Content-aware Rate Control Scheme for HEVC Based on Static and Dynamic Saliency Detection

Xuebin Sun[a], XiaoFei Yang[b], Sukai Wang[a] and Ming Liu[a,*]

[a]Department of Electronic & Computer Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, China
[b]Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai 200000, China

ARTICLE INFO

Keywords:
HEVC
Content-aware
Static Saliency
Dynamic Saliency
Rate Control

ABSTRACT

High efficiency video coding (HEVC) greatly outperforms previous standards H.264/AVC in terms of coding bit rate and video quality. However, it does not take into account the human visual system (HVS), that people pay more attention to specific areas and moving objects. In this paper, we present a content-aware rate control scheme for HEVC based on static and dynamic saliency detection. The proposed strategy mainly consists of three techniques, static saliency detection, dynamic saliency detection, and adaptive bit rate allocation. Firstly, we train a deep convolution network (DCN) model to extract the static saliency map by highlighting semantically salient regions. Compared to traditional texture-based or color-based region of interest (ROI) extraction techniques, our models are more in line with the HVS. Secondly, we develop a moving object segmentation technique to automatically extract the dynamic salient regions for each frame. Furthermore, according to the fusion saliency map, a coding tree unit (CTU) level bit control technique is exploited to realize flexible and adaptive bit rate allocation. As a result, the quality of salient regions is improved by allocating more bits, while allocating fewer bits to the non-salient regions. We verified the proposed method on both the JCT-VC recommended data set and eye-tracking data set. Experiment results show that the PSNR of salient regions can improve by an average of 1.85 dB without adding bit rate burden, which significantly improves the visual experience.

1. Introduction

In recent years, with the development of the high efficiency video coding (HEVC) standard, high-resolution video and large screens have been pouring into people's lives, bringing perfect visual enjoyment, and also posing a huge challenge to the bandwidth of communication channels. The superior compression efficiency of HEVC benefits from its flexible hierarchical coding structure and multiple prediction modes. It uses the well-known rate-distortion optimization (RDO) metric to obtain the optimal division structure and prediction mode. This process is an optimization problem that minimizes the overall reconstructed video distortion D at a given rate R. The RDO method only optimizes the coding performance based on the traditional target metric, ignoring the perceived characteristics of the video content.

However, depending on the use case, it is wise to use different encoding parameter sets or techniques to process different regions according to the video content. For instance, for conversational video, the face area is more important than the background. In facial areas, facial features (such as the eyes, mouth, and nose) appear to be more important than others, resulting in greater importance. Therefore, facial features should have the greatest weight, followed by the face area and background. Additionally, for sports video, the mover attracts most of the viewer's attention. Improving the coding quality in the moving object regions will therefor contribute to an enhanced visual experience. Unfortunately, the static adaptive bit rate budget and the strict R-Lambda model in HEVC fails to take into account the content of the video.

There has been increasing interest in perceptual video coding optimization recently [1] [2]. More specifically, Wu *et al.* [3] proposed an HEVC medical ultrasound video coding method based on a region of interest (ROI) map. They developed an efficient ROI extraction technique based on image texture features. According to the ROI map, the quantization parameter (QP) is adaptively adjusted to accommodate ROI and non-ROI. Yang *et al.* [4] used the Prewitt filter to extract the perceptual features and optimize the RDO process by perceptually adjusting the Lagrangian multiplier. Currently, the existing perceptual-based video coding approaches mainly use feature-based or color-based ROI extraction methods, and few use deep learning methods to extract the static saliency map [5]. Compared with the traditional ROI extraction method, the deep learning method is more in line with the human visual system (HVS) [6].

For sports video, audiences pay more attention to moving people or objects [7] [8]. Compared to the background region, the foreground region needs higher coding quality. Some researchers are committed to improving the quality of moving objects for surveillance video captured with static cameras [9] [10]. As far as we known, there are few HEVC optimization algorithms designed for improving the dynamic salient region quality, especially for sports video captured with moving cameras. To encounter these shortcomings, we develop a content-aware rate control scheme for HEVC based on static and dynamic saliency detection. Figure 1 shows the experimental results of the proposed method compared with the standard HEVC algorithm for Tandem and Butterfly video sequences from the data set of [11]. The bits cost for the current frame is almost the same. For Tandem, viewers pay greater attention to the facial region. While for Butterfly, the flying butterfly attracts more attention. It can be observed that compared with HEVC, the proposed method obtains clearer facial features for the Tandem sequence and clear coding quality for Butterfly. As a result, the proposed method provides a better visual experience.

In this paper, we design a deep convolution network (DCN) model to locate multiple salient regions in each frame of video. Model training only needs to be done offline. We also design an efficient moving object extraction technique for dynamic saliency

*Corresponding author.
E-mail address: sunxuebin@tju.edu.cn, eesxb@ust.hk (X.B. Sun), yangxf@siom.ac.cn (X.F. Yang), swangcy@connect.ust.hk (S.K. Wang), eelium@ust.hk (M. Liu)
ORCID(s): 0000-0002-4867-2282 (X. Sun); 0000-0002-4500-238X (M. Liu)

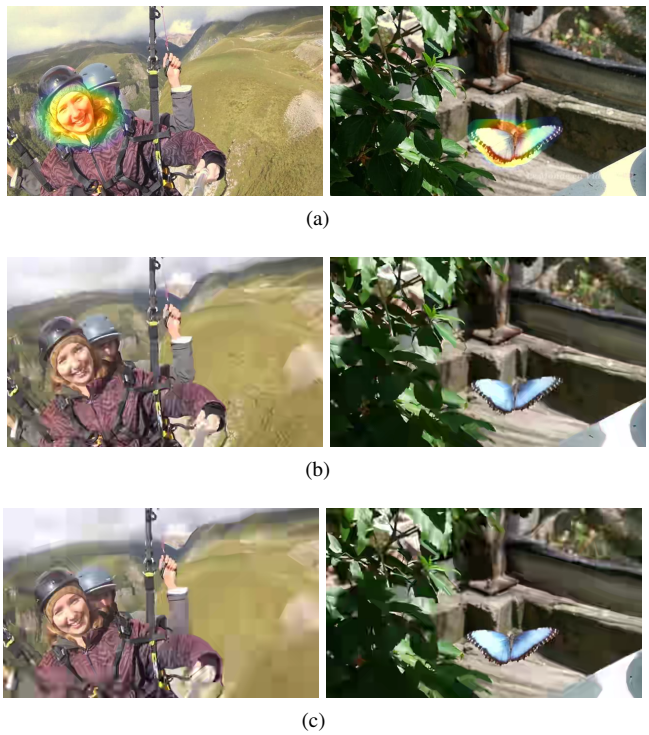


Figure 1: Experiment results of the HEVC algorithm and the proposed method for Tandem and Butterfly sequence: (a) heat map of the gaze locations, (b) HEVC coding result (Tandem: 3055 Bytes, Butterfly: 4260 Bytes), (c) proposed method coding result (Tandem: 3039 Bytes, Butterfly: 4049 Bytes).

detection. Furthermore, a perceptual-based rate optimization method is proposed according to the salient map. The method can adaptively adjust the QP in the RDO process according to the perceived characteristics of the video content. The experimental results show that compared with the original RDO process in HEVC, this method can significantly improve the perceptual coding performance.

The rest of this paper is organized as follows. Section 2 gives an overview of recent perceptual-based HEVC coding algorithms; in Section 3 we present a detailed description of the proposed method; Section 4 is devoted to the experimental results; Finally, we conclude and illustrate future work in Section 5.

2. Related work

Over the past decade, many researchers focuse on perceptual-based image or video coding optimization algorithms. According to the application field, the methods can be generally divided into four categories: conversational video coding, surveillance video coding, medical image or video coding, and conventional video coding.

2.1. Conversational video coding

For conversational video, as the background is usually fixed, the audience pay much attention to the people or objects in the foreground. Therefore, the foreground and background can encode separately according to visual importance. Zhou *et al.* [12] propose a multilevel ROI-based control algorithm for video communication. They segment the current frame into four regions according to the skin color and feature location, and reallocate resource based on visual importance. Experimental results show that their method can improve the peak signal-to-noise (PSNR) of ROI over 0.5dB. Li *et al.* [13] present a novel weight-based R-λ scheme for rate control in HEVC, in which they allocate more bits for the face region to improve the perceived visual quality for conversational videos. Deng *et al.* [14] introduce an ROI-based bit allocation method for HEVC towards conversational video. They perform the robust SURF cascade face detector to extract the ROI, and optimize the rate-distortion model to improve the subjective quality. Xu *et al.* [15] propose an ROI-based HEVC perceptual video coding approach for conversational videos. They endow the unequal importance within the facial region to emphasize its facial features. Their method greatly improves the overall perceived visual quality. Xiong *et al.* [16] develop a face-region-based conversational video coding algorithm. They present an efficient motion-based face detection method to identify face blocks and allocate more bits to encode these regions.

2.2. Medical image or video coding

Medical images such as, ultrasound or endoscopic images, contain large areas of black background areas that are used to record patient information and have no use for a doctor's diagnosis. Therefore, these areas can be encoded with a lesser bit rate. Sanchez *et al.* [17] propose a graph-based rate control algorithm for pathology images, in which the non-ROI is compressed in a lossy manner according to a target bit rate, and the RoI in a lossless manner. Yee *et al.* [18] develop a novel image compression format, known as Better Portable Graphics (BPG), especially for medical image. Firstly, they segment the medical image into two parts: ROI and non-ROI regions. The ROI areas are compressed by lossless BPG compression algorithm, while the non-ROI are compressed by lossy BPG algorithm. Chen *et al.* [19] propose a novel video coding system for telemedicine applications based on HEVC. They perform a two-layer coding approach, in which they use the standard HEVC for the base layer and a lossless coding scheme for the enhancement layer. Wu *et al.* [3] develop an ROI-based coding scheme for medical ultrasound video. Firstly, an effective ROI extraction technique is performed according to image textural features. Then, they design a hierarchical coding method by adjusting the transform coefficient adjustment and QP for the ROIs and non-ROIs.

2.3. Surveillance video coding

For surveillance video, the camera is generally fixed and more attention is paid to moving people or objects. Xue *et al.* [9] introduce a fast ROI-based HEVC coding algorithm for surveillance videos. They use an automatic segmentation to generate the ROI mask. The ROI regions are encoded using normal procedures, while the non-ROI regions are encoded using fast prediction mode selection and fast coding unit (CU) partition techniques. Experimental results show that the proposed method can significantly reduce the computational costs without degrading the visual quality of the ROI. Meuel *et al.* [10] present an ROI coding system for aerial surveillance video. They use an ROI detector to automatically classify a video image on board the UAV in ROI and non-ROI. Then the ROIs are regularly encoded, whereas non-ROI regions are forced to skip mode. To reduce the coding cost of surveillance video, Xing *et al.* [20] propose a surveillance video encoding method with HEVC by dealing with the foreground objects and the background separately. Results

show that their method can remarkably save the total bit-rate. Zhao *et al.* [21] introduce a rate control (RC) scheme based on ROI for unmanned aerial vehicle (UAV) video. A salient object detection method is developed to obtain the ROI map. By using the ROI map, they exploit a bit allocation methods at frame level and large coding unit (LCU) level.

2.4. Conventional Video coding

For conventional video, the ROI is closely related to the human visual perception system. Perceptual-based optimization techniques are designed by allocating more bits for important regions and fewer bits for others. These methods can enhance the visual experience without increasing bandwidth burden. Xu *et al.* [22] propose a saliency-based video coding method with HEVC. The saliency map is obtained according to the motion, color and texture characteristics. By adaptively changing the QP, a quantization control is developed based on the visual saliency map. Wang *et al.* [23] presents a perceptual video coding scheme for HEVC using a saliency map and just-noticeable distortion. Their method can discover the saliency region and define a threshold for unperceived distortions to remove the perceptual redundancy. Bai *et al.* [24] introduce a saliency-based rate control scheme for HEVC. They allocate more bit rate for the ROI while relative lower bit rate ratio for non-ROI. The total bit rate is still close to the assigned target value. In order to improve the visual quality, Wei *et al.* [25] propose a (coding tree unit) CTU-level rate control scheme for HEVC. A saliency map is firstly extracted. The distortion of each CTU is determined by the corresponding saliency map.

2.5. Summary and analysis

The aforementioned works are all great efforts for perceptual-based HEVC coding method. However, they mostly use texture or color features to analyze the interest areas of visual perception. Few of them use the emerging deep learning methods to extract saliency maps. Compared with the traditional ROI extraction algorithm, the neural network method can extract the salient region more accurately. In addition, we observed that few consider the dynamic object as a sensitive region for videos captured by motion cameras. Therefore, our study mainly addresses these issues. In this paper, we propose a perceptual-based HEVC optimization method using DCN and moving object segmentation techniques to enhance the visual experience.

3. The proposed method

The purpose of the perceptual-based video coding method is to improve the coding quality of the video salient regions. The key is to extract the static and dynamic saliency maps of the input video and use these saliency maps to guide the entire video encoding. With the aim to improve the coding quality of the salient areas, an ROI-based bit allocation is designed for perceptual video coding. Therefore, the bit rate can be adaptively assigned according to the perceptual saliency map of each frame. The proposed perceptual-based hybrid optimization algorithm is mainly composed of three parts: static saliency detection, dynamic saliency detection and the bitrate allocation method. These three techniques are discussed in detail following.

3.1. Static saliency map extraction with DCN

Static saliency detection has aroused great interest from researchers in recent years, with the aim of highlighting visually significant areas or objects in an image. It has been widely used in computer vision tasks, for instance, image or video coding [26], content-aware image resizing and image retrieval [27]. During the last decades, a large number of algorithms have been proposed to obtain different saliency cues [28]. Harel *et al.* [29] introduce a graph-based visual saliency detection method. They first form activation maps on certain feature channels, and then normalize them in a way which highlights conspicuity and admits combination with other maps. Hou *et al.* [30] propose an image descriptor as the image signature, which is utilized to approximate the spatial location of a sparse foreground hidden in a spectrally sparse background. Experimental result demonstrates that the approximate foreground positions highlighted by the image signature are very consistent with the positions of eye movement gaze, and they are better predicted than the predictive saliency algorithm in a small part of the calculation cost. Based on spectral residua (SR), Guo *et al.* [31] develop a spatiotemporal saliency detection method. Their method calculates the phase spectrum of the Quaternion Fourier Transform (QFT) of the image to obtain the spatiotemporal saliency map of the image. The phase spectrum of the image is a significant target in the image. Each pixel in the image is represented by a quad: color, brightness, and the motion vector. Their method is independent of prior information, and is suitable for real-time significance detection. Goferman *et al.* [32] propose a context-aware saliency detection method based on four principles observed in the psychological literature: local low-level considerations, global considerations, visual organizational rules, and high-level factors. Experimental results show that their method helps to generate concise, attractive and informative summaries.

Many traditional saliency detection algorithms focus on the design of low-level saliency cues or background priors modeling. However, a low level of significant cues or priors does not produce sufficiently good saliency results, especially when the saliency target is presented in a low-contrast background with visually confusing visual effects. This problem poses a serious problem to the traditional method. The emerging deep convolutional neural network has been successfully applied to various computer vision tasks, for instance, image classification [33], object detection [34] [35] and semantic segmentation [36]. It has shown its powerful ability to extract advanced feature representations, which can well solve the above problems. From another perspective, saliency detection is a task that mimics the human attention mechanism, which is a neurocognitive response controlled by the brain. The goal of a DCN is to mimic the function of the new cortex in the human brain as a layer of filters and nonlinear operations. In order to better detect semantically significant objects, advanced knowledge of object classes becomes very important.

To tackle the aforementioned problems caused by traditional algorithms, we design a deep learning framework for static salient object detection. Our goal is to train a DCN to extract the static salient regions. Unlike traditional object detection models, it is unnecessary to accurately obtain a tight bound on the objects. Conversely, it is crucial to approximately identify and locate multiple objects in a frame. In a classification DCN model, a class is identified by a set of 3D feature maps learning by the network. For example, given an image with $n \times n$ resolution, suppose each layer L contains d_l features. The parameter needed to

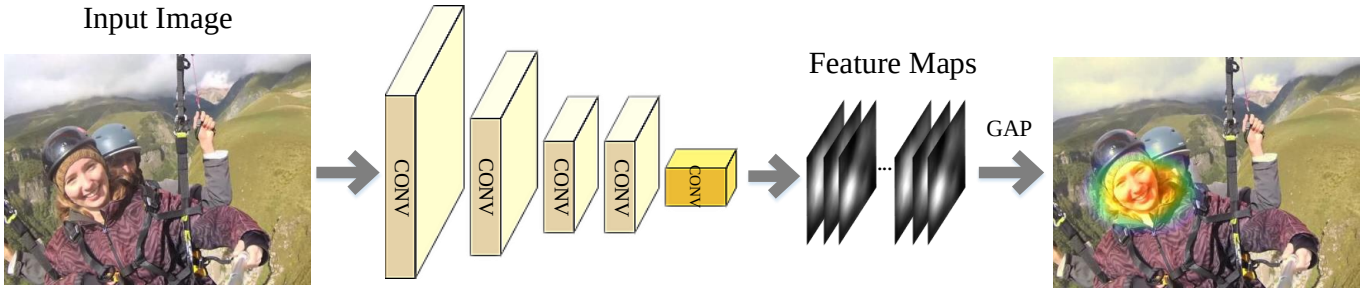


Figure 2: The DCN structure for static saliency map extraction.

Figure 3: Experiment results of the DCN structure: (a) BasketballDrive, (b) PartyScene, (c) FourPeople, (d) Video3.

be calculated for layer L is as follows:

$$\sum_{l \in L} d_l \times C \times \frac{n}{k^l} \times \frac{n}{k^l}, \tag{1}$$

where k denotes the maximum pooling stride size, and C represents the number of classes. For such a network, it is computationally expensive and infeasible to learn a model. Considering the real-time requirement of video compression, we can hardly transfer the DCN technique to the video coding application. In order to reduce the complexity of the algorithm, some optimization strategies need to be exploited.

The majority of CNN models are designed for classification or segmentation applications. For instance, the well-known ImageNet is designed to recognize up to one thousand classes including animals, flowers and other objects. In our model, it is unnecessary to accurately distinguish each category of each species. The number of classes is significantly reduced by folding similar species into a more general category. Obviously, most images contain only a few classes, so building a separate functional diagram for each class is computationally inefficient. In addition, many classes share similar low-level features, even though the number of classes is relatively small. Therefore, parameters are shared across the feature maps for different classes.

The structure of the proposed static saliency map extraction network is illustrated in Figure 2, which is modified on the basis of the classic VGG network [37]. The structure of VGG mainly consists of convolutional layers. Before the last output layer, we perform a global average pool on the convolution feature maps instead of the SoftMax loss function used for classification. Moreover, in order to improve the detection accuracy, the sigmoid active function is used. For a given class c, suppose Z_l^c represents the sum of the activations of layer l for all the feature maps. $f_k(x, y)$ represents the activation of unit k in the last convolutional layer. Therefore, Z_l^c can be calculated by the following formula:

$$Z_l^c = \sum_k \omega_k^c \sum_{x,y} f_k(x, y) \quad = \sum_k \sum_{x,y} \omega_k^c f_k(x, y). \tag{2}$$

We define $S_c(x, y)$ to represent the class importance map for class c. Thus, $S_c(x, y)$ reflects the importance of the activation in the spatial grid, and is given by

$$S_c(x, y) = \sum_{x,y} \omega_k^c f_k(x, y). \tag{3}$$

As depicted in Figure 2, through the global average pool, the last layer outputs the average value of each unit's characteristic graph. Finally, the weighted sum of the convolutional feature map is calculated to obtain multiple saliency maps.

The model was trained using 256 types of artificial and natural objects from the Caltech data set [38], including common animals and plants, buildings, etc. We tested the DCN model on the JCT-VC recommended video test sequence [39]. Figure 3 shows four different frames, as well as the corresponding saliency maps and heat maps generated by the proposed DCN model. It can be seen that the DCN model can deal with different types of video from simple to complex. The static saliency map can accurately be extracted. It is sensible and reasonable to encode the salient regions with high quality. Moreover, it can be observed that the salient areas are arbitrarily shaped and are gradually-changed. This will avoid the square effect when implementing it into HEVC.

3.2. Dynamic saliency map extraction

For sports or surveillance videos, the viewer's attention mainly focuses on the moving objects. Therefore, these regions should

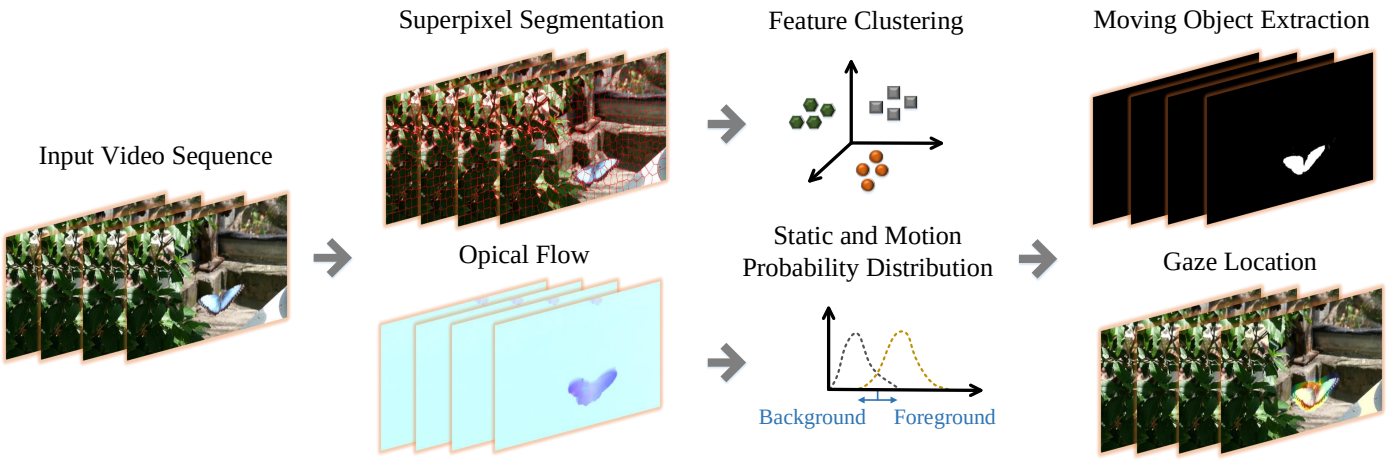


Figure 4: Overview of our dynamic saliency map extraction framework.

have higher compression quality. In this section, we develop a low-complexity technique to extract the moving object region in the foreground, paving the way for further rate control algorithms.

For diverse and complex motions, motion features are not a sufficiently strong cue to extract the moving object region. In some cases, it is not competent, for instance, if only part of the foreground moves in the video while the rest remains still, or the entire moving foreground stops moving for a few frames. Similarly, detecting visual similarity between distant frames can obtain long-term connections, but it is incompetent in judging whether they belong to the sports area or not. This is because in distant frames, moving objects may experience non-rigid deformation, changes in illumination, changes in scale, etc. Therefore, these distant matches can become very noisy and fragmented.

To perform video segmentation, we combine both the motion features and visual similarity. Figure 4 gives the overall workflow of the proposed moving object extraction method. Firstly, we perform a superpixels algorithm for input frames. Meanwhile, we estimate the optical flow field from two consecutive frames, which is used to analyze the moving object of the video frames. Then, we perform an iterative process at each pixel both in space and in time across the video sequence to continuously update the motion area to figure out the final result. The high-level sketch of the proposed technique is described in algorithm 1. The details are described following.

Given an input video sequence $\{F^1, F^2, ..., F^n\}$, we use the SLIC algorithm to over-segment each frame into superpixels [41]. Let $F^k = \{R_1^k, R_2^k, ...\}$ denote the superpixel set of F^k. This number of the superpixels is a good compromise between maintaining a compact frame representation in the graph and maintaining the high precision of the object boundary. This over-segmentation is competent to extract meaningful boundaries even if there is high motion blur or low resolution.

Each superpixel is represented by concatenation of several types of descriptors, including mean RGB value, intensity contrast and orientation feature contrast. Merging the relative spatial coordinates of the superpixel into the descriptor can implicitly punish the spatial distance NN in the NN search. The valid $k-d$ tree search is used to find the approximate NNs. The time radius is set to f frames. f equals to half of its frame rate. Each

Algorithm 1 Outline of the moving object extraction algorithm.

Input:

A video sequence $\mathbf{F} = (F^1, F^2, ..., F^n)$;

Output:

The segmentation result of the video sequence $S_{dynamic} = (s_1, s_2, s_3, ..., s_n)$.

1: Use the simple linear iterative clustering (SLIC) algorithm to oversegment each frame F^k into superpixels set $F^k = \{R_1^k, R_2^k, ...\}$ [40].
2: Use high-dimensional descriptor $d(R)$ to represent each superpixel R.
3: For each region R, search and find M Nearest Neighbors (NNs) in the feature space $d(R)$.
4: Construct a undirected graph G = (V, E) with superpixels as nodes V and the links between adjacent nodes as edges E.
5: Calculate the weight between R and each of its NNs $\{NN_m(R)\}_{m=1}^M$.
6: Accumulate the optical flow gradient magnitudes for each frame within a temporal window of t frames to obtain relatively longer term motion information of the foreground regions.
7: According to the optical flow, assign an initial fg likelihood vote $v^{(0)}(R)$ for each region R.
8: Update the vote of each region R with the weighted average of the votes of its M Nearest Neighbors.
9: Normalize the votes in each frame and obtain the final dynamic saciency map.
10: **end for**
11: **return** $S_{dynamic} = s_m$

superpixel searches for NNs within k frames, including its own (ie $k = 2f + 1$ frames). As fg superpixels tend to be similar to nearby fg superpixels, each superpixel has several good matches (not just one) in a frame. Therefore, we set the number of NNs per superpixel to $m = l(2f + 1)$, where $l = 4$.

For the video sequence, we construct a undirected weighted graph $G = (V, E)$ with superpixels as nodes V and the links between NNs as edges E. The $\omega(R, NN_m(R))$ indicates the

weight of the edge between R and $NN_m(R)$:

$$\omega(R, NN_m(R)) = e^{-\frac{||d(R) - d(NN_m(R))||^2}{\sigma^2}} \tag{4}$$

P denotes a random-walk transition matrix over the graph.

$$W(i,j) = \begin{cases} \omega(R_i, R_j) & R_j \in NN(R_i) \\ 1 & i = j \\ 0 & otherwise \end{cases} \tag{5}$$

After that, we look for frames with a dominant direction of motion, which is likely related to the motion of the background camera. The dominant motion means that the camera either is near to static or the camera is translating. Such frames allow for a more reliable initial separation of fg/bg pixels compared to other frames. Then we calculate a rough motion saliency map for these frames represented by vectors. For the rest of the frames, we just set the initial value of the vector to zero. This frame selection mechanism makes sense, assuming that most videos contain enough frames and simple camera motion (almost static cameras or translating), especially if it is long enough.

We accumulate the optical flow gradient magnitudes for each frame within a temporal window of t frames to obtain relatively longer-term motion information of the foreground regions [42]. In order to reduce the complexity, each frame is down-sampled to a low dimension. The horizontal and vertical motion vector is represented by $MV_x(x,y)$ and $MV_y(x,y)$, respectively. Firstly, we look for the frame with nearly no motion. If the median of the optical flow magnitude is less than 1 pixel, these frames are considered as static frames. Then, by calculating the global histogram of the optical flow direction, we look for the frames with the camera translating in some dominant direction. Furthermore, we calculate a motion saliency map for these frames. For each pixel, we take the stream vector in its surrounding 5×5 patch and calculate its deviation from the estimated dominant motion. In frames with static dominant motion, we calculate the deviation of the stream size from zero, while in frames with dominant translation, we calculate the deviation of the flow direction from the dominant direction:

$$\begin{aligned} \widetilde{MV}(x,y) &= \sqrt{\widetilde{MV}_x(x,y)^2 + \widetilde{MV}_y(x,y)^2} \\ \widetilde{MV}_x(x,y) &= MV_x(x,y) - \overline{MV}_x(x,y) \\ \widetilde{MV}_y(x,y) &= MV_y(x,y) - \overline{MV}_y(x,y) \end{aligned} \tag{6}$$

where $\widetilde{MV}(x,y)$ represent the final motion coefficient, while $\overline{MV}_x(x,y)$ and $\overline{MV}_x(x,y)$ denote the dominant motion vector. These deviations provide a significant score for the pixel. According to the optical flow, assign an initial fg likelihood vote $v^{(0)}(R)$ for each region R. For $t = 1 : T$, we update the vote of each region R with the weighted of its M Nearest Neighbors.

$$\vec{v}^{(t)} = P\vec{v}^{(t-1)} \tag{7}$$

Representative moving object extraction results are illustrated in Fig. 5. As can be seen, the proposed method can deal with various scenarios, and target foregrounds can be segmented accurately. This paves the way to further bit rate control algorithm. As viewers pay greater attention to the moving object, these regions should be coded with high quality. Compared with other unconstrained video fg/bg segmentation methods, the proposed method has higher computational efficiency. The extraction of superpixels and their descriptors takes several seconds per frame, and so to do calculations that approximate the nearest neighbor. Each voting iteration is very efficient because it is simply a multiplication of a very sparse matrix with a voting vector. On a regular PC, the entire run time (including saliency-based initialization) takes about 5 seconds per frame.

3.3. Static and dynamic saliency map fusion

For conventional video, viewers's attention focus on some particular person or object. Additionally, moving objects also attract their attention. Therefore, the static and the dynamic saliency map need to merge together to guide the bit allocation of video coding. In our method, a linear fusion scheme is used, as follows:

$$\begin{aligned} SM_{fusion} = \alpha \cdot SM_{static} + \beta \cdot SM_{dynamic} + \gamma \cdot SM_{mix} \\ SM_{mix} = SM_{static} \cdot SM_{dynamic}, \end{aligned} \tag{8}$$

where SM_{static} represents the static salient map obtained by the DCN model, and $SM_{dynamic}$ denotes the dynamic salient map obtained by the moving object extracion method. SM_{mix} represents the mixed saliency map, and α, β and γ denote weighting factors, defined as follows:

$$\begin{aligned} \alpha &= 1 \\ \beta &= \sqrt{\frac{\delta_{fusion}}{\delta_{dynamic}}} \\ \gamma &= 2 \cdot \sqrt{\frac{\delta_{static}}{\delta_{fusion}} \cdot \frac{\delta_{dynamic}}{\delta_{fusion}}} \end{aligned} \tag{9}$$

where δ_{static}, $\delta_{dynamic}$ and δ_{fusion} represent the standard deviation for the static, dynamic, and mix saliency map, respectively. Normalization is carried out for the fusion saliency map. The final saliency map is obtained and is used for bit allocation.

It is noted that for some special applications, we can also only use the static saliency map or dynamic saliency map independently. For example, for conversational videos, as the movements are very small, we can only use the static saliency map to guide the video coding. Conversely, a sports video is often full of intense motion and has a high frame rate. Sometimes, to save bandwidth, the video can be encoded only according to the dynamic saliency map. As the viewer's attention is mainly focused on moving objects, this does not reduce the viewer's visual experience.

3.4. Perceptual-based rate control scheme

3.4.1. Overview of the HEVC rate control scheme

In video coding, the main goal of rate control is to minimize the distortion of the compressed video at a given bit rate [43] [44]. In order to achieve this goal, the R-λ rate control scheme was employed. The main steps of the scheme are to find the bpp-λ and λ-QP relationships and finally figure out the QP value.

$$\lambda = -\frac{\partial D}{\partial R}, \tag{10}$$

where D denotes the distortion and R represents the bit-rate for one CTU. The Hyperbolic model $D = CR^{-K}$ is utilized in the

Figure 5: Experiment results of dynamic salient regions extraction: (a)BasketballDrill, (b) Soccer, (c) CREW, (d) Hurdles.

rate control scheme to characterize the relationship between R and D [45] [46]. C and K are parameters determined by the video content. Thus, with (9), the relationship between R-λ can be formulated by:

$$\lambda = -\frac{\partial D}{\partial R} = CK \cdot R^{-K-1} = a \cdot R^b, \tag{11}$$

where a and b are parameters related to video content. Since the contents of different CTUs are different, a and b need to be updated with the encoding process of each CTU. Once the R of a CTU is obtained, λ can be used as an output to estimate the QP of the current CTU. Here, the bit rate R can be modeled with bpp:

$$R = bpp \cdot f \cdot w \cdot h, \tag{12}$$

where w and h are the width and height of the video frame. f denotes the frame rate. bpp represents the bit per pixel of the current CTU. With (11), (10) λ can be written as

$$\lambda = \alpha \cdot bpp^{\beta}, \tag{13}$$

where $\alpha = a \cdot (fwh)^b$ and $\beta = b$ are parameters determined by the video content. The values of a and b will be updated with the encoding process. Their initial value is set to 3.2003 and -1.367, respectively. Different alpha and beta initial values have little effect on the R-D performance and bit-rate error.

After obtaining the bpp-λ relationship, the following task is to figure out the λ-QP relationship. The QP value can be obtained by iteratively calculating the QP optimization process. Rate distortion cost is given by:

$$minJ(QP) = D(QP) + \lambda \cdot R(QP), \tag{14}$$

where $D(QP)$ represents the distortion, and $R(QP)$ denotes the rate. The optimal QP can be obtained by solving formula (13). However, this optimization greatly increases the coding complexity. In order to reduce the coding complexity, it is recommended to use the fitting formula instead of multiple QP optimization to determine the QP value QP_j of the j-th LCU:

$$QP_j = \theta_0 \cdot ln\lambda_j + \theta_1, \tag{15}$$

where λ_j represents the smooth λ value of the j-th CTU. θ_0 and θ_1 denote coefficients that fit a linear relationship between QP and $ln\lambda$.

3.4.2. The proposed perceptual-based rate control scheme

In HEVC, each frame is divided into a series of coding units to be processed, and the size of the CU is related to the QP and the quantization step. The values of the QP and the quantization step affect the bits assigned to each CU. Obviously, the more bits is allocated, the higher the quality is. According to the saliency map obtained above, when a CU belongs to a significant area, it will get more attention and it will be allocated more bits to make the encoding quality as high as possible. Conversely, the human eye is not sensitive to the region when it belongs to a non-protruding area so it is possible to save unnecessary bits at these regions. As QP and bit allocation are closely related, our main purpose is to adaptively adjust the QP value to optimize the quantization process and ensure the reconstructed video quality. It is important to choose the appropriate initial QP for a given frame. In the proposed method, the QP for the i-th CTU is a function of the saliency map weight:

$$QP_i = round(\frac{QP_{frame}}{\sqrt{\omega_i}}), \tag{16}$$

where QP_{frame} represents the QP of a given frame, and QP_i denotes the initial QP for the i-th CTU. As an activation function, sigmoid function has been widely used in deep learning. ω_i represents the adjustment factor associated with video content, and is defined as a sigmoid function [47]:

$$\omega_i = a + \frac{b}{1 + exp(\frac{-c(S_{CTU} - S_{ave})}{S_{ave}})}, \tag{17}$$

where S_{CTU} represents the saliency value of the current CTU according to the fusion salient map. S_{ave} denotes the average importance values of the current frame. a, b, and c are defined empirically. a and b determines the coding quality for the salient and non-salient regions, respectively, while c reflects the changing rate of coding quality from non-salient regions to salient regions. Here, we set a = 0.7, b = 0.6, and c = 4, empirically. By solving formula (16) and (17), it can be observed that QP is a function of the saliency map weight. It can be observed that a larger weight of ω_i indicates that the current CTU is more sensitive to viewers' attention, and should be perceived at higher coding quality. When a CTU is less sensitive to viewers' attention, a smaller ω_i is obtained and a large QP will be used.

4. Experimental results

This section firstly evaluates the performance of the proposed saliency model and compares it with the methods by Hadizadeh *et al.* [48] and Zhu *et al.* [49]. Then, the coding performance of the proposed method compared with HEVC is presented.

4.1. Saliency model performance

To evaluate the performance of the proposed saliency model, the eye-tracking dataset presented in study [50] is utilized. The dataset consists of 12 standard video sequences. In these sequences, the categories of the objects and the way they are taken vary widely. These sequences include fast-moving objects, such as the Crew and the Soccer sequences; calm sequence, such as Hall Monitor; complex pictures, such as Bus; and human-centered pictures, such as Foreman and Mother & Daughter.

The quantitative evaluations of the saliency model performance are evaluated in terms of receiver operating characteristics (ROC), area under the curve (AUC), and similarity score (SIM). ROC is the most commonly utilized metric in the deep learning community. AUC is utilized to reflect the overall performance; the larger the AUC, the better the performance. An AUC value of 0.5 indicates that the model has a chance of predicting human gaze. ROC and AUC are calculated according to true positive rate (TPR) and false positive rate (FPR):

$$TPR = \frac{TP}{TP + FN} \tag{18}$$

$$FPR = \frac{FP}{TN + FP}, \tag{19}$$

where TP represents that a positive instance is predicted to be positive, while FN denotes that a positive instance is predicted to be native. FP and TN are the opposite.

There is no doubt that ROC analysis is beneficial. However, it lacks the description of the spatial deviation between the predicted significance map and the actually fixed map. When the predicted protruding position is misplaced close or away from the actual protruding position, it will result in a different performance. For a more comprehensive assessment, we also consider the similarity measure (SIM) in the experiment. The SIM measures the similarity between the two distributions. The similarity is the sum of the minimum values of each point in the distribution obtained by scaling each of the distributions to one. Mathematically, the similarity between the two graphs P and Q can be calculated as follows:

$$SIM = \sum_{i,j} min(P_{i,j}, Q_{i,j}) \quad \sum_{i,j} P_{i,j} = \sum_{i,j} Q_{i,j} \tag{20}$$

Fig. 6 illustrates the average performance of the ROC curves of the proposed method compared with Hadizadeh [48] and Zhu [49] methods. It can be observed that our method obtained a better performance. Zhu [49] reports better results than Hadizadeh [48]. Table 1 depicts the AUC and SIM data for all videos and the final results. The report shows that the method is superior to Hadizadeh [48] and Zhu [49] methods in 12 videos, and in some cases has a larger lead.

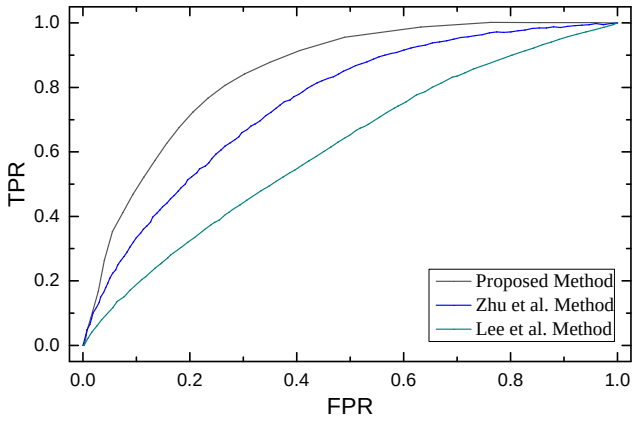


Figure 6: Average ROC curve.

Table 1
Saliency performance of the proposed algorithm versus other methods

Test Sequence	Hadizadeh *et al.* [48]		Zhu *et al.* [49]		Proposed	
	AUC	SIM	AUC	SIM	AUC	SIM
Bus	0.70	0.49	0.81	0.63	0.87	0.68
City	0.73	0.58	0.80	0.75	0.80	0.69
Crew	0.55	0.50	0.71	0.64	0.93	0.78
Foreman	0.65	0.46	0.82	0.66	0.85	0.73
Flower	0.49	0.44	0.72	0.64	0.88	0.74
Hall	0.69	0.38	0.84	0.59	0.82	0.68
Harbor	0.56	0.37	0.71	0.61	0.88	0.64
Mobile	0.67	0.48	0.69	0.56	0.78	0.67
Mother	0.55	0.41	0.80	0.61	0.86	0.71
Soccer	0.64	0.37	0.74	0.60	0.79	0.68
Stefan	0.66	0.41	0.86	0.66	0.84	0.75
Tempete	0.57	0.49	0.74	0.60	0.83	0.69
Average	**0.62**	**0.45**	**0.77**	**0.63**	**0.84**	**0.70**

4.2. Content-aware rate control scheme performance

4.2.1. Dataset and evaluation metrics

To evaluate the performance of the proposed algorithm, various comparisons and subjective visual tests were conducted. We implement the method on the newest version of the HEVC reference model HM 16.7. Eighteen representative sequences provided by JCT-VC [39] are picked for testing. Additionally, we also use the eye-tracking dataset to perform our experiments [50] . Unmodified HM-16.7 encoder is used as a baseline. All frames are coded with intra-mode under QP = 22, 27, 32, 37. The experiments test 50 frames for each sequence. Coding efficiency is measured by $\Delta PSNR$, ΔBR, and ΔT [51]. These metrics are defined as follows:

$$\Delta PSNR = PSNR_{proposed} - PSNR_{HM16.7} \tag{21}$$

$$\Delta BR = \frac{BitRate_{proposed} - BitRate_{HM16.7}}{BitRate_{HM16.7}} \times 100\% \tag{22}$$

$$\Delta T = \frac{T_{proposed} - T_{HM16.7}}{T_{HM16.7}} \times 100\%, \tag{23}$$

where $PSNR_{proposed}$, $BitRate_{proposed}$ and $T_{proposed}$ represents the PSNR, bit rate, and coding time of the proposed method,

Table 2
Performance of the proposed algorithm versus the original algorithm HM16.7.

Sequence		PSRN difference for each region Δ PSNR (dB)			Bit rate Δ BR%	Coding Time Δ T%
		whole	non-salient	salient		
Class A [2560 × 1600]	PeopleOnStreet	-0.62	-0.91	1.90	-1.40	6.18
	Traffic	-0.88	1.12	2.13	0.24	3.90
Class B [1920 × 1024]	BasketballDrive	-0.74	-1.26	0.30	-4.17	9.14
	Cactus	-0.59	-1.06	1.61	-1.68	7.87
	ParkScene	-0.74	-1.74	3.98	-1.72	8.13
Class C [1280 × 704]	FourPeople	-0.99	-1.59	1.23	-3.12	18.09
	Johny	-0.86	-1.44	0.99	0.65	20.53
	SlideEditing	-1.06	-1.98	2.85	-0.15	15.90
	Vidyo1	-1.20	-2.26	1.04	-3.65	20.11
	Vidyo3	-1.18	-2.47	0.07	2.50	19.55
Class D [832 × 448]	BasketballDrill	-1.06	-1.39	3.27	-6.30	23.81
	BQMall	-0.85	-1.34	1.96	-0.93	28.25
	PartyScene	-0.36	-0.73	2.62	-1.59	22.94
	RaceHorses	-0.59	-0.48	3.38	-0.22	27.47
Class E [384 × 192]	BasketballPass	-1.09	-1.55	1.36	-0.47	35.71
	BlowingBubbles	-0.95	-1.30	0.79	-9.01	40.54
	BQsquare	-0.50	-0.75	2.66	0.30	37.50
	RaceHorses	-0.81	-1.02	1.56	-0.55	31.25
Class F Eye-tracking Dataset [352 × 288]	Bus	-0.83	-1.16	2.34	-1.26	31.25
	City	-0.63	-1.31	1.76	0.27	34.88
	Crew	-1.07	-2.45	0.69	0.16	39.79
	Foreman	-0.77	-1.04	1.80	-3.74	39.92
	Flower Garden	-0.61	-0.88	3.69	-2.29	37.50
	Hall Monitor	-0.77	-1.21	0.65	0.36	38.93
	Harbor	-1.03	-1.34	1.78	-1.27	33.33
	Mobile Calendar	-0.77	-1.42	2.85	-1.22	30.61
	Mother Daughter	-1.21	-2.05	0.30	-0.41	40.87
	Soccer	-0.58	-0.93	1.51	-0.49	36.07
	Stefan	-0.76	-1.20	3.09	-0.76	33.51
	Tempete	-0.56	-1.10	1.49	0.31	31.91
Average		**-0.82**	**-1.27**	**1.85**	**-1.39**	**26.85**

respectively. $PSNR_{HM16.7}$, $BitRate_{HM16.7}$ and $T_{HM16.7}$ depicts the PSNR, bit rate, and coding time in HM16.7, respectively. ΔBR indicates an increase in bit rate, and ΔT denotes total code time variation.

As the above objective evaluation metrics do not take into account the video content. We also use the eye-tracking weighted mean square error (EWMSE) metric proposed in [26] to give a subjective quality assessment for the proposed method with saliency map. The EWMSE is defined as follows:

$$EWMSE = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H}(\omega_{x,y}\cdot(F'_{x,y}-F_{x,y})^2)}{W\cdot H\cdot\sum_{x=1}^{W}\sum_{y=1}^{H}\omega_{x,y}}, \quad (24)$$

where $F'_{x,y}$ and $F_{x,y}$ represents the pixel value at location (x, y) for frame F' encoded by the proposed method, and F encoded by the standard algorithm, respectively. W and H denotes the horizontal resolution and vertical resolution, respectively, of the video. $\omega_{x,y}$ denotes the weight of the distortion at the pixel position (x, y), relating to the video content. The $\omega_{x,y}$ can be calculated by following formula:

$$\omega_{x,y} = \frac{1}{2\pi\delta_x\delta_y G}\sum_{g=1}^{G} e^{-\frac{(x-x_{pg})^2}{2\delta_x^2}\cdot\frac{(y-y_{pg})^2}{2\delta_y^2}} \quad (25)$$

where (x_{pg}, y_{pg}) gives the fixed position of the eyeball of the g-th subject in the eye-tracking database described in the study [50]. There were 15 subjects in the eye-tracking database, i.e, G=15. δ_x and δ_y represent two parameters that specify the range or width of the Gaussian function based on the line of sight and viewing angle. The values of δ_x and δ_y can be determined according to the size of the fovea, which is approximately $2-5°$ of the viewing angle. In this study, we specify $\delta_x = \delta_y =$ 64 pixels as the 2° angle of view described in study [26] and [50]. Based on the EWPSNR metric, an equivalent eye-tracking weight $PSNR$ can be derived, as defined by following Equation:

$$EWPSNR = 10log_{10}(\frac{255^2}{EWMSE}) \quad (26)$$

In the proposed experiment, the average value of the EWPSNR is considered as an indicator to measure the subjective

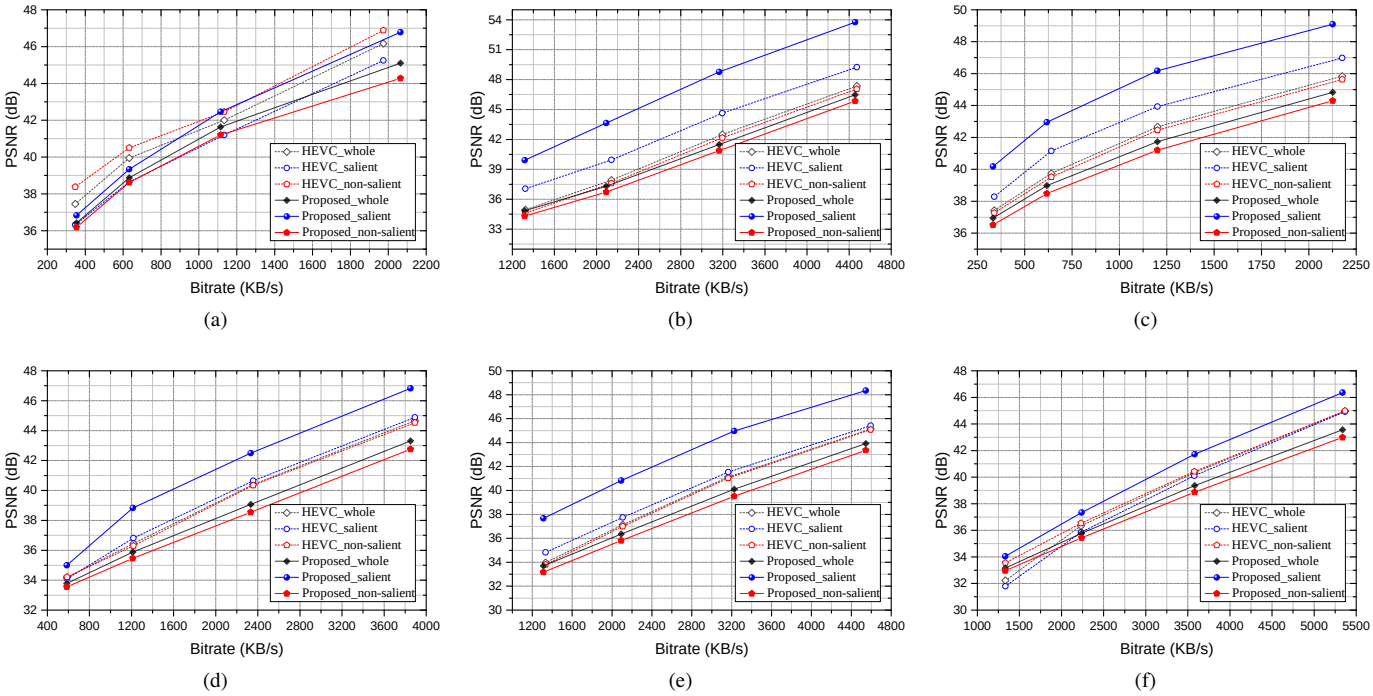


Figure 7: RD curves of six test video sequences: (a) Foreman, (b) Crew, (c) Soccer, (d) Soccer, (e) Stefan, (f) Tempete.

Table 3
Performance of the proposed algorithm versus other methods.

Test Sequence	Hadizadeh *et al.* [48]		Zhu *et al.* [49]		Proposed	
	BD-PSNR	BD-EWPSNR	BD-PSNR	BD-EWPSNR	BD-PSNR	BD-EWPSNR
Bus	-0.61	0.24	-0.39	0.47	-0.88	0.58
City	-0.45	0.16	-0.24	0.64	-0.83	0.43
Crew	-0.34	0.02	-0.15	0.33	-0.94	0.59
Flower Garden	-0.24	0.52	-0.17	0.50	-0.79	1.32
Foreman	-0.51	0.08	-0.26	0.50	-0.84	0.59
Hall Monitor	-2.64	-1.66	-0.05	0.39	-0.82	0.54
Harbor	-0.34	0.32	-0.19	0.44	-0.82	0.75
Mobile Calendar	-0.43	0.54	-0.19	0.73	-0.89	0.64
Mother Daughter	-0.54	-0.31	-0.39	-0.14	-1.04	0.63
Soccer	-0.56	-0.03	-0.47	0.11	-0.87	0.70
Stefan	-0.5	0.42	-0.16	0.72	-0.97	1.21
Tempete	-0.47	0.28	-0.26	0.59	-1.10	0.86
Average	**-0.64**	**0.05**	**-0.24**	**0.44**	**-0.90**	**0.74**

quality of the video. Therefore, a high metric indicates that the subjective quality of the encoded video is better.

4.2.2. Coding performance

Table 2 depicts the performance of the proposed perceptual-based HEVC optimization algorithm compared with the standard HM16.7 under the same setting. In our experiment, we divide each frame into salient and non-salient regions. According to the saliency map, if the value of current pixel larger than the average value of the saliency map, the pixel belongs to the salient region and vice versa. The PSNR for the salient and non-salient regions are calcualted, respectively. It can be seen from the experimental results that the algorithm obtains 1.85 dB PSNR improvement in the salient regions, and 1.27 dB PSNR reduction in non-salient regions. The PSNR of the whole video sightly drop by 0.82 dB. In general, viewers pay more attention to the prominent areas, while the remaining areas rarely attract the viewers' attention. The PSNR reduction in non-significant regions has little effect on the visual experience. Additionally, the bit rate dropped by 1.39% and the encoding time increased by an average of 26.85%. In the experiment, we used low-dimensional processing to process high-resolution images and project them into low-dimensional space, using the DCN and segmentation technique to extract the static and dynamic saliency maps. Therefore, for high resolution video, the encoding time will only increase slightly. However, for low-resolution video, the encoding time increases because most of the time is spent on saliency map extraction.

Video Multi-method Assessment Fusion (VMAF) is a perceptual video quality assessment metric proposed by Netflix [52].

Table 4
Average VMAF scores for each class of video sequences.

VMAF	Class A	Class B	Class C	Class D	Class E
HEVC	95.18	93.60	91.79	93.75	92.77
Proposed	95.74	94.45	92.80	94.76	93.45

VMAF has a high correlation to human perceptual quality. The average VMAF scores for each video class are also calculated, which is illustrated in Table 3. As the HEVC standard does not consider the perceptual information during the coding process, our method obtains higher VMAF scores for the video sequences.

To describe the coding performance intuitively, Fig. 7 illustrates the rate-distortion (RD) curves of six test sequences. It can be observed that the proposed algorithm obtain a higher PSNR value at the salient regions from low to high bitrate compared with HM 16.7. This improvement is achieved at the expense of degrading the coding quality of non-significant areas. The RD curves of each video sequence almost overlap. The experimental results show that compared with HM16.7, this method achieves almost the same coding quality on the whole image from low bit rate to high bit rate. At the same time, it significantly improves the quality of the highlighted areas of the reconstructed video and enhances the viewing experience.

Table 4 shows the performances of our algorithm compared with the methods of Hadizadeh [48] and Zhu [49]. Comparison results are given in terms of BD-EWPSNR and BD-PSNR. It can be observed that the proposed method get 0.74 dB EWPSNR improvement compared with HEVC, 0.69 dB compared with Hadizadeh method, and 0.30 dB compared with Zhu. Experimental results demonstrate that our method can obtain a better visual experience.

To obtain a more intuitive evaluation for conversational video, Figure 8 depicts the reconstructed frames of four video sequences encoded by HM16.7 and the proposed algorithm. The salient maps are also attached. It can be observed that the proposed method has more details on the human face than HM16.7. Also, for the Rollercoaster sequence, it can be observed that our method has high coding quality for the moving object region. According to the salient map, the ω_i and Q_i for the i-th CTU are calculated by solving formula (16) and (17), respectively. If the CTU is more sensitive to viewers' attention, a larger ω and a smaller QP will be obtained, and vice versa. As a result, more bits will be allocated to the salient regions. Figure 9 illustrates the QP and bit heat map for PartyScene sequence. It can be observed that the distribution of QP and bit is consistent with the salient map. Experimental results demonstrate that the proposed content-aware bit allocation algorithm can effectively improve the coding quality for static and dynamic saliency regions. People get a better visual experience.

5. Conclusion and discussion

In this paper, we present a content-aware HEVC coding approach to improve the coding quality of salient regions. It is proved that, in video, viewers' focus their attention on specific areas. Therefore, our method gives inequality importance to emphasize coding quality in the salient regions. Firstly, we train a DCN model to identify multiple semantic regions and generate a static saliency map. Moreover, moving objects in videos also attract viewers' attention. We develop a segmentation technique to extract the moving object region of a video, which will be regarded as the dynamic saliency map. The static saliency and dynamic saliency maps are fused together to guide the video coding. According to the saliency map, we exploit a bit rate allocation scheme by adaptively adjusting the QP value. The visual experience will be improved by allocating more bits to the salient regions and less bits to the non-salient regions.

Compared with feature-based or color-based methods, our technique is more in line with viewers' visual characteristics. Additionally, for sports videos, our algorithm considers moving objects as salient regions. The proposed method can be used for various types of video, for instance, conventional, conversational or sports videos. The visual experience can largely be enhanced.

Currently, the rate control scheme of our proposed method is at the CTU level. Therefore, in low-resolution video, it is difficult to significantly improve the visual quality of the salient region because the size of the salient region may even be smaller than the size of the CTU. Designing a CU-level rate-control scheme is a promising research topic for future work. Additionally, we need to further optimize the salient detection approach to reduce the complexity of the method. Moreover, neural networks are also emerging for dynamic object segmentation. Future study will concentrate on designing a DCN to extract the static and dynamic saliency maps simultaneously.

ACKNOWLEDGMENT

The author would like to thank the editors and anonymous reviewers for their valuable comments. This work was supported by National Natural Science Foundation of China No. U1713211, and the Research Grant Council of Hong Kong SAR Government, China, under Project No. 11210017, awarded to Prof. Ming Liu.

References

[1] Aisheng Yang, Huanqiang Zeng, Jing Chen, Jianqing Zhu, and Canhui Cai. Perceptual feature guided rate distortion optimization for high efficiency video coding. *Multidimensional Systems & Signal Processing*.

[2] Huanqiang Zeng, Aisheng Yang, King Ngi Ngan, and Miaohui Wang. Perceptual sensitivity-based rate control method for high efficiency video coding. *Multimedia Tools & Applications*, 75(17):10383–10396, 2015.

[3] Yueying Wu, Pengyu Liu, Yuan Gao, and Kebin Jia. Medical ultrasound video coding with h. 265/hevc based on roi extraction. *PloS one*, 11(11):e0165698, 2016.

[4] Aisheng Yang, Huanqiang Zeng, Jing Chen, Jianqing Zhu, and Canhui Cai. Perceptual feature guided rate distortion optimization for high efficiency video coding. *Multidimensional Systems and Signal Processing*, 28(4):1249–1266, 2017.

[5] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Semantic perceptual image compression using deep convolution networks. In *2017 Data Compression Conference (DCC)*, pages 250–259. IEEE, 2017.

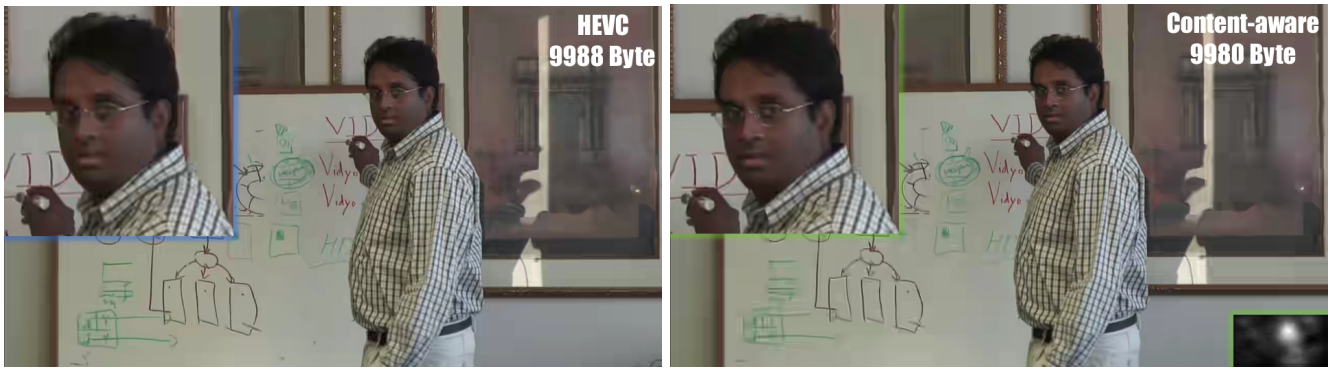
[6] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015.

[7] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 40(1):20–33, 2018.

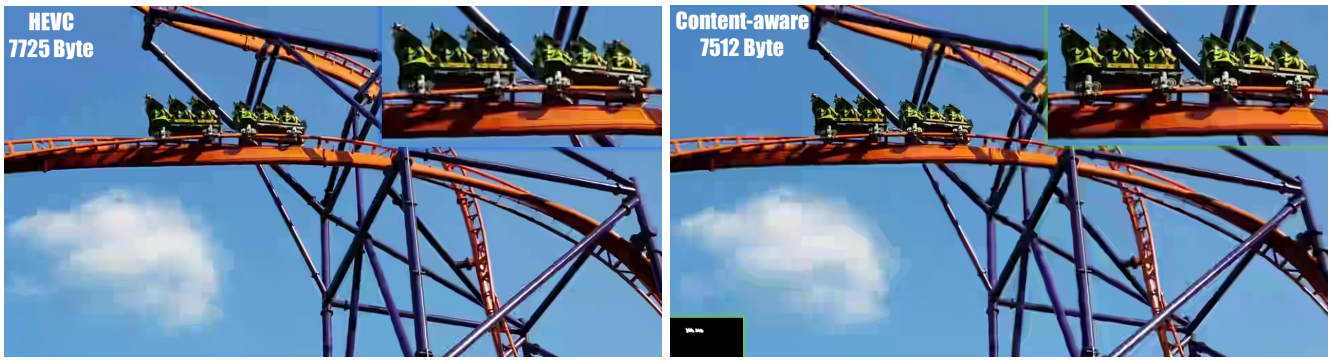
[8] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, volume 2, page 8, 2014.

(a)

(b)



(c)



(d)

Figure 8: Coding performance of the proposed content-aware method compared with HEVC algorithm: (a) BQMALL, (b) PartyScene, (c) Vidyo3, (d) Rollercoaster.

[9] Huaying Xue, Yuan Zhang, and Yunong Wei. Fast roi-based hevc coding for surveillance videos. In *2016 19th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pages 299–304. IEEE, 2016.

[10] Holger Meuel, Julia Schmidt, Marco Munderloh, and Jörn Ostermann. Region of interest coding for aerial video sequences using landscape models. *Edited by Yo-Sung Ho*, pages 51–77, 2012.

[11] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceed-*

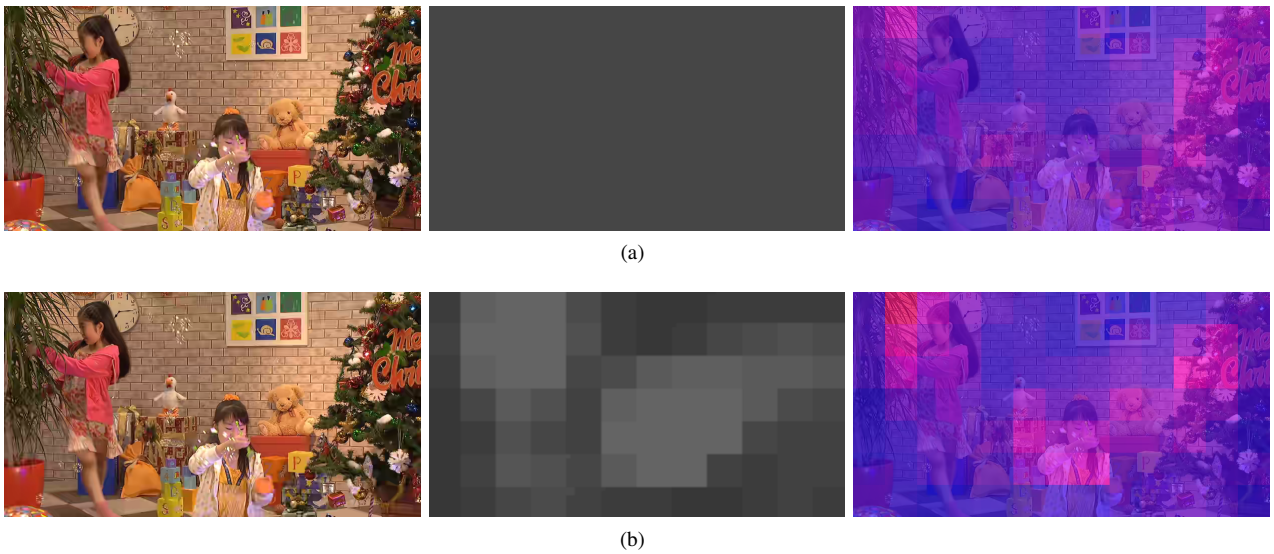


Figure 9: Coding result, QP distribution map, and bit distribution heat map: (a) HEVC (b) Proposed method.

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 724–732, 2016.

[12] QiLui Zhou, Jiaying Liu, and Zongming Guo. A multilevel region-of-interest based rate control scheme for video communication. In *MIPPR 2009: Remote Sensing and GIS Data Processing and Other Applications*, volume 7498, page 74984W. International Society for Optics and Photonics, 2009.

[13] Shengxi Li, Mai Xu, Xin Deng, and Zulin Wang. Weight-based r-λ rate control for perceptual hevc coding on conversational videos. *Signal Processing: Image Communication*, 38:127–140, 2015.

[14] Xin Deng, Mai Xu, and Zulin Wang. A roi-based bit allocation scheme for hevc towards perceptual conversational video coding. In *2013 Sixth International Conference on Advanced Computational Intelligence (ICACI)*, pages 206–211. IEEE, 2013.

[15] Mai Xu, Xin Deng, Shengxi Li, and Zulin Wang. Region-of-interest based conversational hevc coding with hierarchical perception model of face. *IEEE Journal of Selected Topics in Signal Processing*, 8(3):475–489, 2014.

[16] Bing Xiong, Xiaojiu Fan, Ce Zhu, Xuan Jing, and Qiang Peng. Face region based conversational video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(7):917–931, 2011.

[17] Victor Sanchez and Miguel Hernández-Cabronero. Graph-based rate control in pathology imaging with lossless region of interest coding. *IEEE transactions on medical imaging*, 37(10):2211–2223, 2018.

[18] David Yee, Sara Soltaninejad, Deborsi Hazarika, Gaylord Mbuyi, Rishi Barnwal, and Anup Basu. Medical image compression based on region of interest using better portable graphics (bpg). In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 216–221. IEEE, 2017.

[19] Heng Chen, Geert Braeckman, Shahid Mahmood Satti, Peter Schelkens, and Adrian Munteanu. Hevc-based video coding with lossless region of interest for telemedicine applications. In *2013 20th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 129–132. IEEE, 2013.

[20] Peiyin Xing, Yonghong Tian, Tiejun Huang, and Wen Gao. Surveillance video coding with quadtree partition based roi extraction. In *2013 Picture Coding Symposium (PCS)*, pages 157–160. IEEE, 2013.

[21] Chun-lei Zhao, Ming Dai, and Jing-ying Xiong. Region-of-interest based rate control for uav video coding. *Optoelectronics Letters*, 12(3):216–220, 2016.

[22] Zhengrong Xu, Mei Yu, Shuqing Fang, and Shengyang Xu. A new saliency based video coding method with hevc. In *2015 International Conference on Intelligent Systems Research and Mechatronics Engineering*. Atlantis Press, 2015.

[23] Huiqi Wang, Lin Wang, Xuelin Hu, Qin Tu, and Aidong Men. Perceptual video coding based on saliency and just noticeable distortion for h. 265/hevc. In *2014 International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pages 106–111. IEEE, 2014.

[24] Lixun Bai, Li Song, Rong Xie, Jianfeng Xie, and Min Chen. Saliency based rate control scheme for high efficiency video coding. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–6. IEEE, 2016.

[25] Henglu Wei, Wei Zhou, Rui Bai, and Zhemin Duan. A rate control algorithm for hevc considering visual saliency. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 36–42. IEEE, 2018.

[26] Zhicheng Li, Shiyin Qin, and Laurent Itti. Visual attention guided bit allocation in video compression. *Image and Vision Computing*, 29(1):1–14, 2011.

[27] Seung-Hun Nam, Wonhyuk Ahn, Seung-Min Mun, Jinseok Park, Dongkyu Kim, In-Jae Yu, and Heung-Kyu Lee. Content-aware image resizing detection using deep neural network. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 106–110. IEEE, 2019.

[28] Insung Hwang, Sang Hwa Lee, Jae Sung Park, and Nam Ik Cho. Saliency detection based on seed propagation in a multilayer graph. *Multimedia Tools and Applications*, 76(2):2111–2129, 2017.

[29] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, 2006.

[30] Xiaodi Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 34(1):194–201.

[31] Chenlei Guo, Ma Qi, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*, 2008.

[32] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE Trans Pattern Anal Mach Intell*, 34(10):1915–1926.

[33] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[35] Peng Yun, Lei Tai, Yuan Wang, Chengju Liu, and Ming Liu. Focal loss in 3d object detection. *IEEE Robotics and Automation Letters*, 4(2):1263–1270, April 2019.

[36] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[38] Teemu Kinnunen, Joni Kristian Kamarainen, Lasse Lensu, Jukka Lankinen, and Heikki KÃďlviÃďinen. Making visual object categorization more challenging: Randomized caltech-101 data set. In *International Conference on Pattern Recognition*, 2010.

[39] F. Bossen. Common test conditions and software reference configurations. *document JCTVC-L1100*, 2013.

[40] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[41] Marius Leordeanu, Rahul Sukthankar, and Cristian Sminchisescu. Efficient closed-form solution to generalized boundary detection. In *European Conference on Computer Vision*, pages 516–529. Springer, 2012.

[42] Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):500–513, 2010.

[43] Zhenyu Liu, Libo Wang, Xiaobo Li, and Xiangyang Ji. Optimize x265 rate control: An exploration of lookahead in frame bit allocation and slice type decision. *IEEE Transactions on Image Processing*, pages 1–1.

[44] Mingliang Zhou, Xuekai Wei, Shiqi Wang, Sam Kwong, Chi-Keung Fong, Peter Wong, Wilson Yuen, and Wei Gao. Ssim-based global optimization for ctu-level rate control in hevc. *IEEE Transactions on Multimedia*, pages 1–1.

[45] Gary J Sullivan and Thomas Wiegand. Rate-distortion optimization for video compression. *IEEE signal processing magazine*, 15(6):74–90, 1998.

[46] Stéphane Mallat and Frédéric Falzon. Analysis of low bit rate image transform coding. *IEEE Transactions on Signal Processing*, 46(4):1027–1042, 1998.

[47] Zhenzhong Chen and C. Guillemot. Perceptually-friendly h.264/avc video coding based on foveated just-noticeable-distortion model. *IEEE Transactions on Circuits & Systems for Video Technology*, 20(6):806–819.

[48] Hadi Hadizadeh and Ivan V Bajić. Saliency-aware video compression. *IEEE Transactions on Image Processing*, 23(1):19–33, 2013.

[49] Shiping Zhu and Ziyao Xu. Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network. *Neurocomputing*, 275:511–522, 2018.

[50] Hadi Hadizadeh, Mario J Enriquez, and Ivan V Bajic. Eye-tracking database for a set of standard video sequences. *IEEE Transactions on Image Processing*, 21(2):898–903, 2011.

[51] G Bjøntegaard. Calculation of average psnr differences between rd-curves (vceg-m33). In *VCEG Meeting (ITU-T SG16 Q. 6)*, pages 2–4, 2001.

[52] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6, 2016.