibition in the receptive field has been proposed in [7]. This approach lays emphasis on its biological plausibility as no complex number computations are involved which cannot be carried out in human brain. It has also been shown to outperform other transform domain approaches like [6] both in terms of speed as well as accuracy over psychological patterns. This approach is based on the following observation. DCT represents the visual input with periodical signals of different frequency and different amplitude. Therefore, large coefficient of DCT contains the

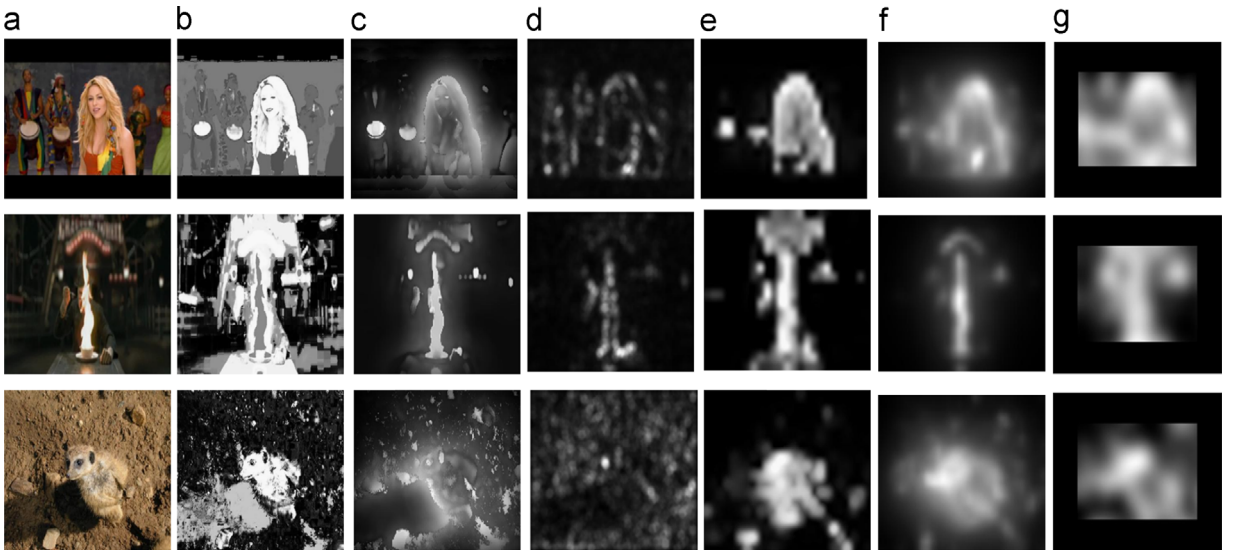

Fig. 8. (a) Input image, (b) global, (c) local [4] and (d) rarity [7] feature maps, (e) our resized saliency map, (f) saliency map obtained from [25] and (g) [26].

Fig. 9. Some more results. Left to right for each set: Original image, ground truth, resized saliency map obtained from our saliency algorithm.

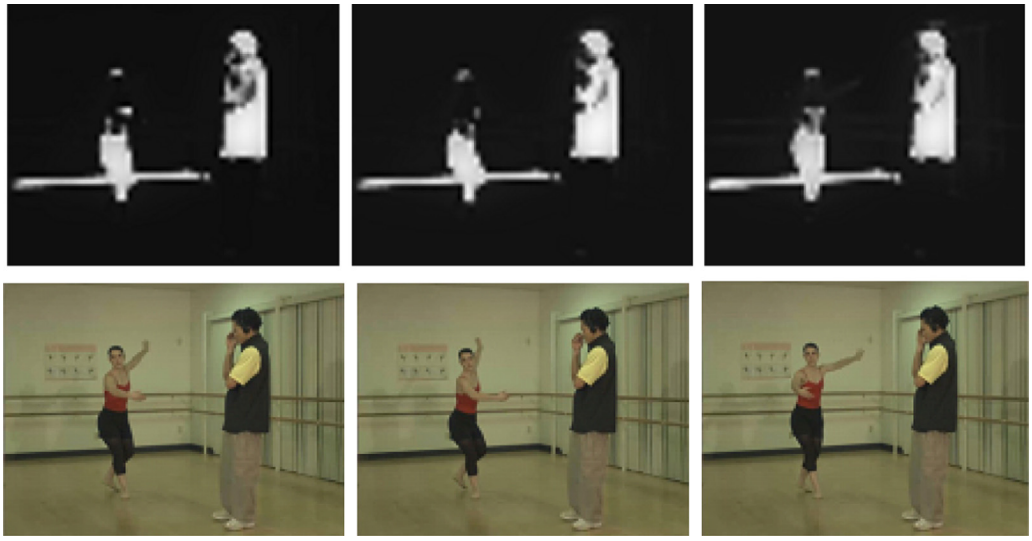

Fig. 10. Consistency between successive frames saliency maps and coded results for dance sequence.

Fig. 11. Consistency between successive frames saliency maps and coded results for ESPN sequence.

information of statistical homogeneity. Pulse DCT (PCT) only retains sign of the DCT coefficients (Eq. (10)). By flattening the magnitude, PCT simulates the lateral suppression among neurons with similar features.

RGB color space is used to compute feature map as it allows a wide range of colors. The feature map is computed as follows. If M_I, M_R, M_G, M_B are the intensity and broadly tuned color channels, then the feature map F_x for each channel is calculated as

$$P_x = sign(\mathfrak{C}(M_x)) \tag{10}$$

$$F_x = abs(\mathfrak{C}^{-1}(P_x)) \tag{11}$$

where $\mathfrak{C}$ = DCT, $\mathfrak{C}^{-1}$ = inverse DCT, $sign(.)$ = signum function. To balance all original maps, the weight factor of each feature map is calculated as $w_x = max(M_x)$ and then the maps are combined to obtain a cumulative feature map fM as

$$F = w_R F_R + w_G F_G + w_B F_B + w_I F_I \tag{12}$$

$$fM = G * F^2 \tag{13}$$

We apply the PCT algorithm to smoothed images to produce our rarity feature map. We hypothesize that a Gaussian blurred image simulates the scene viewed from a distance, and thus finer edge details in a cluttered background are not noticed leading to a sparser feature map. An example is shown in Fig. 6. This map is also normalized to the range [0,1].

2.4. Learning to integrate the feature maps

The steps followed for combining the 3 feature maps are as follows. First, we selected 30 images, of size 300×400,

encompassing the success and failure cases of each of the 3 feature maps. Five viewers were asked to mark each part of the image they considered salient. In accordance with [2], our images (mostly taken from MSRA database [22]) had well-defined salient regions and hence the markings turned out to be exactly the same for almost all images.

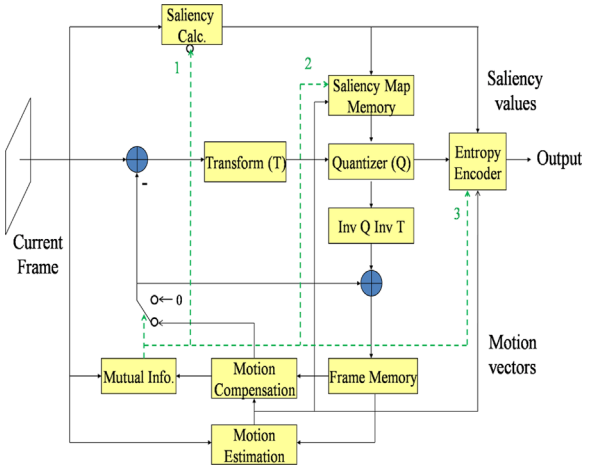

Fig. 12. Our video compression architecture incorporating saliency propagation.

A majority rule was applied for the rest. We could not use the ground truth of the MSRA database as they provide only rectangular bounding salient regions whereas we required the exact salient region boundaries. We came across a publicly available ground truth database prepared by Achanta [23] for the MSRA image set. This database closely matches our own ground truth, establishing the credentials of our experiments. Then, an MB level, 3 dimensional training data (total 450×30 points) was prepared taking average values of each of the 3 feature maps over each MB of size 16×16. It is important to note that since all video coding operations are MB based, we learn saliency at MB level to save on unnecessary computation. A target class label '1' was assigned to an MB if more than half of the pixels of that MB were marked salient; else class label '0' was assigned. This ground truth preparation is illustrated in Fig. 7.

Next, we trained an RVM over this training data as a binary classification problem [10]. Here we must point out that we are not really interested in a binary label (salient/non-salient) but the relative saliency value of each MB which will later be used for bit allocation. A potential advantage of RVM over SVM, which is desired here, is that it provides posterior probabilities. Also, RVM has better generalization ability and its sparser kernel function leads to faster decisions [24]. The probabilistic outputs of the RVM for data points from unseen images formed our final saliency map. The saliency map is also obviously obtained at a 16×16 level.

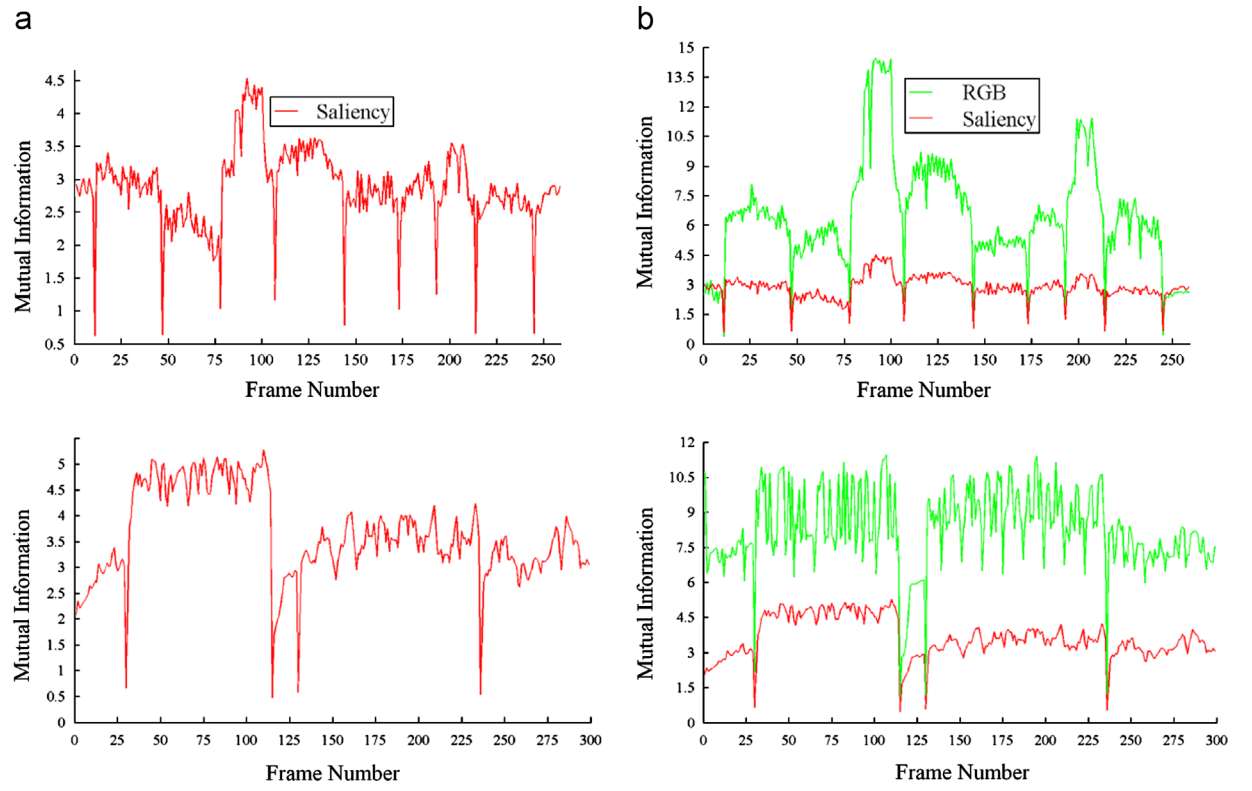

Fig. 13. (a) MI plot for saliency maps, (b) MI plots of RGB and saliency overlaid. Above: Airtel ad sequence with 9 cuts. Below: ESPN news sequence with 4 cuts.

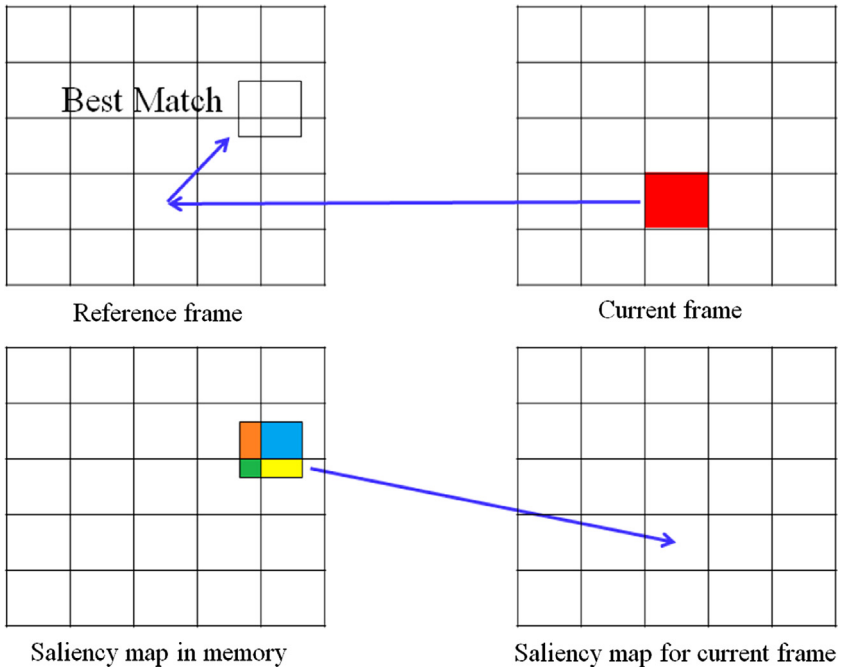

Fig. 14. Image illustrating a weighted averaging of saliency values, the orange, blue, yellow, green colors denote the amount of overlap and hence weights.

36	36	36	36	36	35	36	36	36	36	36
36	36	36	36	25	23	27	36	36	36	36
36	36	32	36	23	23	23	36	36	36	36
36	36	36	27	23	23	23	36	36	36	36
36	36	36	31	23	23	23	35	36	36	36
36	36	36	27	23	23	23	23	27	36	36
36	34	36	23	23	23	23	23	23	36	36
36	36	31	23	23	23	23	23	23	34	36
36	36	23	32	23	23	23	24	23	36	36

Fig. 15. Illustration of QPs for 16×16 MBs for Claire sequence. QP $= 24$.

2.5. Results

To test the scheme, we generated testing data from 120 images of size 300×400 (450×120 points) and evaluated the saliency maps obtained against ground truth. Achanta's ground truth [23] overlaps with only 60 images in our dataset since he prepared it for images randomly sampled from the MSRA database. Some results and comparisons with [25] and [26] are shown in Fig. 8. A comparison with [4] and [7] is implicit in these results as our local and rarity feature maps respectively. It is apparent that our approach is better or at least at par with these other high-ranking approaches. Some more results are shown in Fig. 9.

3. Our compression architecture

We wish to employ saliency for the purpose of video compression. However, computation of feature maps for each video frame can prove to be computationally very expensive if we rely on video compression techniques such as those proposed in [6,12,14] as they necessitate calculation of saliency map of each frame.

There exist consistency between salient regions for successive frames. Figs. 10 and 11 are illustration of this characteristic observed in video sequence. So we propose here the use of temporal redundancy inherent in videos to propagate saliency values. Ideally the saliency map should be re-calculated only when there is a large change in saliency. However, to measure this change, we require the saliency for the next frame which is unavailable. Hence, we also propose a workaround to identify the frames for which re-computation of saliency map is indispensable. A block diagram of the architecture is shown in Fig. 12 which is discussed in detail in the following subsections.

3.1. Relation between MI and saliency

First, we describe the need for the MI (Mutual Information) computation unit in the architecture as shown in Fig. 12. The idea is that we perform a re-calculation of saliency map on the basis of MI value between successive frames. An elegant information theoretic shot detection algorithm has been proposed by Cernekova et al. in [27] and an improved version of the same using motion prediction is presented in [28]. The authors compute the MI between consecutive frames and argue that a small value of MI indicates existence of a cut, which is quite self-explanatory. MI is computed as follows.

If a video sequence has gray levels varying from 0 to N-1, at every frame f_t three $N \times N$ matrices $C^R_{t,t+1}, C^G_{t,t+1}$ and $C^B_{t,t+1}$ are created carrying information on the gray level transitions between frames f_t and f_{t+1}. For the R component, the element $C^R_{t,t+1}(i,j)$, with $0 \le i \le N-1$ and $0 \le j \le N-1$, corresponds to the probability that a pixel with gray level i in frame f_t has gray level j in frame f_{t+1}. Hence, $C^R_{t,t+1}(i,j)$ is the number of pixels which change from gray level i in frame f_t to gray level j in frame f_{t+1}, divided by the number of pixels in the video frame. The mutual information $I^R_{t,t+1}$ for the R component is then calculated as

$$I^R_{t,t+1} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} C^R_{t,t+1}(i,j) log \frac{C^R_{t,t+1}(i,j))}{C^R_t(i) C^R_{t+1}(j)} \quad (14)$$

The total MI is defined as

$$I_{t,t+1} = I^R_{t,t+1} + I^G_{t,t+1} + I^B_{t,t+1} \quad (15)$$

In our case, we take f_{t+1} as the current frame and f_t as the motion compensated previous frame (also called predicted frame). A motion compensated frame undoes the effect of motion, leading to higher MI for same scene frames by amending the $C^X_{t,t+1}(i,j)$ matrices. We experimented with this method over some video sequences, with saliency map of each frame pre-computed, and plotted the MI distributions for color as well as saliency. MIs for an Airtel ad sequence with 9 scene changes and an ESPN news sequence with 4 scene changes are plotted in Fig. 13. One can immediately point-out the shots by locating steep valleys in the plots. It is apparent that not only does this method effectively capture changes in saliency as shown in Fig. 13(a), but also that the RGB and saliency plots follow a very similar distribution as evident from Fig. 13(b).

We briefly reason out our choice of MI as the suitable indicator of saliency changes. From Eq. (14) it can be observed that the $C^X_{t,t+1}(i,j)$ matrix takes into account the spatial correlation between pixels in the consecutive frames. Since we are considering motion compensated previous frame, this matrix can drastically change, causing a significant drop in MI, only when there is an abrupt addition/removal of an object to/from the same scene or in case of scene change. In both cases, the saliency map ought to be re-computed. Other popular scene change detection algorithms [29] cannot mirror saliency as well. Color histogram difference is inappropriate as it cannot distinguish images with different structures but similar color distribution, and edge change ratio does not take into account color, a primary visual cue for saliency.

Table 1
Video sequences.

Sequence	Frame size	# Frames	Raw size (MB)
Dance	1024×768	100	115.2
Airtel	480×360	400	101.2
ESPN	448×256	300	49.2
Claire	176×144	494	18.3
Coastguard	176×144	300	11.1

3.2. Re-computation of saliency values

Fig. 13(b) implies that we can detect the frames requiring re-computation of saliency by calculating MI over the color channels. The frame where a large change is detected should be coded in intra (I) mode; and saliency re-computed for this frame and stored in the saliency map memory of Fig. 12. Now the use of motion compensated frame becomes even more meaningful, as re-computation of saliency would be futile if there is a drop imputable to motion in MI between the original frames. The method has been found to work best on natural video sequences, with no special effects and animations, as they yield MI plots with steep valleys and easily identifiable thresholds for shot declaration.

3.3. Propagation of saliency values

For predictive (P) coded frames, we make use of motion vectors to approximate saliency values. In the motion estimation stage of video compression, an MB is selected in the current frame and a best match is searched for in the reference frame. This best match may or may not exactly overlap an MB in the reference frame, but we have the saliency values for only non overlapping 16×16 MBs. Therefore, we take a weighted average of the saliency values of each of the MBs under the best match region in the reference frame, as the saliency value for the MB in current frame. The weights correspond to the amount of area overlap as shown in Fig. 14.

Returning back to the architecture of Fig. 12, the MI computation unit has to thus provide input to 3 other units dependant on the coding mode chosen (I/ P) for the current frame. First, it directs the saliency calculator whether to compute saliency (I) or not (P). Second, it notifies the saliency map memory whether to use motion vectors (P) or not (I) and third, it directs the entropy encoder if the motion vectors are to be encoded and

Table 2
Compression results with Gaussian blurring and proposed method. QP=24.

Sequence	H.264 with RDO			Gaussian blurring			% Gain	Our result			% Gain
	File size	PSNR	Rate	File size	PSNR	Rate		File size	PSNR	Rate	
Dance	523 KB	52.51	1286.56	435 KB	51.54	1070.51	16.8	410 KB	52.51	1008.87	21.6
Airtel	820 KB	50.06	503.99	695 KB	44.69	427.41	15.2	748 KB	50.14	460.55	8.9
ESPN	1.26 MB	47.58	1063.03	824 KB	45.74	674.29	34.6	1.0 MB	46.95	927.88	20.6
Claire	122 KB	43.38	60.42	115 KB	42.00	56.93	5.7	117 KB	43.51	55.03	4.0
Coastguard	557 KB	45.74	456.10	251 KB	36.48	205.30	54.9	391 KB	45.76	320.52	29.8

Table 3
Compression results with Gaussian blurring and proposed method. QP=28.

Sequence	H.264 with RDO			Gaussian blurring			% Gain	Our result			% Gain
	File size (KB)	PSNR	Rate	File size (KB)	PSNR	Rate		File size (KB)	PSNR	Rate	
Dance	280	57.44	684.16	246	50.71	604.15	12.1	234	51.46	575.49	16.4
Airtel	422	47.91	258.91	383	45.56	234.74	9.2	409	47.76	251.49	3.0
ESPN	673	44.77	551.76	449	43.86	367.28	33.3	509	44.77	417.02	24.3
Claire	67	41.59	33.01	64	40.71	31.73	4.5	74	41.86	36.75	−10.4
Coastguard	291	44.10	237.80	150	36.39	122.21	48.4	218	43.77	178.91	25.08

Table 4
Compression results with Gaussian blurring and proposed method. QP=30.

Sequence	H.264 with RDO			Gaussian bluring			% Gain	Our Result			% Gain
	File size (KB)	PSNR	Rate	File size (KB)	PSNR	Rate		File size (KB)	PSNR	Rate	
Dance	215	50.88	527.77	193	50.23	427.99	10.2	184	50.88	451.78	14.4
Airtel	305	44.80	187.04	285	44.12	174.77	6.5	300	46.77	187.0	1.6
ESPN	447	43.29	365.89	314	42.73	256.80	29.7	378	44.76	309.02	15.4
Claire	50	40.79	24.51	48	40.08	23.58	4.0	55	41.06	27.18	−10.0
Coastguard	194	43.12	158.89	110	36.32	89.64	43.2	149	43.80	121.37	23.19

Table 5
Compression results with Gaussian blurring and proposed method. QP=32.

Sequence	H.264 with RDO			Gaussian blurring			% Gain	Our result			% Gain
	File size (KB)	PSNR	Rate	File size (KB)	PSNR	Rate		File size (KB)	PSNR	Rate	
Dance	169	50.22	414.41	153	49.70	374.97	9.4	158	50.26	387.18	6.5
Airtel	226	45.68	138.30	214	45.69	131.45	5.3	238	45.70	145.94	−5.0
ESPN	299	41.87	244.61	223	41.57	182.36	25.4	275	42.12	225.13	8.0
Claire	38	39.85	18.49	36	39.30	17.83	5.2	42	40.11	20.54	−10.5
Coastguard	130	42.16	105.97	81	36.23	65.88	37.6	111	42.93	90.59	14.6

Table 6
Comparison of performance with VDSI [16] and proposed method for Football sequence.

QP	H.264 without RDO		VDSI [16]		Our result	
	PSNR	Rate	PSNR	Rate	PSNR	Rate
22	40.45	2690	38.90	2346	40.07	1992
28	36.03	1206	35.06	1220	37.56	986
30	34.62	1065	33.84	973	36.93	781
32	33.27	830	32.63	766	36.34	631

Table 7
Comparison of performance with VDSI [16] and proposed method for Stefan sequence.

QP	H.264 without RDO		VDSI [16]		Our result	
	PSNR	Rate	PSNR	Rate	PSNR	Rate
22	40.34	3250	36.18	1930	40.21	2500
28	35.60	1408	32.56	910	36.12	1000
30	33.93	1027	31.40	708	35.15	770
32	32.35	741	30.21	548	34.24	560

transmitted (P) or not (I). The only constituent of Fig. 12 which remains unexplained as yet is the input from saliency map memory to the quantizer. This is elucidated in the next section.

3.4. Selection of quantization parameters

Once the saliency map is obtained, bits may be non-uniformly distributed across a frame, allocating more bits to the salient regions and lesser to those regions which are not attended to by the human eye. A number of authors have proposed bit allocation strategies aimed at achieving maximum perceptual quality at fixed bandwidth or bit-rate. A region weighted rate distortion function has been proposed in [30] and a method based on target bit count assignment for ROI and NROI used over a quadratic rate-quantization model is presented in [9]. We however wish to use saliency for compression purpose. A very simple, widely adopted approach to compression is selective blurring, as adopted by [12,13]. The idea here is that a threshold is applied to the saliency map to get the salient regions and then the non-salient regions are Gaussian blurred, reducing high frequency content and hence achieving compression. Selective blurring gives high compression but yields quite obvious distortions in low-saliency regions. Hence we adopt a softer technique of tuning the QP for each MB in accordance with its saliency. This has been shown to give better subjective visual

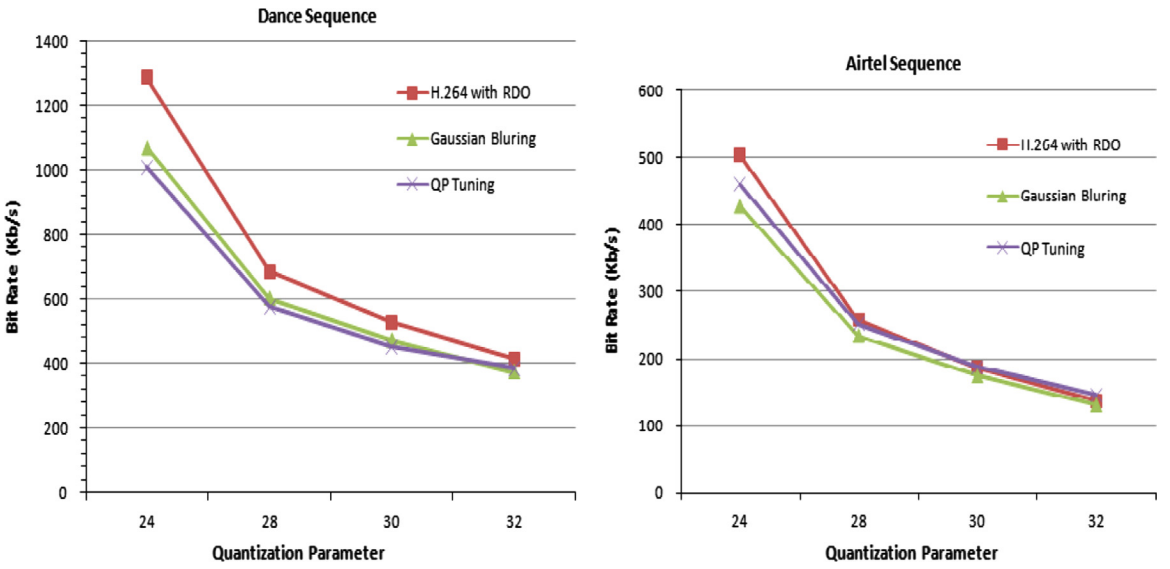

Fig. 16. Comparison between rate and distortion for dance and Airtel sequence.

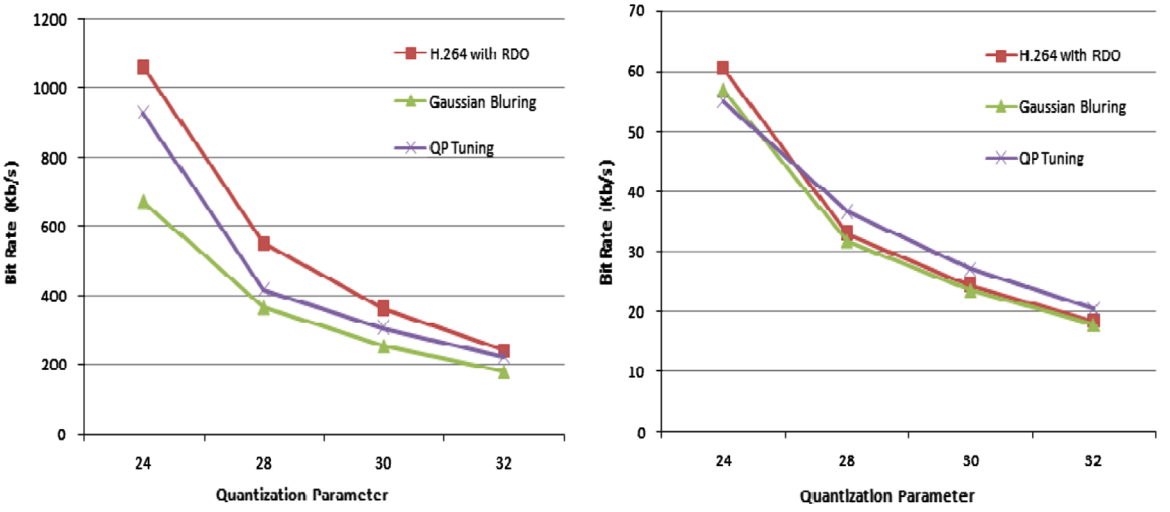

Fig. 17. Comparison between rate and distortion for ESPN and Claire sequence.

quality under the same bandwidth compared to selective blurring. For this, we require a function which can optimally tune the QPs of salient and non-salient MBs to achieve compression, i.e, reduce rate (R) without any significant loss of perceptual quality, i.e, constant distortion (D). In [14], this is posed as a global optimization problem and solved using the method of Lagrange multipliers. The final result for quantization step Q_{istep} for the ith MB having a saliency value w_i is given as

$$Q_{istep} = \frac{Ws}{w_i S} Q_{step} \tag{16}$$

where W is the sum of saliency values over all MBs, s is the area of MB_i (16×16 here), S is the area of entire frame and Q_{step} is a fixed value depending on the amount of distortion tolerable. This formula implies that the quantization step size should be inversely proportional to the saliency value which is completely justified since we would like to allocate more bits (and hence smaller quantization step size) for a salient (high weight coefficient) MB.

We present here a short verification of how this formulation achieves compression without compromising on perceptual quality. Assuming a R–D function [31] for an MB_i is given by

$$D_i = \sigma_i^2 e^{-\gamma R_i} \quad \text{or} \quad R_i = \frac{1}{\gamma} log\left(\frac{\sigma_i^2}{D_i}\right) \tag{17}$$

where σ_i^2 is variance of encoding signal and γ is a constant coefficient. Ignoring the constant term γ and taking $\sigma_i^2 = 1/\alpha$ we get

$$R_i = log\left(\frac{1}{\alpha D_i}\right) \tag{18}$$

Now, the average rate R is calculated as $\sum_{i=1}^{N} sR_i/S$, where N is the number of MBs. Noting that $D_i \propto Q_{istep}$, we

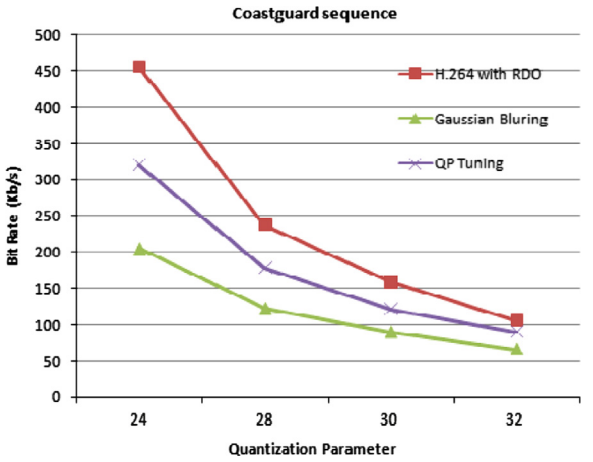

Fig. 18. Comparison between rate and distortion for Coastguard sequence.

get after replacing Q_{istep} by Eq. (16):

$$R = \frac{s}{S}\log\left(\left(\frac{S}{\alpha s Q_{step}}\right)^{N}\frac{w_1.w_2\cdots w_N}{(w_1+w_2+\cdots+w_N)^N}\right) \tag{19}$$

which reduces to

$$R = \frac{Ns}{S}\left[log\left(\frac{1}{\alpha Q_{step}}\right) + log\left(\frac{(w_1.w_2\cdots w_N)^{1/N}}{w_1+w_2+\cdots+w_N}\right) + \log\left(\frac{S}{s}\right)\right] \tag{20}$$

From the above equation it is clear that the first term denotes the rate if every MB was quantized with the same parameter Q_{step}, the second term is always ≤ 0 by the AM-GM inequality and the third term is a constant. Thus $R$is reduced. It can also be readily observed from Eq. (16) that overall D ($\sum w_i D_i/W$) remains constant.

4. Results

We present here the results for videos compressed in two different ways. First is through Gaussian blurring of raw video frames as front end pre-processing to the standard H.264/AVC JM encoder [32], leaving the encoder design intact. Second is through encoding of the theory of Section 3 into the JM reference software. A comparison of the video sizes, PSNR and Bit Rate obtained (in .264 format) is made against the standard JM output with rate distortion optimization (RDO) turned ON for 5 video sequences. A comparison of the PSNR and Bit Rate obtained (in .264 format) is made against the standard JM output with rate distortion optimization (RDO) turned OFF, with VDSI (Visual Distortion Sensivity Index) [16] and with the saliency based fidelity adaptation scheme [17] for 2 video sequences. Subjective quality assessment in the form of mean opinion scores (MOS) is also presented in the following sections.

4.1. Gaussian blurring

This is a simplest and most widely adopted approach to attentional video compression. In [12], a few virtual foveas track salient objects and then a Gaussian blur is applied across the frame with the amount of blur increasing with distance from these salient spots. We apply a Gaussian blur to each pixel with the amount of blur inversely proportional to the saliency of that pixel. Specifically, we take $\overline{S} = 1-S$. If $\overline{S} \leq T$ (0.3 here) then the region is left untouched. For regions with $\overline{S} > T$ we use $\sigma = 10 \times \overline{S}$ as the variance of the Gaussian filter. Larger the $\overline{S}$ value, larger is the variance, ranging from $\sigma = 3$ for most salient (among remaining) to 10 for least salient.

4.2. QP tuning

As mentioned in Section 3.4, the formula for quantization step size Q_{istep} for each MB is given by

$$Q_{istep} = \frac{Ws}{w_i S} Q_{step} \tag{21}$$

Now, in the extreme case, if saliency for an MB is equal to 0, then $Q_{istep} \to \infty$. However, H.264 allows only 52 different values of QP (0–51) and only the 52 corresponding values of Q_{step} (0.625–224) [33]. So Q_{istep} cannot go beyond 224. Also, MB encoding at $Q_{istep} = 224$ (QP=51) gives rise to unpleasant artifacts in the form of blockiness and hence this extreme must be avoided. We limit Q_{istep} to 40 (QP=36) to ensure maintenance of perceptual quality. However this might lead to an unbalance in the equations of overall distortion and average rate (Eq. (20)). Hence we also limit the minimum value of Q_{istep} or QP. In our experiments the minimum value of QP is set to one less than the QP corresponding to Q_{step}.

We smoothen the saliency map before computation of quantization steps. This serves two purposes, first, it ensures that the salient objects/regions are covered completely and second, it ensures a smooth transition from salient to non-salient regions. An example of the QPs computed for a frame of the Claire sequence is shown in Fig. 15.

The results for 5 video sequences of Table 1 are shown in Tables 2–5. The results for H.264 have been generated for the following encoder configuration. Baseline profile (I and P picture types only, interlace, per-picture adaptive frame/field, in-loop de-blocking filter, 1/4-sample motion compensation, tree-structured motion segmentation down to 4×4 block size, CAVLC), RDO turned ON and frame rate of 30 fps. The results for Gaussian blurring and QP Tuning have also been generated for the baseline profile encoder configuration with RDO tuned off.

The results for 2 video sequences are shown in Tables 6 and 7. These results have also been generated for the baseline profile encoder configuration with RDO tuned off.

Table 8
MOS scores.

Sequence	H.264 with RDO	Gaussian blurring	QP tuning
Dance	4.3	3.77	4.5
Airtel	3.8	2.6	3.85
ESPN	4.25	2.85	4.0
Claire	4.2	3.65	4.2
Coastguard	3.7	2.7	3.7

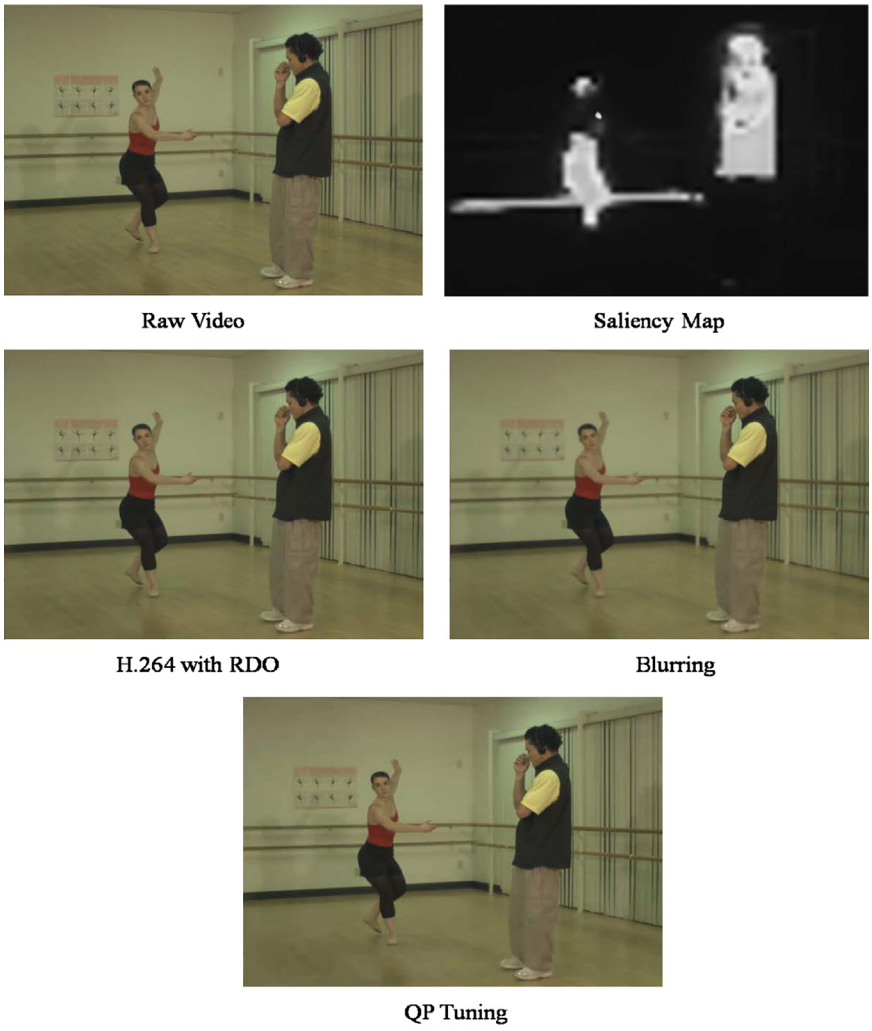

Fig. 19. Coded results for dance sequence.

4.3. Discussions

It can be observed from Tables 2–5 that a compression gain between 1.6% to as high as 29.8% over H.264, can be achieved with our algorithm. The compression gain increases with a decrease in QP. This is because at lower QP, there is a larger difference between the QP values corresponding to background regions in QP tuning method and H.264 with RDO method. Similar argument holds for Gaussian blurring. The gain achieved with QP tuning is at least as much as that achieved with blurring for most cases. It is important to realize here that in QP tuning approach, an additional overhead is incurred in encoding of the different QP values for each MB. There is no such overhead in the other approach. Some comments on the large variation in gain for the 5 sequences chosen are as follows. The gain for Claire (QCIF) sequence is less because it is a talking head sequence with a smooth, static background where neither blurring, nor quantization can benefit as there is already very little high-frequency non-salient content. Gain is in fact negative for QP tuning approach. This can explained by the equation unbalance discussed in Section 4.2 and encoding of the salient regions, which occupy a significant proportion of frame area, at a QP lower than the QP corresponding to Q_{step}. ESPN and Airtel are multiple scene change sequences where the first frame following each shot is coded as an I frame, and the rest all as P frames. The background varies a lot, from smooth to clutter over the various scenes. However for Airtel, the salient object/region occupies majority of the frame area in most of the frames leading to lower gains. Coastguard is a single scene sequence with fair amount of background clutter. Dance is a high resolution (XGA), single scene sequence with fair amount of background clutter. Needless to say, the gain improves with an increase in the background clutter and the ratio of the area of non-salient to salient regions.

As an objective measure PSNR is used. PSNR is computed only for the salient region of the video frames and compared with the H.264 with RDO and Gaussian blurring method. As we can see from the table PSNR of QP Tuning method is high as compared to Gaussian blurring and

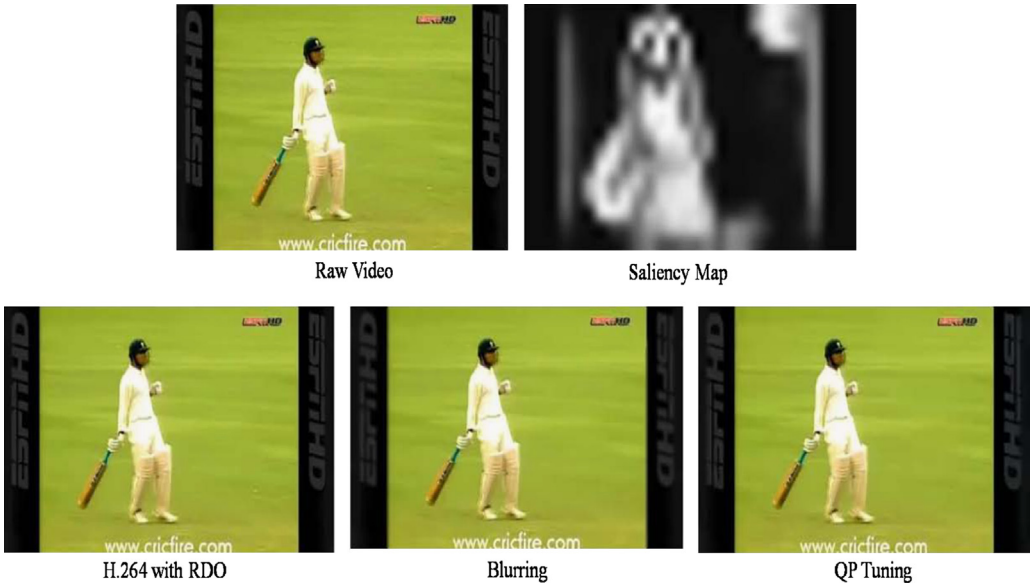

Fig. 20. Coded results for ESPN sequence.

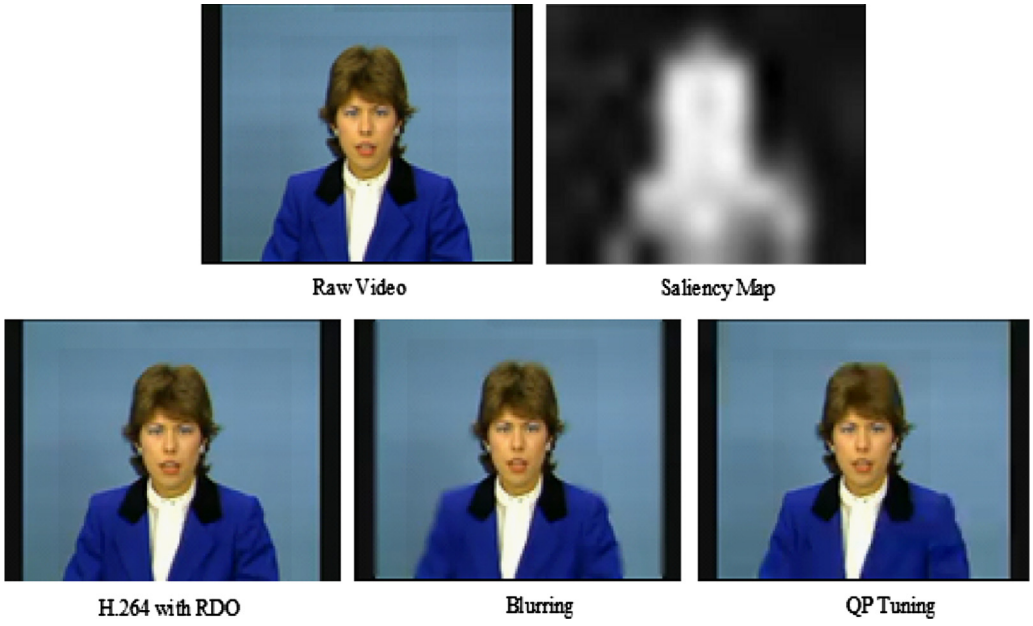

Fig. 21. Coded results for Claire sequence.

same as H.264 with RDO method. The overall bit rate is less as compared to other methods. As more bits are allocated to salient region and less number of bits to non salient regions so there is no perceptual quality degradation. It can be observed from Tables 6 and 7 that the PSNR is higher than the H.264 with RDO tuned off and VDSI [16] scheme for the sequence used in VDSI [16]. As compared to saliency based fidelity adaptation [17] scheme for Football sequence at QP 28, PSNR is high and there is 12% gain in bit rate also by the proposed method. Although saliency based fidelity adaptation [17] scheme requires saliency detection and saliency map computation but our method computes more effective saliency maps using three features Global, Local and Rarity.

Figs. 16–18 gives the comparison between the distortion and rate per video using uniform bit allocation (constant QP), Gaussian blurring scheme and QP Tuning scheme for 5 video sequences. The rate curve resulted from proposed scheme shows that at same QP value bit allocation is less without any loss in perceptual quality. The subjective quality assessment MOS scores are given in Table 8. The scores in this table are the average MOS scores obtained from 15 viewers totally in-cognizant of the project. The viewers were allowed to see the sequences

multiple times and rate them for their visual quality on a scale of 1–5, 1 being highly unpleasant and 5 being very pleasant. It is evident that Gaussian blurring scores very poorly (unacceptably low) on MOS. Viewers have assigned the least MOS score to the Gaussian blurred Airtel sequence. This is majorly because the sequence contains text in every scene which starts appearing a few frames after the shot and hence does not get marked as salient. A pronounced blurry text immediately tends to attract a low subjective score. On the other hand, QP tuning scores significantly high on MOS receiving scores similar to H.264 with RDO in most cases, and even higher in one. This might be due to the fact that in this approach, the visually salient regions are encoded at a lower QP than the fixed QP in standard H.264, resulting in better visual quality of these regions. We may thus assert that QP tuning clearly outperforms variable Gaussian blurring, and successfully achieves our goal of compression without degradation of viewing experience (Figs. 19–21).

5. Conclusions

A vast amount of research has gone into modelling of the HVS with each model having its own merits and shortcomings. The potential which lies in an integration of these models has been demonstrated by the accuracy of our results. A simple and effective learning based approach for such a unification has been presented where an RVM trained over 3 dimensional feature vectors pertaining to global, local and rarity measures of conspicuity outputs probabilistic values which form the saliency map. Though we make use of only 3 features, this model is easily extendible to more features if desired. Here we have avoided the use of some obvious high level features like skin color and face recognition so as to not bias the learning toward a particular feature and keep the algorithm generic. However, such features, which attract human attention more than anything else can be easily incorporated any time, if required. We computed saliency at MB level to save computation, however our model is equally applicable at pixel level. Our model outperforms various other eminent approaches in terms of accuracy of detection and concedes a very low false negative rate at the cut-off point of the ROC curve. This makes our algorithm perfectly befitting the purpose of video compression.

A compression framework approximating the saliency of P frames, saving a lot of computation and speeding-up compression has been proposed. Thresholding of MI computed over low level features between successive frames indicates the frames requiring re-computation of saliency and their subsequent coding mode as intra. The motion vectors computed during motion estimation stage propagate the saliency values for MBs in P frames. The saliency map directs bit allocation over frame MBs commensurate to saliency values. A significant amount of video compression gain over H.264 has been achieved through 2 different approaches. QP tuning emerges as the clear winner outdoing variable Gaussian blurring both in terms of % gain, as well as subjective visual quality assessment of compressed video sequences.

References

[1] M. Nicolaou, A. James, A. Darzi, G.-Z. Yang, A study of saccade transition for attention segregation and task strategy in laparoscopic surgery, in: Proceedings of the 7th International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2004, pp. 97–104.

[2] U. Engelke, A. Maeder, H.-J. Zepernick, Analysing inter-observer saliency variations in task-free viewing of natural images, in: Proceedings of the 17th IEEE International Conference on Image Processing, IEEE Signal Processing Society, 2010, pp. 1085–1088.

[3] 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 20 (1998) 1254–1259.

[4] R. Huang, N. Sang, L. Liu, Q. Tang, Saliency based on multi-scale ratio of dissimilarity, in: Proceedings of the 20th International Conference on Pattern Recognition, IEEE Computer Society, 2010, pp. 13–16.

[5] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2007, pp. 1–8.

[6] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE Transactions on Image Processing 19 (2010) 185–198.

[7] Y. Yu, B. Wang, L. Zhang, Pulse discrete cosine transform for saliency-based visual attention, in: Proceedings of the IEEE 8th International Conference on Development and Learning, IEEE Computer Society, 2009, pp. 1–6.

[8] A. Desolneux, L. Moisan, J.-M. Morel, Computational gestalts and perception thresholds, Journal of Physiology-Paris: Neurogeometry and Visual Perception 97 (2003) 311–324.

[9] J.-C. Chiang, C.-S. Hsieh, G. Chang, F.-D. Jou, W.-N. Lie, Region-of-interest based rate control scheme with flexible quality on demand, in: Proceedings of the IEEE International Conference on Multimedia and Expo, IEEE, 2010, pp. 238–242.

[10] M.E. Tipping, The relevance vector machine, in: S.A. Solla, T.K. Leen, K.-R. Müller (Eds.), Advances in Neural Information Processing Systems, vol. 12, MIT Press, Cambridge, MA, 2000, pp. 652–658.

[11] M. Tipping, Sparse bayesian learning and the relevance vector machine, Journal of Machine Learning Research 1 (2001) 211–244.

[12] L. Itti, Automatic foveation for video compression using a neurobiological model of visual attention, IEEE Transactions on Image Processing 13 (2004) 1304–1318.

[13] N. Tsapatsoulis, C. Pattichis, A. Kounoudes, C. Loizou, A. Constantinides, J.G. Taylor, Visual attention based region of interest coding for video-telephony applications, in: Proceedings of the 5th International Symposium on Communication Systems, Networks and Digital Signal Processing, IEEE, 2006.

[14] Z. Li, S. Qin, L. Itti, Visual attention guided bit allocation in video compression, Image and Vision Computing 29 (2011) 1–14.

[15] H. Hadizadeh, I.V. Bajic, Saliency-preserving video compression, in: Proceedings of the IEEE International Conference on Multimedia and Expo, IEEE Computer Society, 2011, pp. 1–6.

[16] Y.H.Y. Chih Wei Tang, Ching Ho Chen, C.J. Tsai, Visual sensitivity guided bit allocation for video coding, IEEE Transactions on Multimedia 8 (2006) 11–18.

[17] S.P. Lu, S.H. Zhang, Saliency-based fidelity adaptation preprocessing for video coding, Journal of Computer Science and Technology 26 (2011) 195–202.

[18] W.J.H. Gary, J. Sullivan, Jens Rainer Ohm, T. Wiegand, Overview of the high efficiency video coding standard, IEEE Transactions on Circuits and Systems for Video Technology 22 (2012) 1649–1668.

[19] T. Liu, J. Sun, N.-N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2007, pp. 1–8.

[20] A. Kapoor, K.K. Biswas, A case-based reasoning approach for detection of salient regions in images, in: Proceedings of the 7th Indian Conference on Computer Vision, Graphics and Image Processing, ACM, New York, USA, 2010, pp. 48–55.

[21] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.

[22] MSRA salient object database, ⟨http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm⟩, 2007.

[23] R. Achanta, Ground truth of 1000 images, ⟨http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/GroundTruth/binarymasks.zip⟩, 2009.

[24] X. Xiang-Min, M. Yun-Feng, X. Jia-Ni, Z. Feng-Le, Classification performance comparison between RVM and SVM, in: Proceedings

of the IEEE-2007 International Workshop on Anti-counterfeiting, Security, Identification, 2007, pp. 208–211.
[25] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: B. Schölkopf, J. Platt, T. Hoffman (Eds.), Advances in Neural Information Processing Systems, vol. 19, MIT Press, Cambridge, MA, 2007, pp. 545–552.
[26] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems, vol. 18, MIT Press, Cambridge, MA, 2006, pp. 155–162.
[27] Z. Cernekova, I. Pitas, C. Nikou, Information theory-based shot cut/fade detection and video summarization, IEEE Transactions on Circuits and Systems for Video Technology 16 (2006) 82–91.
[28] L. Krulikovska, J. Pavlovic, J. Polec, Z. Cernekova, Abrupt cut detection based on mutual information and motion prediction, in: Proceedings of the IEEE International Symposium on Electronics in Marine, IEEE, 2010, pp. 89–92.
[29] R. Lienhart, Comparison of automatic shot boundary detection algorithms, Storage and Retrieval for Image and Video Databases SPIE 3656 (1999) 290–301.
[30] W. Lai, X.-D. Gu, R.-H. Wang, W.-Y. Ma, H.-J. Zhang, A content-based bit allocation model for video streaming, in: Proceedings of the IEEE International Conference on Multimedia and Expo, vol. 2, IEEE, 2004, pp. 1315–1318.
[31] V. Bhaskaran, K. Konstantinides, Image and Video Compression Standards: Algorithms and Architectures, 2nd edition, Kluwer Academic Publishers, Norwell, MA, USA, 1997.
[32] H.264/AVC reference software, ⟨http://iphome.hhi.de/suehring/tml/download/⟩, 2011.
[33] I.E. Richardson, The H264 Advanced Video Compression Standard, 2nd edition, John Wiley & Sons, Ltd, 2010.