Visual Saliency Estimation Via HEVC Bitstream Analysis

Submitted by:

XU DAI

For the degree of

Master of Philosophy

The University of Sheffield

Department of Electronic and Electrical Engineering

September 8th, 2017

Abstract

Since Information Technology developed dramatically from the last century 50's, digital images and video are ubiquitous. In the last decade, image and video processing have become more and more popular in biomedical, industrial, art and other fields. People made progress in the visual information such as images or video display, storage and transmission. The attendant problem is that video processing tasks in time domain become particularly arduous.

Based on the study of the existing compressed domain video saliency detection model, a new saliency estimation model for video based on High Efficiency Video Coding (HEVC) is presented. First, the relative features are extracted from HEVC encoded bitstream. The naive Bayesian model is used to train and test features based on original YUV videos and ground truth. The intra frame saliency map can be achieved after training and testing intra features. And inter frame saliency can be achieved by intra saliency with moving motion vectors. The ROC of our proposed intra mode is 0.9561. Other classification methods such as support vector machine (SVM), k nearest neighbors (KNN) and the decision tree are presented to compare the experimental outcomes. The variety of compression ratio has been analysis to affect the saliency.

Acknowledgements

First and foremost, I would like to show my deepest gratitude to my supervisor Dr. Charith Abhayaratne, a respected scholar who has provided me with valuable guidance in my research period. Without his enlightening instruction, impressive kindness and patience, I could not have completed my study.

Also, I would like to thank everyone in our research group and my friend Zheng Hui for their support. Finally, I would like to thank my parents and my every family members. They give me very big support.

Contents

List of Figures

List of Tables

Chapter 1. Introduction

With the rapid development of information technology, image video information is growing and expanding. Computers are a useful tool in information analysis and large amounts of data processing. However, the speed of multimedia data growth is much greater than the speed of computer processing performance [1]. In addition, People only concern a part of a given image (or a video). This is because the amount of data in the image/video is beyond the processing power of the human eye [2]. Therefore, the ability to predict where humans' attention becomes a popular research content.

In recent years, network bandwidth and storage have increased rapidly, but it is far from meeting the requirements for the transmission and storage of massive video data. Therefore, the efficient compression of video information is one of the important technical measures to solve this contradiction. Video compression technology has been concentrated in the past two to three decades. From the first video coding standard H. 261 / MPEG-1 [3], the second generation of video coding standard H. 264 / AVC [3], and the third generation of video coding standard High Efficiency Video Coding (HEVC), the efficiency of video compression for each generation greatly increased.

This thesis focus on the saliency estimation in HEVC compression domain.

1.1. Motivations

With the rapid development of science and information technology, video resolution is from early 176 x 144, 352 x 288, 416 x 244, 720 x 480, 1280 x 720, and 1920 x 1080. Now ultra-high-definition (3840x2160) videos also cut a striking figure. With the increasing amount of video data, the existing video processing technology is not mature enough to deal with such a huge amount of data, real-time video processing tasks become particularly arduous.

Currently, video saliency detection has been widely applied to many video processing applications such as video compression based on video highlighting, video classification, video watermarking, video transcoding, and video scaling. Visual attention analysis simulates the human eye vision system by automatically detecting the salient region of the acquired image. In image processing, the research of video saliency is practically important.

Computer
Webpage
Television
Compressed Video
Gaming console
Mobile/ Cell phone

Figure 1-1 The applications of compressed video.

1.2. Aims and Objectives

The main aim of this thesis is to estimate video saliency using Bayesian modeling of HEVC features in the compressed domain.

Our model is used in video saliency estimation. This can be summarized by the following objectives:

To research the state-of-the-art of saliency estimation methods in both image and video domain. The gaps of the existing methods are analyzed.

To study the High Efficiency Video Coding (HEVC). The decoder part of HEVC is introduced. The good standing of HEVC decoder can help us to find the relationship between HEVC bitstream features and saliency.

Video saliency estimation via HEVC feature. To propose a video saliency model within features extracted from HEVC compressed domain. All extracted features are analyzed and evaluated in different HEVC compressed mode. A naïve Bayesian classifier is trained for saliency region estimation in videos.

Different machine learning classification algorithms such as support vector machine (SVM), k nearest neighbors (KNN) and the decision tree are presented to compare the experimental outcomes. The variety of compression ratio has been analysis to affect the saliency.

1.3. Contribution

A new proposed video saliency model in the compressed domain is proposed in this thesis. The relative intra and inter features such as block size, residuals, intra mode difference, motion vectors are extracted from HEVC decoder. Both intra mode saliency and inter saliency are achieved. The novelty of this model is to achieve saliency map in HEVC compressed domain without fully decoded bitstream. This method can be used in image/video searching and visual interface design.

Bayesian model is used for training and classification for the salient regions for each frame. The testing videos are divided into two parts: half of testing videos are used to training and half of them are used to classification. The basic theory of Bayesian model is classifying features to maximum probability category. The variety of compression ratio has been analysis to affect the saliency.

1.4. Thesis Outline

The reminder of the thesis is structured as the following chapters. The contents of each chapter are summarized.

In chapter 2, the state-of-the-art for visual saliency is reviewed and presented. The chapter includes background of visual attention and visual attention models (VAM) in single image or a video. The comparison of video saliency estimation in the compressed domain is also provided. The applications of visual attention model are presented. General features used in the saliency estimation are introduced in detail. Then an overview of HEVC coding system is provided. Some important HEVC decoder block coding steps is introduced briefly.

Chapter 3 analyzes and evaluates all extracted features are in different HEVC compressed mode. Then, a new video saliency method in video compressed domain is proposed. The intra saliency estimation is achieved using naïve Bayesian classification by relative HEVC features. The inter saliency is achieved by intra saliency moving via motion vector. Different machine learning classification algorithms such as support vector machine (SVM), k nearest neighbors (KNN) and the decision tree are presented to compare the experimental outcomes. The variety of compression ratio has been analysis to affect the saliency. The results are evaluated by ROC curve.

Chapter 4 includes a brief summary of the whole thesis. Some suggested further work is provided.

Chapter 2. Literature Survey

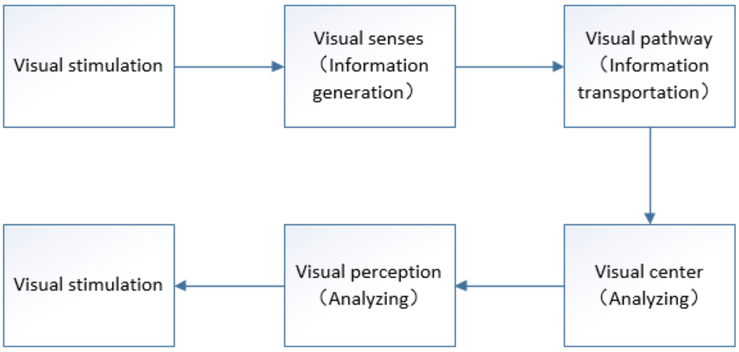
This chapter contains background of human visual system and visual attention models. Some existing image and video saliency models are presented. The applications of visual attention in different areas are discussed. The image/video features are introduced.

2.1. Visual attention mechanism

With the rapid development of information technology, images and videos have become an important carrier of information. Processing and analysis digital image and video data efficintly has become the central issue.

2.1.1. Human Visual System

On the human visual perception system and visual nervous system, psychology and other related fields experts have carried out long-term exploration [4] and research. Through deeply research and exploration, visual sensory information in the visual nervous system is in accordance with a fixed path to be transmitted. The input is visual stimulation, the output is visual perception. The human visual system is mainly composed of visual sensory [5], visual pathway, visual central nervous system [6] and visual perception central organization.



Visual stimulation
Visual senses（Information generation）
Visual pathway（Information transportation）
Visual center（Analyzing）
Visual perception（Analyzing）
Visual stimulation

Figure 2-1: Information perception processing of Human visual perception system.

The average diameter of person's eyes is about 24 mm [7]. The human eye is approximately spherical and consists of two parts: the eye wall and the eyeball. Cornea and sclera are located in the outer layer of the eye wall, in which the cornea with refractive effect [54], can be reflected the light to the eyes, scleral protects the eyeball. The middle layer of the eye wall consists of iris and choroid, and the inner retina is composed of cone cells and rod cells. Visual information is transmitted as follows: visual stimulation from the light sensory cells, effect on the retina [8], and then through the optic nerve, optic tract and subcortical center, finally reach the visual cortex, causing visual perception. The so-called visual sensation refers to the light brightness; visual perception refers to the color, shape and other characteristics.

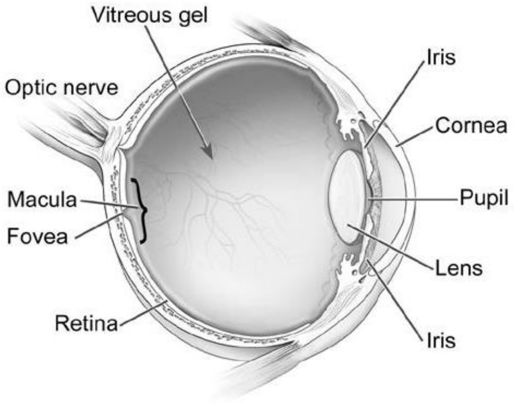


Figure 2-2Human eye structure schematic diagram [57].

The cornea of the eye is a transparent, highly curved refraction window. The cornea of the eye is a transparent, highly curved refraction window, and then partially blocked by the opaque iris surface. The pupil changes with the intensity of the light. Under normal lighting conditions, pupil is in a 4contraction state, to avoid the blurring caused by spherical aberration.

2.1.2. Visual attention mechanism model

Visual attention is essentially a biological mechanism which can select from the complex environment, and gradually remove the relatively unimportant information [9]. In this way, the complex external scene can be simplified and decomposed. The advantage of this mechanism is that: it allows us to focus the important information and objects rapidly in an environment [53].

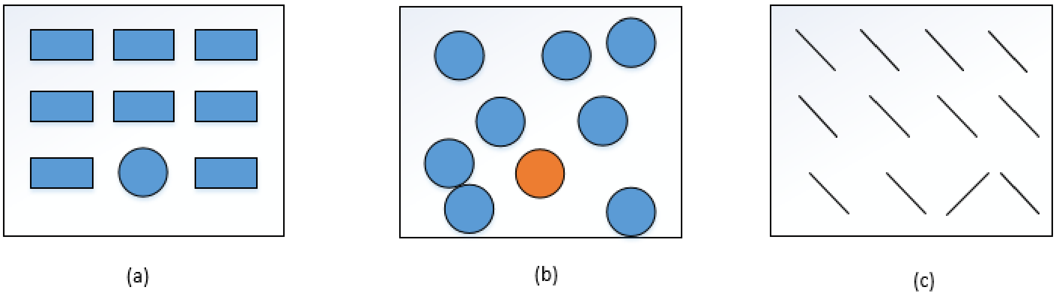


Figure 2-3 Illustration of Visual Attention. a) Intensity contrast b) Colour contrast c) Orientation contrast.

Researchers do a lot of exploration about the applications of visual attention in physiological sciences, psychological science, neuroscience and information science field. The essence and characteristics of visual attention include six aspects [10]:

- Selectiveness: Selecting part of information.
- Concentration: Excluding unrelated stimuli.
- Search: Looking for part of the targets.
- Activation: To cope with all possible stimuli.
- Set: To accept and respond to specific stimuli.
- Vigilance: Keeping long attention.

Perry and Hodges study the mechanism of visual attention in the perspective of neuroscience. The manifestations of visual attention are [15]:

- Selective Attention and Shifting: The characteristic is to focus on dealing with a relevant stimulus at the same time while ignoring or discarding other irrelevant stimuli.
- Sustained Attention: It is characterized by a long period to maintain concentration attention.
- Divided Attention: It is characterized by distributing attention at the same time, focuses on dealing with multiply related stimuli.

Harris argues that "concentration" and "vigilance" are the most basic features of the attention mechanism and, based on which visual attention is divided into four types:

- Selective Attention: Used to select part of the visual information to meet the brain's limited information processing needs.
- Parsing Attention: Used to separate the target from the background for pattern recognition.
- Directing Attention: Used to guide the emergency interruption, the normal detection, and maintenance of visual attention and other acts of switching;
- Alertness Attention: Used to wake up the potential visual information processing.

2.2. Psychology and neurobiology of visual attention

The main research of visual selective attention mechanism on psychology and neurobiology in the following aspects.

2.2.1. Bottom-up and top-down selective attention

The visual selective attention mechanism can be generalized in two aspects: bottom-up selective attention [11] and top-down selective attention [11]. Bottom-up [12] selective attention is driven by pure external stimuli, such as a strong contrast. The top-down selective attention is controlled by the subject, which is controlled by high-level brain information such as knowledge, expectations, and goals [12]. In the same scene, different people get the results of attention are different. Current research on top-down factors is limited and often manifested in obtaining knowledge related to target objects. Other top-down factors such as motivation, expectation, emotion, etc. are more difficult to control and analyze.

2.2.2. Explicit attention and implicit attention

In general, explicit attention refers to the transfer of the region of interest, which is associated with the movement of the eye. Because the center of the human eye has a high resolution but a low resolution in the surroundings, the explicit attention happens when the area of selective visual attention falls away from the place where the current fixation point is distant or even outside the field of view [13, 52]. In addition, there is another situation, which is not accompanied by the transfer of gaze. In 1890, William James pointed out that we can move attention without eye movement [52]. This phenomenon is implicit attention. For example, when you walk down the street to see the front, you will not bump the pedestrian next to you. Implicit attention is more efficient than explicit attention.

2.3. Visual searching

Visual searching is an important tool in the study of visual attention. In the visual psychology experiment, the subjects are required to find certain requirements of the target around many interference objects [14]. One of the parameters of the visual searching efficiency is the reaction time. The research demonstrated that the response time almost unchanged as the number of interfering objects increases sometimes. However, in some cases, the reaction time increases rapidly as the number of interfering objects increases.

There are many theories to explain the essential difference between the selective attention mechanism in the efficient and inefficient search. The influential theory is Treisman's characteristic fusion theory. Feature fusion theory states that when the target can be distinguished by a feature, and the other disturbances have the same characteristics, the target is detected easily, quickly and in parallel. That theory holds that each object is processed in parallel in an efficient visual searching, while the object is processed serially during an inefficient searching. When the difference between the target object and the interfering objects is a combination of multiple features, the search belongs to serial searching.

2.3.1. Inhibition mechanism in visual attention

The inhibition of selective attention is widely studied. One of the most common studied phenomena is the inhibition of return. It means that the respondent has slowed down the target response that had previously been repeated in the same position. It has been argued that the inhibition of return mechanism plays an important role in maintaining spatial selectivity [15]. It enables the tester to return to the location where the target information has been extracted. Inhibition of return has important implications, for example, people do not repeatedly detect the same interfering object in visual search. inhibition of return [22].

2.3.2. Resolution and multi-scale of selective visual attention

The resolution of selective visual attention refers to the ability to distinguish a single object in a number of closely arranged objects. It is common to use a plurality of interfering objects around the target to investigate how far the interfering object will affect the test object [16, 56]. It is used to study the resolution of the selective attention. When the distance between two objects is less than the resolution, people will not notice the single object. This phenomenon is called the Crowding Effect [51]. Another problem with this is the multi-scale problem. Human visual perception is carried out at multiple scales. For example, when you want to get a book, you will first notice the bookcase and then find your book. The scale of searching is different.

2.4. Visual attention mechanism model

Visual saliency have become a prerequisite for many computer vision algorithms. This includes selecting a block in the scene or selecting a combination of some features in the area. Selective visual attention mechanism is an effective way to improve the real-time complexity and to solve the problem of information explosion [17].

2.4.1. Filter model

Broadbent proposed a filter model in 2001 [18]. The number of visual information input channels is greater than before. However, only one channel is accessed through the filter into the advanced analysis phase. This filter reflects the selection of visual attention. Once the information exceeds human's acceptable capacity, the filter will lock the redundant information. Only that information through the filter can be analyzed.

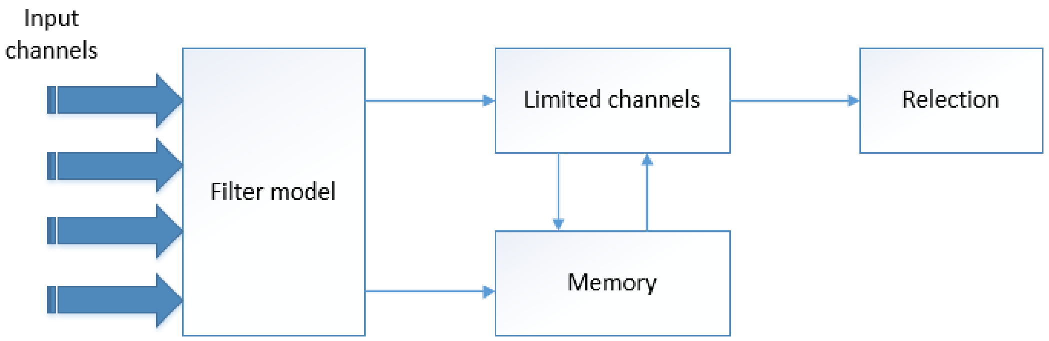


Figure 2-4 Illustration of Filter Model.

2.4.2. Response selection model

Deutsch proposed response selection model [19]. All the visual information of multiple input channels can enter the advanced analysis phase. The perceptual processing is obtained. Visual attention is not the choice of visual stimulus, but rather the choice of response to stimulation.

2.4.3. Resource allocation model

Kahneman proposed the resource allocation model [20]. It thought visual attention is essentially a resource allocation mechanism. It allocates human limited information processing capacity (resources) according to a resource allocation scheme under various constraints. The selectivity of visual attention is manifested through this resource allocation scheme.

2.4.4. Binary theory

The binary theory was proposed in the 1990s [21]. Visual information processing methods are divided into control processing and automatic processing. The characteristics of control processing are:

- It needs to pay attention to participation.

- Under the conscious control.
- Limited capacity.
- Slow and flexible.

The characteristics of automatic processing are:

- It does not need to pay attention to participation.
- Free from conscious control.
- Great capacity.
- Faster but lack of flexibility.

2.5. Basic features of image and video

2.5.1. Brightness

The luminous intensity of the visual scene is called brightness. The human eye has a strong sensitivity to the brightness, which determines the brightness is an important factor of saliency detection. In the field of digital image processing, grayscale [23] can be used to represent the brightness. The brightness level of the image is presented using black with different saturation. The brightest represent by white and the darkest represent by black. Each grayscale object has a luminance value in the range 0% -100%. 0% represent white and 100% represent black. For 8-bit images, the luminance values are quantized and the range can be normalized to [0,255] intervals. The 256 discrete gray scale represents the different brightness of the image.

2.5.2. Color

For different wavelengths of visible light, the human visual system has different subjective feelings. The different performance lies in different colors. Color is one of the most important elements of the image. The human visual system is more sensitive to color. In the image processing, the color characteristics can be described by the existing color space. Different color spaces are available for different applications. Color TV uses YUV as the color space, in order to use the brightness signal Y to solve the compatibility issues. So that black and white TV can also receive color TV signals.

2.5.3. Three primary colors

The human eye has a different sensitivity for different colors because the human eye has several kinds of conical photoreceptor cells. These cells are most sensitive to green, yellow-green and violet light. Their wavelengths are 534nm, 564nm, and 420 nm respectively [24]. If the stimuli of yellow-green photoreceptor cells is greater than the stimuli of green photoreceptor cells, people will recognize yellow; if the stimuli of yellow-green photoreceptor cells is much higher than the stimuli of green photoreceptor cells, people will recognize red. Although the three photoreceptors are not the most sensitive to green, red and blue. These three colors can stimulate the photoreceptors. Therefore, the green, red and blue as the basis of color perception. These three colors called primary colors.

2.5.4. RGB color space

Based on the principles of the human visual system described above, red, green, and blue colors are set as the reference colors of RGB color space [25]. Variety of different colors are formed by different weights between three colors. RGB color space almost includes all perceived colors.

2.5.5. YUV color space

In YUV color space, "Y" is the luminance, that is, the gray scale value. "U" and "V" are the color components. The specific part of the RGB signal is superimposed to establish the luminance signal. The RGB color space and the YUV color space conversion are shown below [26]:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & -0.00093 & 1.401687 \\ 1 & -0.3437 & -0.71417 \\ 1 & 1.77216 & 0.00099 \end{bmatrix} \begin{bmatrix} Y \\ U-128 \\ V-128 \end{bmatrix}. \quad (2\text{-}1)$$

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.5 \\ 0.5 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix}. \quad (2\text{-}2)$$

2.5.6. Texture

Compared with intensity and color features, texture belongs to the advanced saliency characteristics. The external manifestation of the surface change or distribution of the object is called texture. The texture is often expressed as color or light of a regular change [27]. It can be seen that the human vision system can quickly determine the surface with different textures. However, it is difficult to know how the human visual system is handled. In addition, it is difficult to use language or text to describe in detail [28]. One of the most popular view states that the texture primitives form a texture according to a regular distribution such as zebra or tiger body stripes. Usually, this law has a certain uniformity, repeatability, and directionality [62, 63]. The above properties are also the basis of texture analysis.

Texture analysis methods are generally region-based because texture has a strong regional nature. Texture analysis refers to the use of a certain image processing technology is to extract the texture parameters, which can be a quantitative or qualitative description of the texture processing. Texture analysis can be divided into three types: structural methods, statistical methods, and spectrum methods. Structural methods refer to the analysis of the structure of the region to get the texture elements, and then use the texture elements to describe the texture of the image. Statistical methods refer to the analysis of the texture properties of the color distribution in the observation area. The main algorithm is random field model, random classification model and gray level co-occurrence matrix. Spectrum methods refer to the Garbo transform, Fourier transform or wavelet transform to obtain the coefficients to describe the texture. The Tamura parameter and the gray level co-occurrence matrix are more effective spectrum method.

2.5.7. Motion

Motion as a description of the video content cannot be ignored in computer vision. Motion is also an indispensable feature for extracting video prominence regions [28]. People tend to concern objects with fast and strenuous moving. The features for the image mainly focus on color, brightness, texture; the main features of the video focus on the motion.

For non-compressed domain videos, the main methods of motion detection are optical flow [62] and frame difference [63]. The optical flow method can be considered as the motion of the image luminance mode in the video sequence, also it can be considered as the representation of the velocity of the motion on the surface of the object. The frame difference method is the subtraction between the adjacent frames; the difference is taken as the motion information.

For the compressed domain videos, inter frame predictive coding is used to remove temporal redundancy between adjacent frames. Inter frame prediction coding mainly uses the time domain correlation between successive frames to eliminate the time redundancy by motion compensation. Each frame in the video is divided into blocks, and each block is encoded. The target frame encoder requires the use of blocks of previously encoded video frames. The relative position difference between the target code block and the reference code block is called motion vector. The motion vector is the motion information of the compressed domain video. Therefore, the motion vector in the bitstream can be extracted. In addition, using motion vector need to take into account the camera movement and other factors. In the event of camera shake, the background has motion relative to the foreground, thus affecting the motion vector accuracy.

2.6. Saliency model

In visual attention models, saliency map is used to define the visual attention area. Saliency map is a two-dimensional map of the same size as the original image, where each pixel value represents the visual attention of the corresponding image. Many methodologies in visual attention estimation of both Static or dynamic scene are presented in this section.

2.6.1. Static image saliency

The first image saliency estimation algorithm is proposed by Koch and Ullman [11]. Although this model has not yet been implemented at the beginning of the presentation, it provides an algorithmic basis for future implementation. Many of current visual attention models are based on Koch and Ullman's work. The main idea of this saliency estimation algorithm is to combine a number of features in parallel with a feedforward model. WTA (Winner-Take-All) competitive neural network is used to determine the most salient area. Then a suppression of return mechanism is used to move the gaze into next the most salient area. Koch and Ullman [5] proposed the WTA neural network (the neural network that determines the most salient area in topographic map) and the concrete description of the implementation. This WTA neural network approach has a strong biological motivation and shows how the human brain may achieve VA mechanism. However, for computer systems, the WTA is not necessary because there is a simpler way to determine the most salient area.

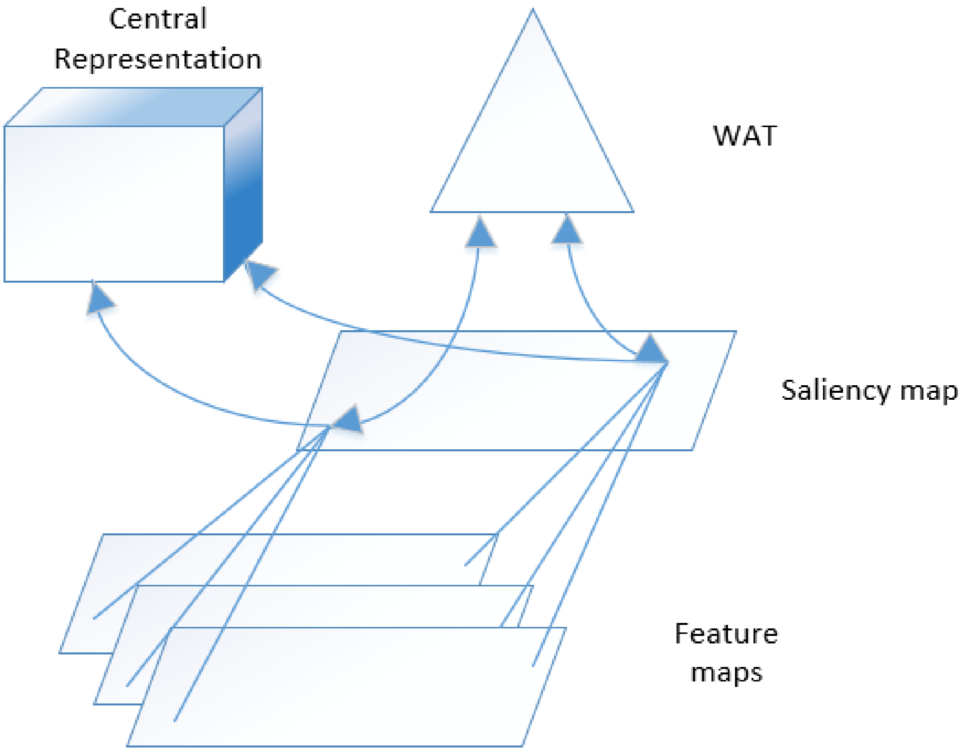


Figure 2-5 Illustration of Koch & Ullman selective processing model.

The first completed implementation of Koch and Ullman's hypothesis is proposed by Koch & Ullman in 1994 [12]. Since then, visual attention estimation becomes more and more popular. The Neuromorphic Vision Toolkit (NVT) is one of the most well-known systems which is derived by Koch & Ullman. NVT has been developed over the years by Itti research groups. This model has developed under the efforts of the research team centered on Itti et al. Their models and implementation methods become the basis of many research groups. The feature maps, saliency maps, winner-take-all (WTA) neural network and Inhibition-of-return (IOR) are derived from Koch & Ullman model.

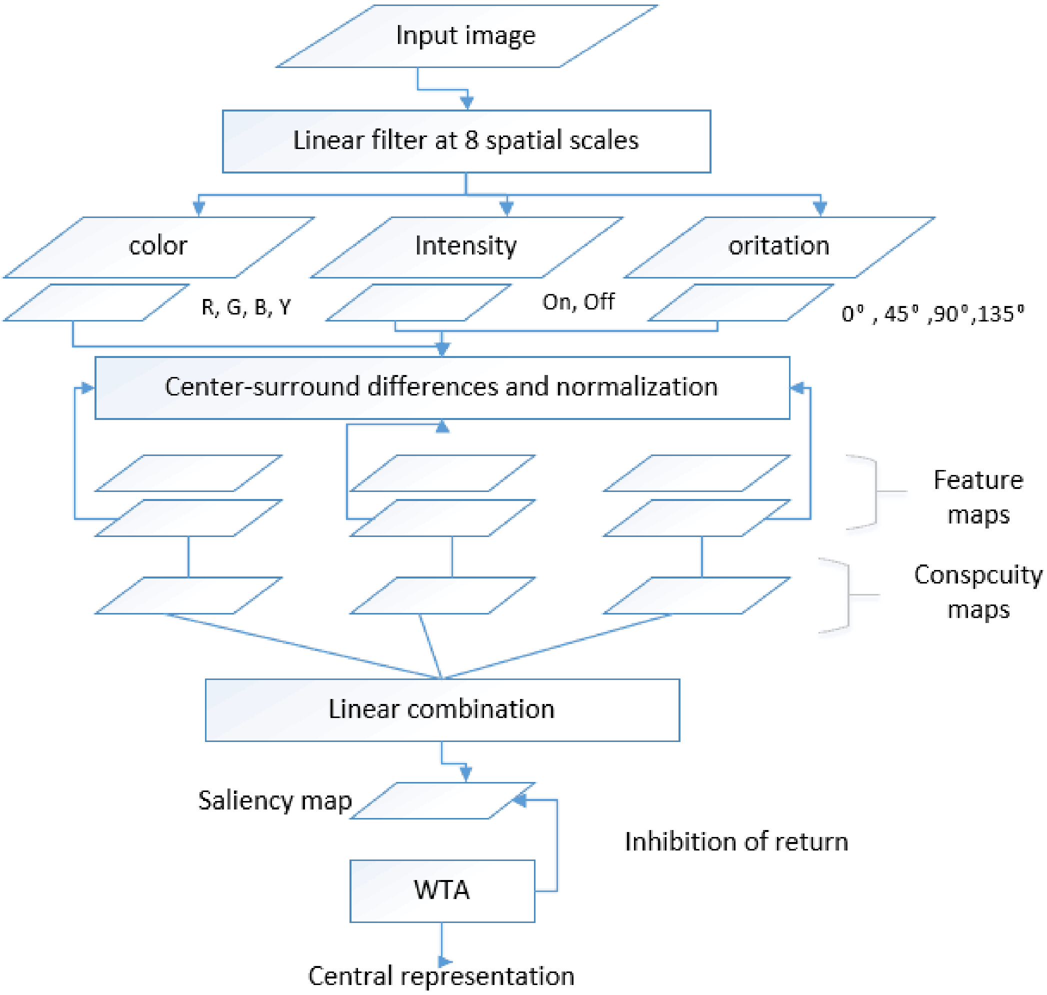


Figure 2-6 llustration of Itti selective visual attention model.

This model can be described as followings:

1. Feature maps extraction. They represent the gray, color, and gradient directions of the original image at six different scales. Each feature is calculated by the center-surrounding differences operator.
2. Then superimposing these feature maps as three conspicuity maps by weight normalizing. Then adding three conspicuity maps as a saliency map via linear combinations.
3. The inhibition of return system suppresses the salient area in this saliency map so that the attention position autonomously points to the next position.

Every feature map is calculated using a center surrounding structure similar to the biological receptive field. Central surrounding structure means that typical visual neurons are sensitive to small areas located in the center. In addition, stimuli in the wider, weaker regions around their central regions will suppress the response of the visual neurons. It is clear that such precise structures for local spatial discontinuities are particularly suitable for detecting areas that are prominent around their surroundings, which are also general principles of use in the retina, external geniculate and visual cortex.

Let r, g, b correspond to the red, green and blue channels of the input image, the gray scale image I is:

$$I = \frac{(r+g+b)}{3}. \quad (2\text{-}3)$$

In order to separate the chrominance signal from the intensity, gray scale is used to normalize the r, g, b channels, because the little brightness change in Chroma channel is difficult to recognize. So the normalization only applied in the position which the gray scale is greater than the maximum 1/10 on the original image, and other locations of the r, g, b value is assigned to zero. According to the normalized r, g and b, the established of four wide tuning of the color channel are:

$$\text{Red: } R = r - \frac{(g+b)}{2}. \quad (2\text{-}4)$$

$$\text{Green: } G = g - \frac{(r+b)}{2}. \quad (2\text{-}5)$$

$$\text{Blue: } B = b - \frac{(r+g)}{2}. \quad (2\text{-}6)$$

$$\text{Yellow: } Y = \frac{(r+g)}{2} - \frac{|r-g|}{2-b}. \quad (2\text{-}7)$$

Further, according to these color channels, four Gaussian pyramids $R(\sigma), G(\sigma), B(\sigma), Y(\sigma)$ can be established to multiple Gaussian blur, where $\sigma \in (0, 1, \dots, 8)$ is the scale. The Gabor pyramids $O(\sigma, \theta)$ used to represent the local orientation information, where $\sigma \in (0, 1, \dots, 8)$ and $\theta \in (0, 45^0,\ 90^0,\ 135^0)$. Flicker pyramid $F(\sigma)$ can be yielded.

Center surrounding operation is implemented as a difference between the fine and course of a given feature. Center-surrounding differences defined two scales: fine scale c and coarser scale s. Considering three kinds of features: grayscale, color and orientation. If the central peripheral differential operation is $\ominus$, the gray scale feature can be obtained from the following equation:

$$I(c,s) = |I(c) \ominus I(s)|. \quad (2\text{-}8)$$

Which $c \in \{2,3,4\}$ and $s = c + \delta,\ \delta \in \{3,4\}$.

In a human visual system, this feature is detected by neurons that are sensitive to bright central dark surroundings or dark central bright surroundings.

The second feature is related to color. In the human visual cortex, there are four kinds of space and color double opponent: red/green, green/red, blue/yellow and yellow blue color pairs. In this model, the corresponding feature map to the red/green or green/red color pairs is:

$$RG(c,s) = |R(c) - G(c)) \ominus (G(s) - R(s))|. \quad (2\text{-}9)$$

And the feature map corresponding to blue/yellow or yello/blue color pair is obtained from the following:

$$BY(c,s) = |B(c) - Y(c)) \ominus (Y(s) - B(s))|. \quad (2\text{-}10)$$

Local orientation information is achieved by applying Gabor pyramids $O(\sigma,\theta)$. Orientation feature maps $O(c,s,\theta)$ is obtained by:

$$O(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)|. \quad (2\text{-}11)$$

For each pixel of the input image, saliency map uses a scalar to characterize its saliency and to guide the selection of attention points based on the spatial distribution of the saliency. The difficulty of combining feature maps is that features are incomparable. Since all features are combined together, some salient targets that appear in some feature maps may be overwhelmed by a large number of noise or not salient objects. In the initial thesis of Itti et al., a normalization operator $N(.)$ is proposed to enhance the feature map with less salient peaks and to weaken the feature maps with large numbers of salient peaks. For each feature map, this operator includes:

1. Normalizing the feature values to a range of $[0\ldots M]$ to eliminate the amplitude difference depending on the features.
2. Calculating the global maximum M of the map and the average of the other local maximum $\bar{m}$.

3. Multiplying the feature map by $(M - \bar{m})^2$.

Only considering the local maximum $N(.)$, the useful regions of the feature maps can be focus on while ignoring the uniform areas. The difference between the global maximum and all local maximum mean value reflects the difference between the most interested area and the average interested area. If the difference value is large, the region of most interested will be highlighted; if the difference value is small, it indicates that the feature maps do not contain any salient region. The biological basis of $N(.)$ is that: It approximates the cortical inhibition mechanism.

The feature maps are grouped into three conspicuity maps. As shown in the following:

$$\bar{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} N(I(c,s)). \quad (2\text{-}12)$$

$$\bar{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} [N\big(RG(c,s)\big) + N(BY(c,s))]. \quad (2\text{-}13)$$

$$\bar{O} = \textstyle\sum_{\theta \in \{0,45^0,90^0,135^0\}} N(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} N(O(c,s,\theta))). \quad (2\text{-}14)$$

$\bar{I}$ means intensity, $\bar{C}$ means color and $\bar{O}$ means orientation. And $\oplus$ means accumulation point by point.

The reason for the establishment of three normalized saliency feature description is: similarity features are highly competitive while different features will independently play a role in the feature map. These three feature descriptions are further normalized and summed to obtain a saliency S:

$$S = \tfrac{1}{3}(N(\bar{I}) + N(\bar{C}) + N(\bar{O})). \quad (2\text{-}15)$$

In the later work of Itti et al. [12], A number of feature combinations were compared, and it was found that normalizing each feature graph to a fixed dynamic range would show poor detection performance when detecting prominent targets in complex background. A possible way to improve performance is to obtain a linear combination of the feature maps. Although this method can highly improve the detection performance, it brings a difficulty of different models used in one application. Itti et al. [12] proposed a new feature combination strategy, on the basis of normalization, to enhance the feature maps which have less salient peaks by the local nonlinear competition. This competition pattern is similar to the nonclassical inhibition observed by electrophysiology.

2.6.2. Video saliency estimation

Researchers extend spatial attention studies to videos that contains a lot of motion. Cheng et al. [30] proposed a model of video saliency that combines motion. As shown in Figure 2-7, this video visual attention model analyzes the horizontal and vertical motion of the pixels in every frame. The input video is segmented into shots that contain motion gradient information. Each frame is then divided into frame segments that do not coincide with each other. For each frame segment, the motion information of the recorded camera is used to generate saliency after wards. At the same time, each feature model (such as brightness, color, motion, etc.) are calculated. Depending on the motion information of the camera, combinations of these feature maps. Finally, the overall saliency map was constructed. The area of visual attention in the video is represented by the saliency distribution.

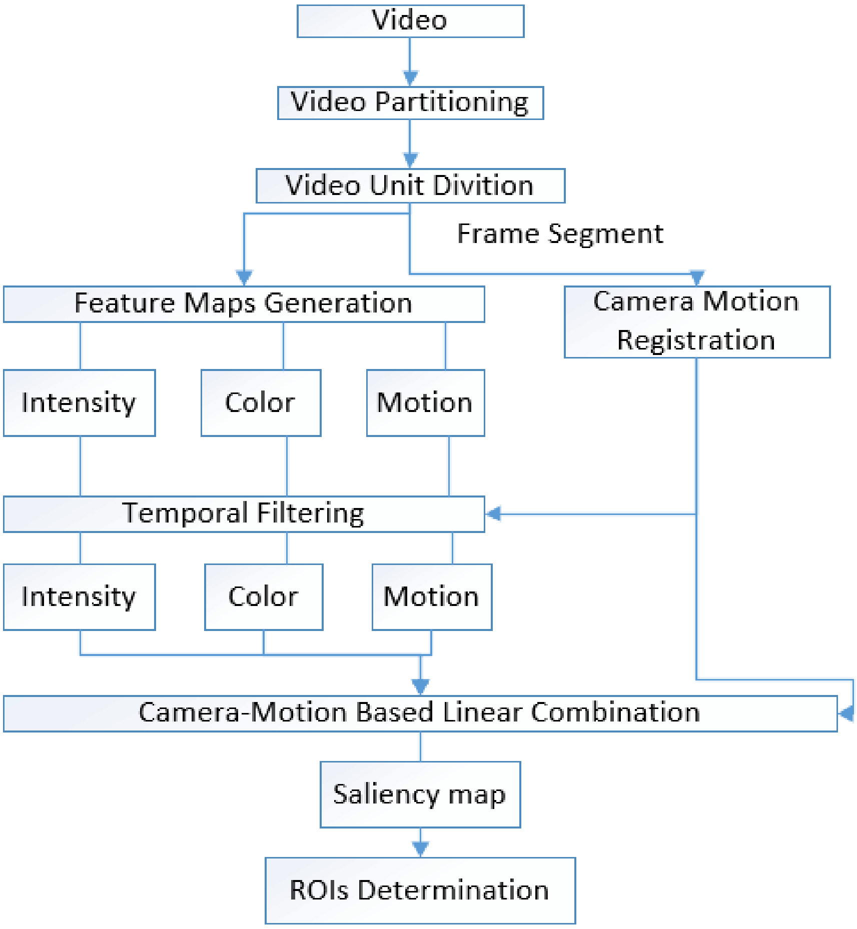


Figure 2-7 Illustration of Cheng motion saliency visual attention model.

Bioman et al. [32] proposed a method of detecting irregularities in the temporal space domain of the video. The basis of this method is to compare 2-D and 3-D texture training data sets of the video blocks instead use the motion information, to obtain the information of the irregular motion information in the video. Meur et al. [33] proposed a time-space domain model based on visual attention, by analyzing affine parameters to generate saliency map.

Muthuswamy et al. [29] proposed a motion saliency detection model. The discrete cosine transform (DCT) coefficient and motion information are used to identify the motion saliency in MPEG-2. In MPEG-2, I frame refers to intra information compression; B and P frame refers to motion compensation compression with the reference I frame. The DCT coefficient of every macro block is obtained according to the established partial decoder in the bitstream. Residual and motion vector are used to DCT coefficient de-quantization. The luma and chroma components of DCT coefficient are used to estimate spatial saliency. The motion saliency map is obtained by accumulating the motion vector map to improve the spatial saliency map.

Fang et al. [60] proposed video attention estimation model in MPEG-4. The features such as color information, luminance, and texture are extracted from DCT coefficients. The motion vectors are extracted from the bitstream. The intra prediction I frame saliency is estimated by color information, luminance, and texture which is relative with Itti's image saliency model. The P and B frame saliency is estimated by applying Gaussian model to weight the static saliency and motion map.

Khatoonabadi et al. [61] proposed a model to measure saliency using operational block description length (OBDL). This OBDL with minimum bits encoding in H.264/AVC compressed domain. When the prediction occur large errors, the residuals are large requiring more bits to code. OBDL can be calculated directly from decoder. According with OBDL, block size information from DCT coefficient can be extracted. Then Markov random field (MRF) is used to classify the block into saliency group or non-saliency group. The above methods can be summarized as following table.

	Compressed method	Features	Performance evaluation
Muthuswamy's model	MPEG-2	The DCT coefficient (luma and chroma components), motion vector	Average Equalized Error Rate (EER) value
Fang's model	MPEG-4	The DCT coefficient (color information, luminance and	Receiver Operating Characteristic (ROC) curve

		texture)	
Khatoonabadi’s model	H.264/AVC	operational block description length (OBDL)	Area Under Curve (AUC) of ROC

Table 2-1 The comparison of video saliency estimation in compressed domain.

As discussed with previous methods we can see that all techniques cannot cover the complex of HEVC features such as coding tree blocks splitting and bit locations. Moreover, none of the methods find the exact effect of features in the compressed domain to determinate the visual saliency. In fact, the relationship between the features in the compressed domain and the visual saliency can be achieved by training and testing the groundtruth and the video dataset. Therefore, this thesis presents a visual attention model of video by using HEVC features.

2.6.3. Applications of saliency model

The visual attention model has been applied in many fields. Baccon et al. [34] proposed the use of visual attention techniques to select spatial-related visual information to control the direction of movement of the robot. Driscoll et al. [35] constructed a pyramid-type artificial neural network by calculating the two-dimensional saliency of the surrounding environment to control the focus of the camera. Chen et al. [36] applied visual attention techniques to small picture displays. The salienct regions have higher priority than other regions. Ouerhani et al. [37] and Stentiford [38] applied the attention model to image compression; the high saliency regions have higher compression quality.

2.7. High Efficiency Video Coding (HEVC)

2.7.1. Video coding system

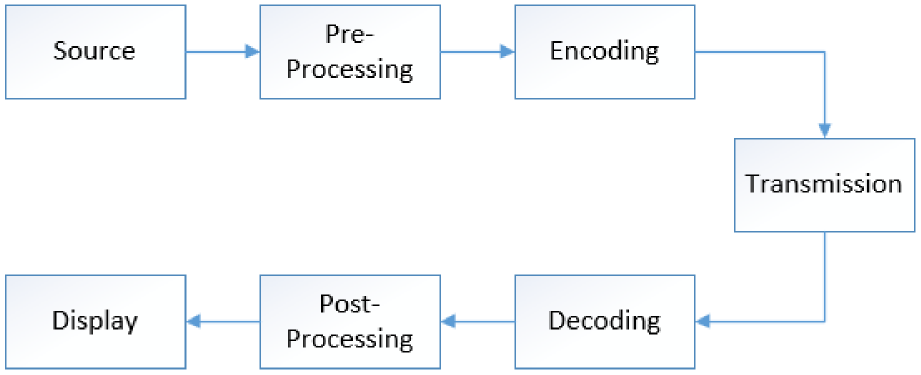


Figure 2-8 Video coding system.

Figure 2-8 provides an overview of HEVC video coding system. These functions of blocks are indicated as below:

- Video source: A video sequence is in a digital format acquired.
- Pre-Processing: Some pre-operations such as trimming [31], color correction or de-noising.
- Encoding: Operation of transform input sequence as a bitstream which can fit for the transmission scenario.
- Transmission: Transmit the coded bitstream to the receiver.
- Decoding: Processing bitstream into the reconstruction video. The reconstructed video is not the input sequence, as the encoding will occur compression loss.
- Display: The final video to view. Some color format of the video needs to change in this processing.

2.7.2. HEVC Introduction

The HEVC video coding compression standard is mainly developed by two major international organizations: ITU-T (International Telecommunication Union

Telecommunication Standardization Department) and ISO / IEC (International Organization for Standardization / International Electrotechnical Commission). ITU-T developed H.261 [32] and H.263 [32]; ISO / IEC developed MPEG-1 and MPEG4 [33]. These two organizations have developed H.262 / MPEG-2 video and H.264 / MPEG-4 AVC together. The co-developed video standards have been widely used, especially H.264 / MPEG-4 AVC [68]. Its applications include high-definition satellite television broadcasting, cable television, video capture / editing system, portable cameras, video surveillance, network and mobile Internet video transmission, Blu-ray discs, and real-time video applications such as video chat, video conferencing and telepresence systems. H.264 / MPEG-4 AVC basically covers all digital video applications and replaces other video compression standard.

However, with the increase in service diversification, the development of high-definition video and the emergence of ultra-high-definition format (4k × 2k or 8k × 4k), the market requires better video compression coding standards than H.264 / MPEG-4 AVC. In addition, with the rise of mobile devices and tablet PCs, demand for video services is increasing. Video quality and resolution requirements are increasing as well. These cause challenges to existing network bandwidth. Therefore, HEVC (High Efficiency Video Coding) as a new generation of video coding standards came into being. HEVC is composed of ITU-T VCEG (Video Coding Expert Group) [37] and ISO / IEC MPEG (Moving Picture Experts Group) [37]. JCT-VC began its first meeting in April 2010, to collect new video coding standard proposals from major companies, universities and research institutions in the world. The first version of HEVC was released in January 2013. The basic framework and content of HEVC were determined. HEVC continued to expand its content and functionality to adapt different application requirements such as multiple color space formats, Screen Content Coding (SCC), 3D video encoding, scalable video coding and so on. ISO / IEC will refer to HEVC as MPEG-H Part2 (ISO / IEC 23008-2), ITU-T may refer to HEVC as H.265.

The design goal of HEVC is to reduce the bit rate by 50% [69] compared to H.264 / AVC at the same image quality [34]. There are two main reasons for this design proposal: high-resolution video and parallel processing. Figure 2-9 shows the HEVC decoder structure, main including block segmentation, intra prediction, motion compensation, transform and quantization, deblocking, and sample adaptive offset (SAO).

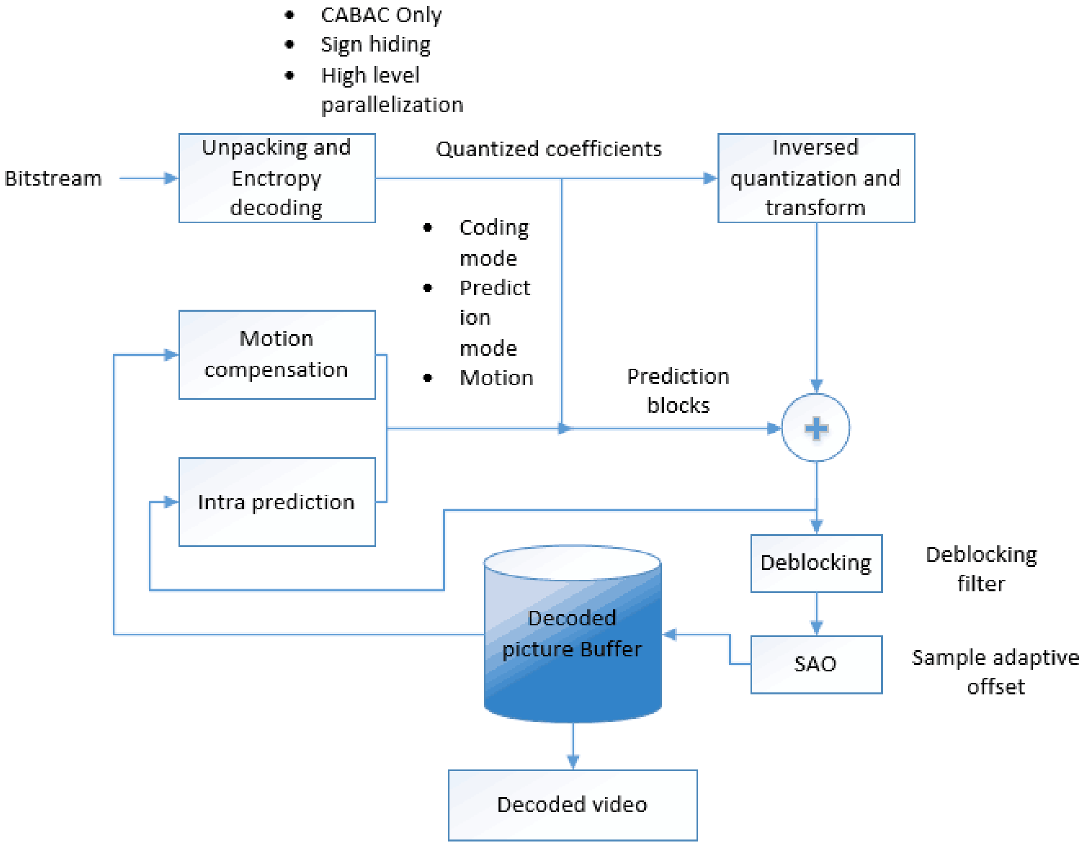


Figure 2-9 HEVC video decoder processing.

2.7.3. HEVC Coding structure

2.7.3.1. Group of Pictures

The video sequence is composed of several consecutive frames, and when it is compressed, the video sequence is divided into several GOPs (Group of Pictures). GOP is divided into: closed GOP and opened GOP.

Closed GOP is shown in the following figure 2-10, each GOP begins with a IDR (Instantaneous Decoding Refresh), and each GOP is independently coded.

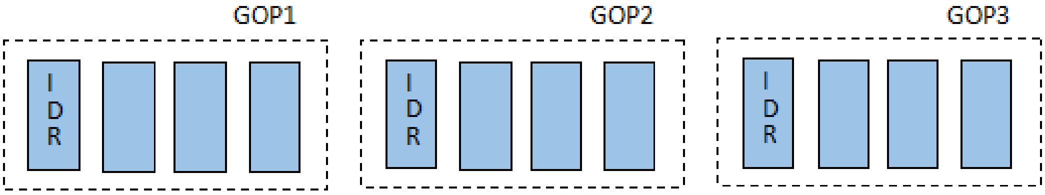


Figure 2-10 Closed GOP structure.

Open GOP is shown in figure 2-11, the first intra frame of the first GOP is an IDR frame, and the first intra frame in the subsequent GOP is a non-IDR frame [48]. That is, the inter frame in the subsequent GOP may cross the non-IDR frame and use the encoded frame in the previous GOP as the reference frame.

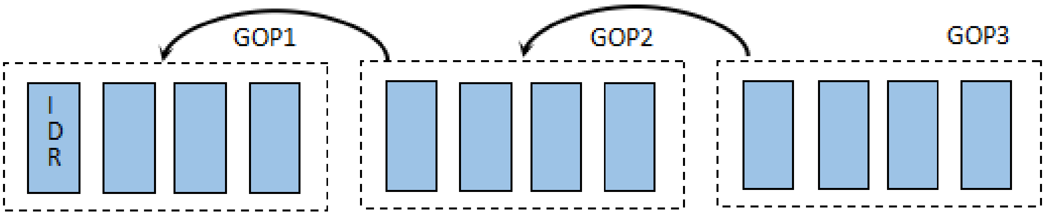


Figure 2-11 Opened GOP structure.

2.7.3.2. Slices

Each GOP is divided into multiple slices which are independently encoded. This main purpose is to resynchronize in the event of data loss. Each slice consists of one or more slice segment (SS) [35, 49]. In HEVC, a slice contains only one fragment by default. That is a frame is a slice, but also a slice segment. Slices can be coded by applying different coding types [49]:

1. I slice: All coding blocks in I slice are intra prediction coded.
2. P slice: Coding blocks in P slice can be coded in both intra prediction and inter prediction with only one motion compensation signal.
3. B slice: Coding blocks in B slice can be coded in both intra prediction and inter prediction with one or two motion compensation signals.

2.7.3.3. Tile

Tile is a new concept in HEVC. A frame can be divided into several slices, and it can be divided into several tiles, which divides an image from horizontal and vertical directions into several rectangular regions [36, 50]. The main purpose of tile division is to enhance the ability of parallel processing without introducing new error diffusion. Tile provides a greater degree of parallelism (at the image or sub-image level) than CTB, without the need for complex thread synchronization. The division of tile does not require uniform distribution of horizontal and vertical boundaries and can be grasped according to the requirements of parallel computing and error control. In general, the CTU data contained in each tile is approximately equal. At the encoding time, all tiles are processed in the order of scan. The number of CTUs in a tile and the number of CTUs in a slice does not affect each other. In the same image, some slices contain multiple tiles and some tiles contain multiple slices can simultaneously exist.

Slice and tile are divided for the purpose of independent coding. The shape of the tile is substantially rectangular, and the shape of the slice is striped. A slice consists of a series of slice segments (SSs), a SS consists of a series of CTUs. Tile is directly composed of a series of CTUs.

2.7.4. Prediction blocks

HEVC first divides a frame into a number of two-dimensional symmetric coding structure and then processing. CTU (Coding tree unit) [35] is the core of HEVC symmetric coding structure. CTU is similar with the "macroblock" in H.264 / AVC. The size of the CTU is not strictly limited. The size of CTU can be 64 × 64, 32 × 32, 16 × 16, and 8 × 8 [36]. The larger size can be better compression rate. CTU contains a luma CTB (Coding tree block) and two chroma samples CTBs. CTB cannot decide whether HEVC process intra or inter prediction. Thus CTBs can be partitioned into smaller blocks. CTU can be spilt into luma and chroma CBs (coding blocks). The largest CU is called the LCU (Largest Coding Unit), the smallest CU is called the SCU (Smallest Coding Unit). The size of the LCU and SCU is generally limited to an integer power of 2 and normally greater than or equal to 8.

If the size of the LCU and the maximum depth of the recursive segmentation is known, the sizes of CUs in the LCU are known. If the size of the LCU is 64 × 64 and depth is 4, the CUs size can be 64 × 64 (LCU) [36], 32 × 32, 16 × 16, 8 × 8. If the LCU size is 16 × 16 and depth is 2 [70], the CU size is 16 × 16, 8 × 8.

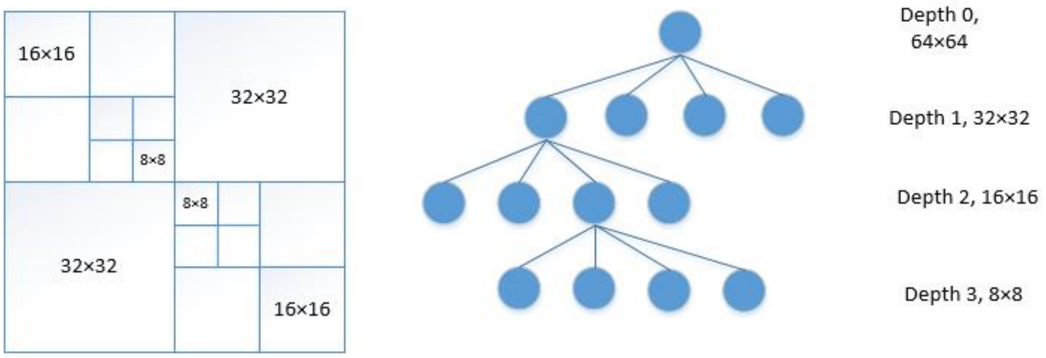


Figure 2-12 CU quad-tree syntax.

The un-limitation size of coding block is conducive to improve the efficiency of HEVC. The coding blocks (CBs) can be divided into luma and chroma prediction blocks (PBs). All the operations related to the prediction are processing in PUs. Intra mode difference, inter prediction, motion vector difference, reference frame index and motion compensation are based on PU processing. PU size is limited by the size of CU. After CU division, the size of PU is processing. There are three types of prediction mode in HEVC: Skip, Intra and Inter [50]. Predictive types are the main factors that affect PU segmentation. If the size of CU is 64 × 64, PU size is 64 × 64 in the skip mode. However, in intra mode, PU size may be 64 × 64 or 32 × 32. In Inter mode, PU size may be 64 × 64,64 × 32, 32 × 64, 32 × 32, 64 × 16, 64 × 48, 16 × 64 and 48 × 64 [37]. Residuals appear between the original blocks and its prediction blocks.

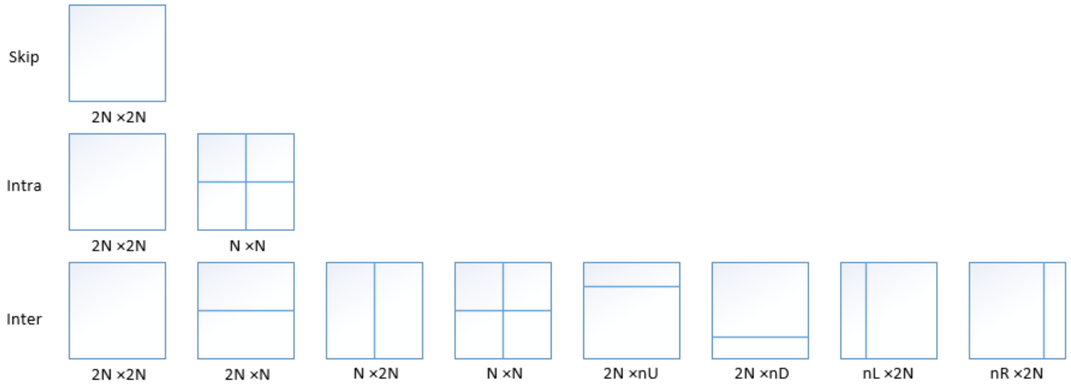


Figure 2-13 The block partitions in HEVC.

Residuals appear between the original blocks and its prediction blocks. The residuals of CBs can be split into smaller transform blocks (TBs). The size of transform blocks can be from 32*32 to 4*4.

HEVC also defines the TU (Transform Unit) as the basic unit of transformation and quantization. TU size may be greater than PU, but will not exceed the size of CU. TU must be two-dimensional symmetry. TU size is N × N or N / 2 × N / 2 [48] and depends on the PU split. The purpose of this segmentation design is to avoid TUs across the boundaries of PU. CU, PU, TU are independent and interrelated, this design is more in line with the texture features of images. Coding, prediction, transformation is more flexible [66] than H.264.

2.7.5. Intra prediction

The intra prediction of HEVC is similar to H.264 / AVC. It is based on data from neighboring blocks to perform predictive reconstruction in various ways. When encoding high-definition video, larger coding units will be applied. In order to make the intra prediction more accurate, the prediction modes of HEVC brightness component up to 35. Including two non-directional predictions: DC and Planar, and another 33 kinds of directional prediction [38]. As shown in Figure 2-14, there are five prediction modes for chrominance components: horizontal, vertical, DC, DM (Derivation Mode) and LM (Linear Mode) [67], where the DM mode determines the chrominance prediction mode according with the luminance prediction mode. The LM mode predicts the chromaticity of the current block based on the luminance and chrominance linear model relationships of neighboring blocks.

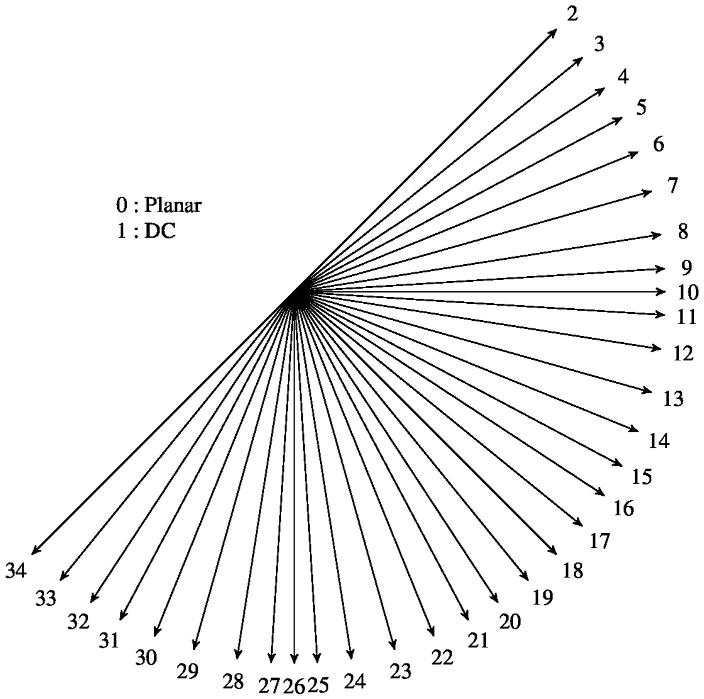


Figure 2-14 Intra prediction modes and directions of angular prediction. [58]

2.7.5.1. Planar prediction Mode

The Planar prediction is suitable for reconstruction of smoothing content. The JCT-VC first proposes this prediction scheme. First, the lower right corner pixel of the block is written to the coding, and then interpolates the rightmost column and the bottom row according to the adjacent block reconstruction pixel. The predictions of other pixels are obtained by bilinear interpolation. In 2011, JCT-VC proposes another planar prediction method [39]. The pixel at the bottom right corner is interpolated by the adjacent blocks instead transmission to the decoding part. In addition, the bilinear interpolation is changed to the average of horizontal and vertical linear interpolation.

2.7.5.2. Linear Mode (LM) prediction

LM (linear model) is new chroma prediction mode in HEVC. The specific calculation of chroma prediction is:

$$Predc[x, y] = \alpha . Recl'[x, y] + \beta \quad (2\text{-}16)$$

Where $Predc[x, y]$ is the chrominance prediction signal of the current block, $Recl'[x, y]$ is the luminance reconstruction signal of the current block. α and β are derived from the relationship between the luminance and chrominance signals of adjacent blocks.

If the video source is in YUV 4: 2: 0 format, the sampling rate of the chrominance signal is half that of the luminance signal. When using the LM prediction, the chrominance and luminance signals have a phase difference of 1/2 pixel. Therefore, it is necessary to first sample the luminance signals to match the size and phase of the chrominance signals. In the LM prediction mode, the reconstructed luminance signals down sampled in the vertical direction, and secondary sampled in the horizontal direction:

$$Recl'[x, y] = (Recl[2x, 2y] + Recl[2x, 2y + 1]) \gg 1 \quad (2\text{-}17)$$

By using least squares method, the relationship between the reconstructed luminance signal and the chrominance signal after the down sampling can be fitted to derive the parameters α and β.

$$\alpha = \frac{I \sum_{i=0}^{I} Recc(i)Recl'(i) - \sum_{i=0}^{I} Recc(i) \sum_{i=0}^{I} Recl'(i)}{I \sum_{i=0}^{I} Recl'(i)Recl'(i) - (\sum_{i=0}^{I} Recl'(i))^2} = \frac{A_1}{A_2} \quad (2\text{-}18)$$

$$\beta = \frac{\sum_{i=0}^{I} Recc(i) - \alpha \sum_{i=0}^{I} Recl'(i)}{I} \quad (2\text{-}19)$$

$Recc(i)\ and\ Recl'(i)$ represent reconstructed chrominance signals and reconstructed downsampled luminance signals. I is the total number of adjacent block sampling points.

Only the left and upper sides of the current block are sampling points. In the Intra configuration, LM mode enabled to increase the BD-rate of Y, C_b and C_r by 0.8 %, 7.8% [51] and 5.9% [52].

2.7.6. Inter prediction

As the PU segmentation may use four kinds of asymmetric way (2N × nU, 2N × nD, nL × 2N, nR × 2N), the motion vector is also allowed to asymmetric block as a unit in the inter prediction. This technique is called AMP (Asymmetric Motion Partition). The asymmetric shape of the region can be more flexible for motion estimation.

Conventional video encoders generally use predictive coding for motion vector coding. In addition, the difference between the MV prediction value and the actual value is encoded [40]. This spatial motion vector predictive coding method is also known as MVP (Motion Vector Prediction).

AMVP (Advanced motion vector prediction) is proposed instead of MVP in HEVC. The motion vector prediction candidate blocks are not limited to the spatial domain, also within the time domain. These candidate blocks to form a collection. The AMVP scheme will find the optimal MV matching in this collection, and then encoding the index of the optimal matching block, reference frame subscript, and MVD (Motion Vector Difference), so as to save the space cost more effectively. If MVD is 0, HEVC will enable the merge mode so that the current block and the candidate block share a motion vector. HEVC generally uses both AMVP and merge to achieve optimal MVP encoding efficiency.

2.7.7. Transform and quantization

2.7.7.1. Large scale Transform

H.264 / AVC only contains 4 × 4 and 8 × 8 transformation modes. HEVC added 16 × 16, 32 × 32 two large scale transformation [65]. For HD video, the use of large scale frequency domain transformations will achieve better coding result. Because blocks represent content that is typically part of a particular object or a small portion of the background in HD video, the contents of blocks are mostly uniform texture patterns and subtle color changes. The calculation [41] can be expressed as:

$$H = \begin{bmatrix} 64 & 64 & 64 & 64 & 64 & 64 & 64 & 64 & 64 & 64 & 64 & 64 & 64 & 64 & 64 & 64 \\ 90 & 87 & 80 & 70 & 57 & 43 & 25 & 9 & -9 & -25 & -43 & -57 & -70 & -80 & -87 & 90 \\ 89 & 75 & 50 & 18 & -18 & -50 & -75 & -89 & -89 & -75 & -50 & -18 & 18 & 50 & 75 & 89 \\ 87 & 57 & 9 & -43 & -80 & -90 & -70 & -25 & 25 & 70 & 90 & 80 & 43 & -9 & -57 & -87 \\ 83 & 36 & -36 & -83 & -83 & -36 & 36 & 83 & 83 & 36 & -36 & -83 & -83 & -36 & 36 & 83 \\ 80 & 9 & -70 & -87 & -25 & 57 & 90 & 43 & -43 & -90 & -57 & 25 & 87 & 70 & -9 & -80 \\ 75 & -18 & -89 & -50 & 50 & 89 & 18 & -75 & -75 & 18 & 89 & 50 & -5 & -89 & -18 & 75 \\ 70 & -43 & -87 & 9 & 90 & 25 & -80 & -57 & 57 & 80 & -25 & -90 & -9 & 87 & 43 & -70 \\ 64 & -64 & -64 & 64 & 64 & -64 & -64 & 64 & 64 & -64 & -64 & 64 & 64 & -64 & -64 & 64 \\ 57 & -80 & -25 & 90 & -9 & -87 & 43 & 70 & -70 & -43 & 87 & 9 & -90 & 25 & 80 & -57 \\ 50 & -89 & 18 & 75 & -75 & -18 & 89 & -50 & -50 & 89 & -18 & -75 & 75 & 18 & -89 & 50 \\ 43 & -90 & 57 & 25 & -87 & 70 & 9 & -80 & 80 & -9 & -70 & 87 & -25 & -57 & 90 & -43 \\ 36 & -83 & 83 & -36 & -36 & 83 & -83 & 36 & 36 & -83 & 83 & -36 & -36 & 83 & -83 & 36 \\ 25 & -70 & 90 & -80 & 43 & 9 & -57 & 87 & -87 & 57 & -9 & -43 & 80 & -90 & 70 & -25 \\ 18 & -50 & 75 & -89 & 89 & -75 & 50 & -18 & -18 & 50 & -75 & 89 & -89 & 75 & -50 & 18 \\ 9 & -25 & 43 & -57 & 70 & -80 & 87 & -90 & -90 & -87 & 80 & -70 & 57 & -43 & 25 & -9 \end{bmatrix}$$

2.7.7.2. Alternative DST

For 4 × 4 size TU, HEVC provides an optional DST based transformation mode. The transformation matrix is shown as below:

$$H = \begin{bmatrix} 29 & 55 & 74 & 84 \\ 74 & 74 & 0 & -74 \\ 84 & -29 & -74 & 55 \\ 55 & -84 & 74 & -29 \end{bmatrix}$$

DST has better coding adaptability for regions that the residual amplitudes increased. And DST can save about 1% of the bit rate [42]. In addition, DST conversion is only used to 4 × 4 luma transform blocks.

2.7.7.3. Transform Skipping

In order to improve the efficiency of video coding, HEVC also involves some other coding techniques. TSM (Transform Skip Mode) [43] is one of the technologies adopted by HEVC. Due to the anisotropic characteristics of video content, the traditional Hybrid video encoder cannot achieve the best coding results. It will be better to encoder prediction residual directly without frequency transformation.

Since the correlation between blocks of intra prediction is not as high as inter prediction, the prediction residual of intra prediction is generally large, especially for coding blocks.

The use of 2D frequency domain transforms facilitates energy concentration [42,43]. If the video source is a screen image, the content is mostly repetitive lossless matching data, the intra prediction residuals will be smaller or zero, in which case if the frequency domain transform is still used, it will be reduced coding efficiency. In this case, transform units will skip the transformation in TSM. In the subsequent CABAC entropy coding stage, the statistical properties of residual data can be modified to obtain better coding results.

The motion compensation residual signal generally exhibits different characteristics in both vertical and horizontal directions. Therefore HEVC can select different TSM [44] to skip the horizontal/vertical transform according to the specific situation in inter prediction. The TSM mode also includes the option of enabling both horizontal and vertical transformations. For some screen videos, BD-rate performance can be increased by up to 30% after enabling TSM with less modification to HEVC encoder.

2.7.7.4. Quantization

The processing of quantization is to obtain a simpler representation of the transform coefficients. Quantization is the main reason of distortion in compression, so choosing the appropriate quantization step size to balance distortion and bit rate becomes the key problem. The quantization step size in HEVC is marked by the quantization parameter (QP), with a total of 52 levels (0 to 51). Each QP corresponds to an actual quantization step size. The large value of QP means the quantization will result in the lower bit rate, and the greater distortion will be. HEVC uses the Rate Distortion Optimized Quantization (RDOQ) technique to select the optimal quantization parameter for a given bit rate to minimize the distortion of the reconstructed image.

2.7.8. In-loop filtering

2.7.8.1. Deblocking Filter

Due to the error caused by the quantization of the frequency domain transform and the prediction deviation caused by the motion compensation, the block based coding processing appear effect that PU and TU boundaries are not aligned after the deblocking Filter prediction/transformation/quantization step. Therefore, the hybrid video encoder will eliminate that effect. The general practice is adding deblocking filter in the block boundary [45]. HEVC to block filter basically follows the H.264 / AVC method [64], such as filtering method and boundary strength decision mechanism. The only difference is that HEVC adopted More flexible block segmentation scheme. In H.264 / AVC, deblock filter is used on 4*4 blocks. However, deblock filter is only used on 8*8 blocks in HEVC.

2.7.8.2. Sample adaptive offset (SAO)

Sampling Adaptive Offset (SAO) is a new technology in HEVC after the deblocking filtering [46]. The principle of SAO is based on the differences between the reconstructed frame and the original frame. SAO can improve the coding performance of 2% to 6%, coding complexity increased about 2% [48]. There are two kinds of offset method in SAO: Band Offset (BO) and Edge Offset (EO). Band offset is divided into several bands according to the pixel intensity. The same offset sample value is used for each band.

Edge offset is mainly used to compensate for the edge of the pixel in the frame, by comparing the center sample and the adjacent two samples to obtain the type of that pixel. There are four patterns of adjacent samples edge offset, as shown in Figure 2-15.

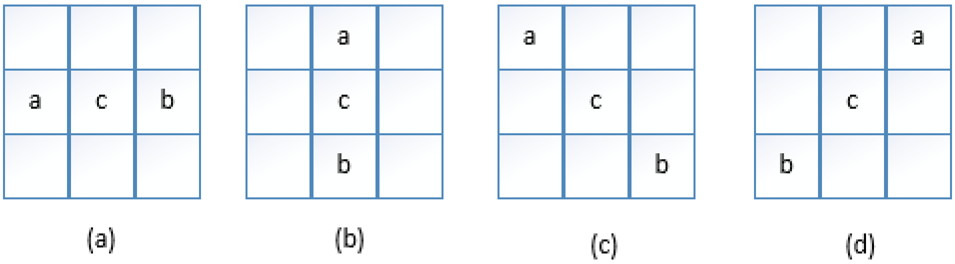


Figure 2-15 Four patterns of adjacent samples in SAO. C is the center sample and a and b is two adjacent samples. (a) horizontal location, (b) vertical location, (c) 135° diagonal location, (d) 45° diagonal location.

SAO initialization is performed in units of frames at the encoder. By analyzing the distortion between the reconstructed data and the original data, the parameters of SAO are configured and the SAO type is determined. SAO processing is performed for each LCU [47]. This prior information collection and finishing process only appears in the encoder part. At decoding, each LCU processing is independent. LCU does not need to access the frame buffer data can be decoded.

2.7.9. Decoded picture buffer

The decoded picture buffer (DPB) used to keep the decoded frames. There are three types of frames in the decoded picture buffer: short-term reference frames [58], long-term reference frames [58], and un-reference frames. Frames, which are used for prediction, are in the reference picture set (RPS) [58]. Un-reference frames will be released from DPB.

Chapter 3. Visual saliency estimation via HEVC bitstream analysis

3.1. Introduction

At present, most of the video saliency detection model is based on the pixel domain. The pixel-based video saliency detection model must decompress video to the spatiotemporal domain first and then performing feature extraction. The process of partial decoding bitstream is not only computationally complex but also time consuming. Most of the video on the network is stored in a compressed form. Compressed video is widely used in internet multimedia applications [54] because they can reduce storage space and significantly accelerate the transmission speed. Video processing based on compressed domain is gaining attention because of its no decoding or partial decoding requires. In recent years, video saliency detection in the compressed domain has also been a preliminary study, but its detection effect has yet to be improved. Therefore, the temporal and spatial domain video saliency detection algorithm based on the compressed domain is expected to appear.

In this chapter, a novel visual saliency model using Bayesian model in the compressed domain is presented. The relative intra and inter features such as block size, residuals, intra mode difference, motion vectors are extracted from HEVC decoder. Intra mode saliency (spatial saliency) is achieved by applying Bayesian modelling of relative intra features. Then inter saliency is achieved by adding the spatial saliency from the reference block with temporal saliency from motion vector. The aim of this model is to achieve saliency maps in HEVC compressed domain without fully decoded bitstream. The results of our model is high accuracy and quick. This method can be used in image/video searching and visual interface design.

3.2. Experimental Datasets

The video dataset is from Grundmann’s et al. [59] which contains 10 videos. Each video contains about 100 frames. The ground truth data is also included in this dataset. There are 4 types of videos in the dataset: moving a fast object with a static background, moving a fast object with a movable background, moving slowly with a static background and moving slowly with the movable background. Figure 17 shows the thumbnails of test sequences in this dataset. This dataset can be found at the following link:

http://www.cc.gatech.edu/cpl/projects/videosegmentation

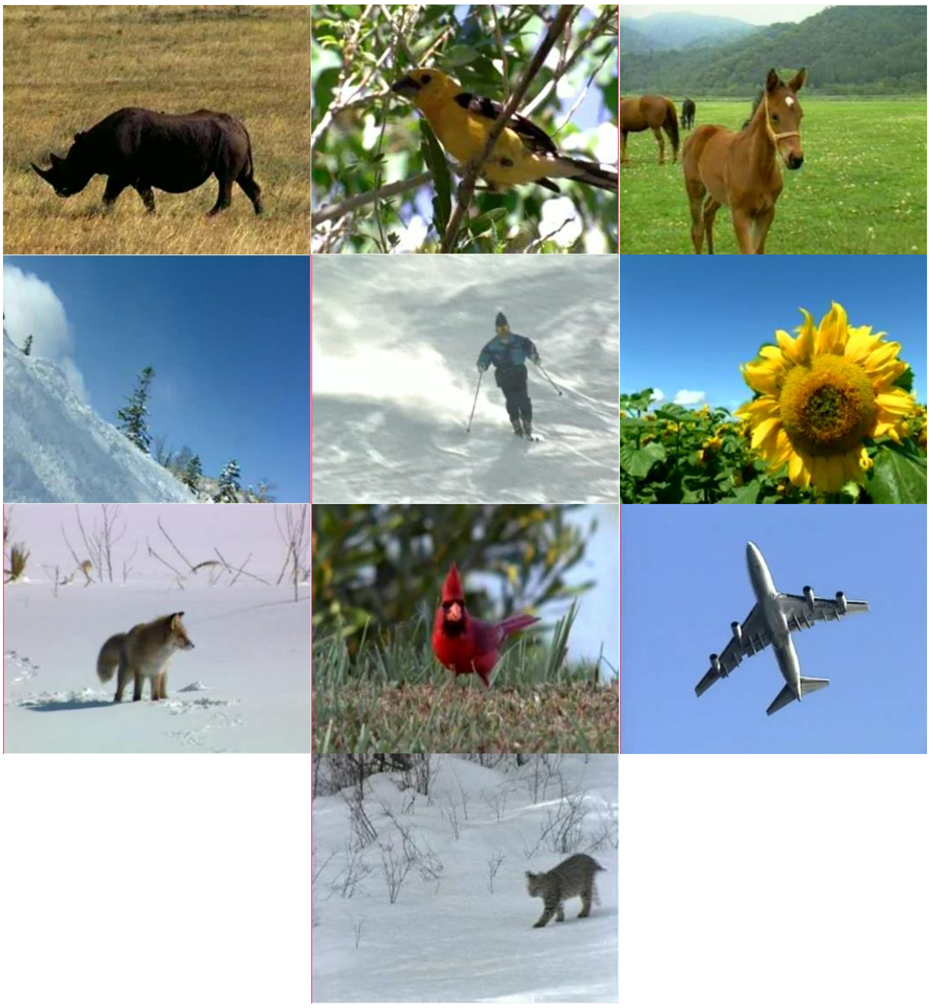


Figure 3-1 The thumbnails of test sequences in dataset.

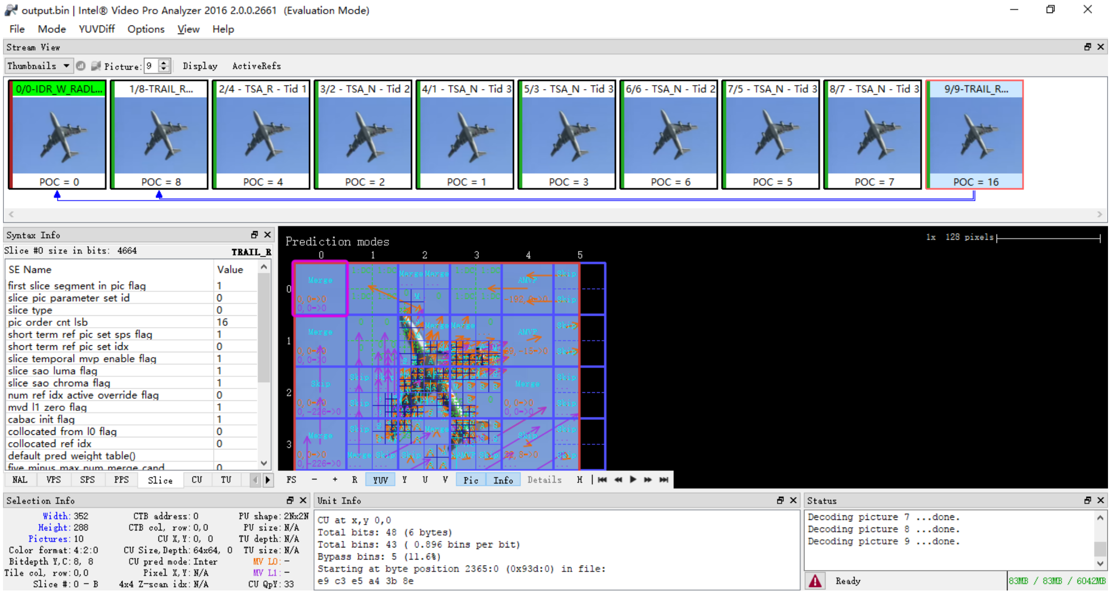
Another dataset is from Perazzi's et al. [75], which contains 28 videos. Each video contains about 80 frames. The ground truth data is also included in this dataset. This dataset can be found at the following link:

http://davischallenge.org/code.html

We combined videos from Grundmann's et al. [59] and Perazzi's et al. [75] together. Therefore, we separate the current dataset into four sets with (1) static background, (2) dynamic background, (3) objects moving with dynamic background, and (4) moving camera.

3.3. Developer tools

Intel Video Pro Analyzer 2016 evaluation mode is used in the beginning. Intel Pro Analyzer can be used to explore and debug both encoder and decoder processing. It supports HEVC, AVC and MPEG-2. The splitting maps, block information and motion compensation can be shown as intuitive view to help improve performance.



Syntax Info

Slice #0 size in bits: 4664 TRAIL_R

SE Name	Value
first slice segment in pic flag	1
slice pic parameter set id	0
slice type	0
pic order cnt lsb	16
short term ref pic set sps flag	1
short term ref pic set idx	0
slice temporal mvp enable flag	1
slice sao luma flag	1
slice sao chroma flag	1
num ref idx active override flag	0
mvd l1 zero flag	1
cabac init flag	1
collocated from l0 flag	0
collocated ref idx	0
default pred weight table()	
five minus max num merge cand	0

Selection Info

Width:	352	CTB address:	0	PU shape:	2Nx2N
Height:	288	CTB col, row:	0,0	PU size:	N/A
Pictures:	10	CU X,Y:	0, 0	TU depth:	N/A
Color format:	4:2:0	CU Size,Depth:	64x64, 0	TU size:	N/A
Bitdepth Y,C:	8, 8	CU pred mode:	Inter	MV L0:	-
Tile col, row:	0,0	Pixel X,Y:	N/A	MV L1:	-
Slice #:	0 - B	4x4 Z-scan idx:	N/A	CU QpY:	33

Unit Info

```
CU at x,y 0,0
Total bits: 48 (6 bytes)
Total bins: 43 ( 0.896 bins per bit)
Bypass bins: 5 (11.6%)
Starting at byte position 2365:0 (0x93d:0) in file:
e9 c3 e5 a4 3b 8e
```

Status

```
Decoding picture 7 ...done.
Decoding picture 8 ...done.
Decoding picture 9 ...done.
Ready                83MB / 83MB / 6042MB
```

Figure 3-2 Intel Video Pro Analyzer 2016 evaluation mode.

3.4. HEVC encoding profile and configuration

The HEVC is designed to be configurable and flexible to suit different type of video. The reference software HM 16 (HEVC Test Model) shipped with three different types of configure files. They are intra, random access, and low delay.

The following table lists the main parameters of a configuration file for HEVC and highlighted the major differences between these three configurations.

	Intra	Random access	Low delay
Profile	main	main	main
MaxCUWidth	64	64	64
MaxCUHeight	64	64	64
MaxPartitionDepth	4	4	4
QuadtreeTULog2MaxSize	5	5	5
QuadtreeTULog2MinSize	2	2	2
QuadtreeTUMaxDepthInter	3	3	3
QuadtreeTUMaxDepthIntra	3	3	3
IntraPeriod	1	32	-1 (only first)
DecodingRefreshType	none	Clean random access	none
GOPSize	1	8	4
Result (encode and decode a video which has 70 frames)			
Number of intra frames	70	8	1
Number of inter frame	0	62	69
Encoding time(seconds)	169	351	454
Decoding time(seconds)	617	175	170
File Size(KB)	128	24	23

Table 3-1 HEVC configuration.

All the result are computed using one computer in the same video sequence for 10 frames. As we can see, that all intra mode has the longest decoding time and the largest file size while random access and low delay configuration has the relatively smaller and similar decoding time and file size. We will discuss the pros and cons using a different configuration.

3.4.1. Intra

The intra configuration using intra period of 1, which means every frame is encoded as an intra frame the decoding order is the same as the frame order in a video by setting the decode refresh type to none.

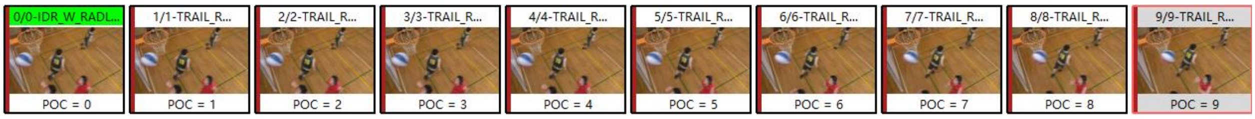


Figure 3-3 All intra-decoded sequence.

In the all intra mode, all frames are decoded in the intra mode. This means any frame in the sequence can be decoded without the information from the previous frame providing a robust video stream. The downside is also evident. Firstly, the decoding time is significantly longer than other two methods. Secondly, the encoded bitstream is also larger than other two methods. In this mode, only intra HEVC feature can be extracted, useful information such as motion vector is missing.

3.4.2. Low delay

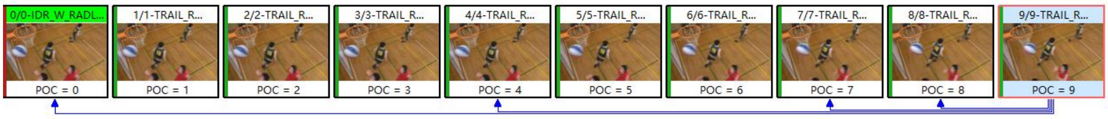


Figure 3-4 Low delay decoded sequence.

In low delay configuration, inter period is set to -1, which means only the first frame is encoded using intra mode, the rest are all encoded in inter mode. Similar to the all intra configuration, the frames are encoded in the same order as they are displayed. The decode time is significantly shorter and the file size is much smaller, comparing to the intra only mode. However, this reduction in size and decoding time comes at a cost. Each frame is relying on its reference frame and tracing all the way back to the first intra frame. For example, to decode the 100th frame in a video sequence, all the previous 99 frames need to be decoded. For an online streaming application, if one frame information is lost, the future frames cannot be decoded anymore.

3.4.3. Random access

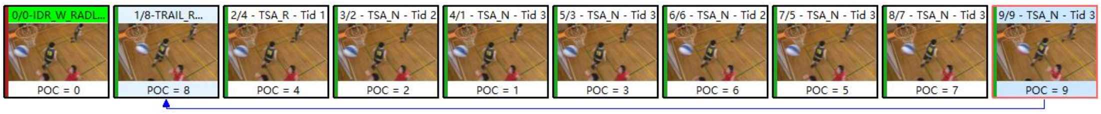


Figure 3-5 Random access decoded sequence.

The major difference between random access and the other two configurations is the order of decoding. While the other two decode the frames in the same order of the display of the frames, random access decodes the frames using a technic called clean random access. Clean random access (CRA) is a new feature introduced by HEVC that can decode a picture using an independently coded picture at the location of a random access point without decode any pictures in the bitstream before the access point. This feature brings the advantage of random access which allowing channel switching, video frame seeking, and online streaming, as all these applications need a way of fast decoding without a start from the first frame.

Researchers have been trying to using the features extracted from HEVC to estimate the saliency region of a given frame in a video sequence. Xu [71] et.al proposed a way of using the bit allocation, split depth and motion vector to estimate the saliency region, it should be noticed that in their approach, the bitstream needs to be encoded using low delay configuration. Guo [72] selected some high-level features such as decoded frame texture from the HEVC bit stream as well as some low-level features such as motion vector. However, features such as motion vector are not available in all frame. Methods rely on such features can only be used on limited HEVC configurations. In the following part, we will discuss the correlation between features and salient in details.

3.5. HEVC Feature analysis

From bit stream to reconstructed video, HEVC bitstream is decoded in steps.

At the beginning of decoding a frame, features are decoded from a bitstream. Those features include frame type, CU information, and PU information. The PU information describes how a frame is divided into prediction blocks and contains the prediction information about the size of a PU, the location of a PU, and the prediction method of the block. That information is further decoded into prediction and residual data to reconstruction the original frame.

The following chapter is focusing on discussing the performance of different of features in each steps estimating the salient.

3.5.1. Low-level features

Low-level features include information to reconstruct the prediction data and residual data, such as ADI, motion vector and residual coefficients. There are other common features that are shared between residual and prediction data, such as block depth.

3.5.1.1. Block depth

According with the previous HEVC structure, HEVC relies on dividing a video into multiple Coding Tree Units (CTUs). Block depth is a feature describing the number of division in a prediction block. According to configuration parameters, block depth ranges from 1 to 4, which means a prediction block can be recursively divided from 0 to 3 times. After partial decoding of the bitstream, block depth can be obtained. Based on the configuration of a HEVC encoder, the extracted block depth can be different for a video.

In intra main configuration, all frames are encoded in all intra mode. Each frame can be considered as an individual video and has no dependency between other frames within the video.

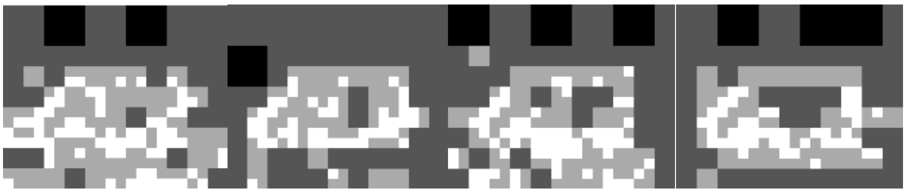


Figure 3-6 Block depth feature map in intra main configuration. First row: original frames in intra mode. Second row: block depth map in intra mode.

The feature map indicates that there is a correlation between salient and block depth. In the background, the frame is more likely to be divided into large blocks, hence has smaller block depth. In the saliency region, the frame is divided into smaller blocks and has larger block depth. However, one would notice that in the center of the saliency area, as the skin texture is repeated, the block depth becomes smaller than its surroundings. The shows that, the block depth can be correlated with the changes in color, orientation, and intensity. The Changes of color, orientation and intensity are used as features in Itti's method and have achieved a good result.

Figure 3-7 shows the number of saliency and non-saliency pixels in a video frame. It should be emphasized that the size of saliency region vary in different frames, and should be unified when analyzing them. For example, in figure 3-7, before unification, the number of saliency pixels are much smaller than the non-saliency pixels. When plotting without unification, the result can be biased. After unification, it is clear that in this video, a pixel is more likely to be a saliency pixel when it has a higher block depth.

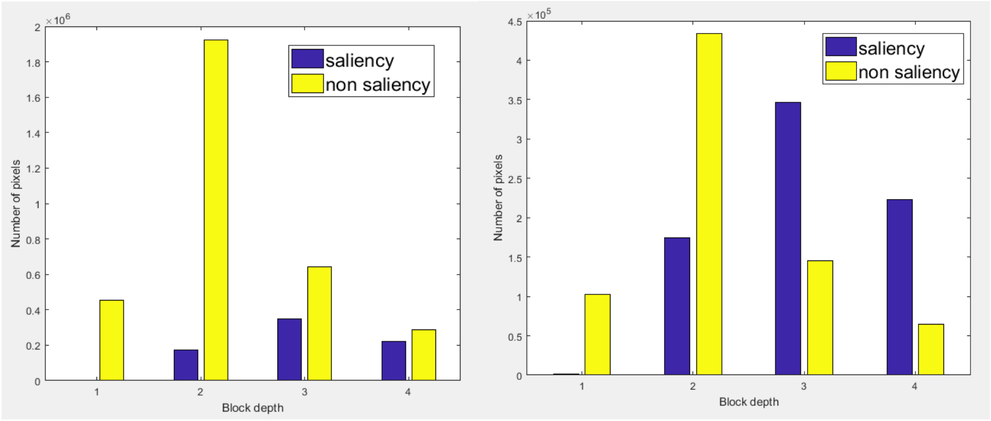


Figure 3-7 Number of saliency/non-saliency pixels in different block depth in intra main configuration. Left: Before unify. Right: After unify.

In low delay mode, only the first frame is an intra frame, the rest of all the frames are inter frame. All frames are encoded and decoded in the same order as they are in the video. No random access is enabled.

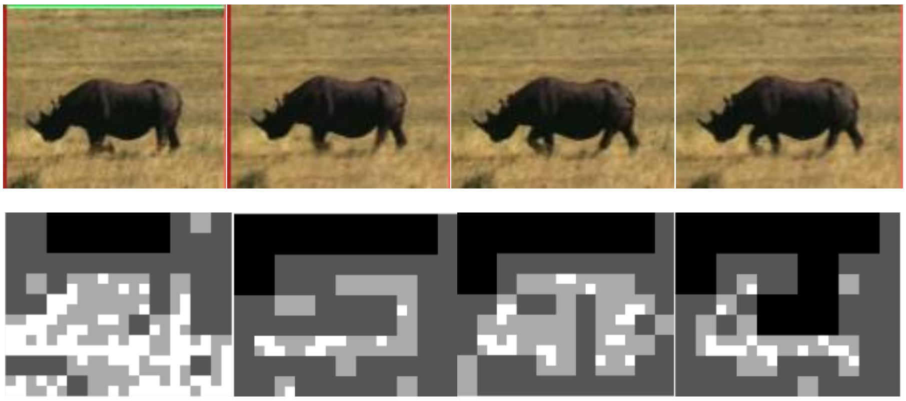


Figure 3-8 Block depth feature map in low delay configuration. First row: original frames in low delay mode. Second row: block depth map in low delay mode.

Figure 3-8 shows the block depth feature maps extracted from the same video as in figure 3-6 using a low delay configuration. It is clear that the block depth maps are different from the block depth maps from figure 3-6, although they are from the same frames in a video.

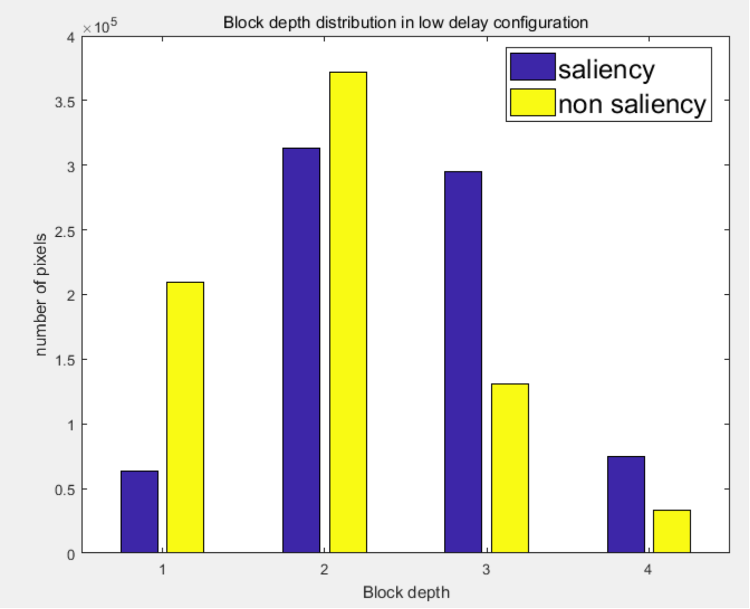


Figure 3-9 Number of saliency/non-saliency pixels in different block depth

Comparing to the all intra mode, the features extracted from a video encoded with random access configuration show less correlation with saliency pixels. It is noticeable that the number of saliency pixels in block depth 2 is close to the number of non-saliency pixels in block depth 2. This makes it difficult to estimate whether a pixel is a saliency or non-saliency by block depth alone.

The natural of low delay configuration will bring in another problem. Because all the frames in a video are temporally continuous and low delay configuration decodes them in a temporal order, frames will have little difference between themselves. For example, in the following sequence. POC 2 has no difference from POC 1. This will cause the encoder to set the block depth of all blocks to 1. This will also happen in the frames whose saliency object has not moved. Those outliers should be taken into consideration when trying to train a machine-learning model.

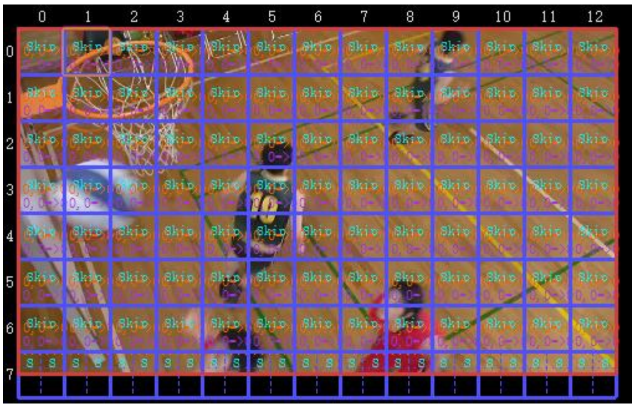


Figure 3-10 Example of all skip inter frame.

The random access configuration divides a video into smaller chunks of x frames. For every n frames, there is an intra frame and n-1 inter frames.

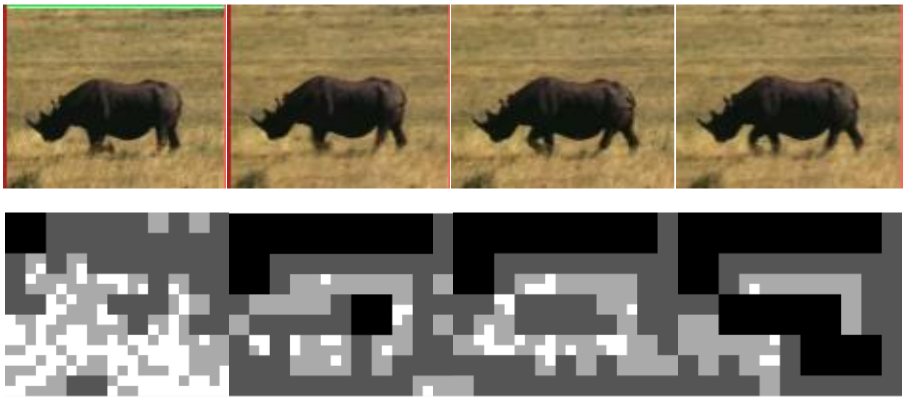


Figure 3-11 Block depth feature map in random access configuration. First row: original frames in

random access mode. Second row: block depth map in random access mode.

The result from random access is similar to the result from low delay configuration. If a frame is similar to its reference frame, the encoder will not divide the block into smaller blocks. The distribution of saliency and non-saliency pixels are shown in figure 3-12. It is similar to low delay configuration. Compared to intra configuration, the block depth can be considered as a weak feature in random access configuration.

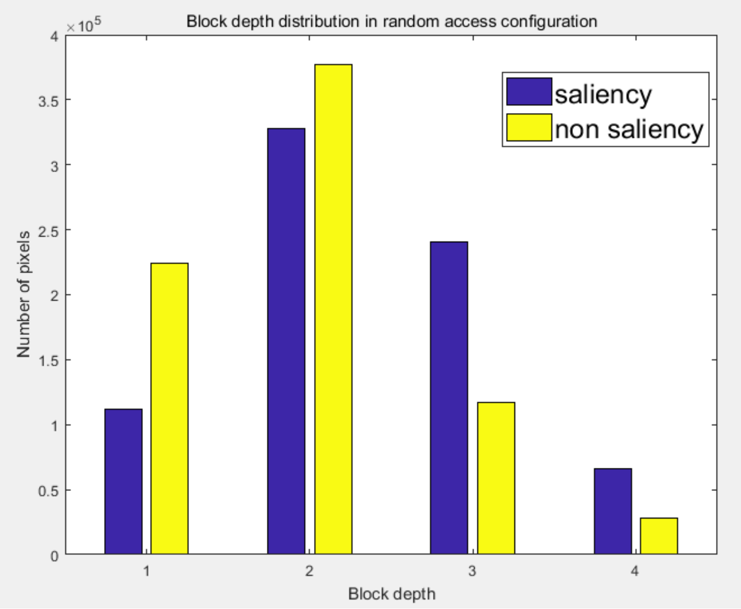


Figure 3-12 Number of saliency/non-saliency pixels in different block depth

By comparing three different configurations, the intra main configuration shows the best correlation between saliency and block depth. This is mainly because the video compression algorithms are based on the concept of using reference. All the frames are encoded in intra mode in intra main configuration while in other two configurations only a small amount of frames are encoded in all intra mode. In intra mode, HEVC will encode the frame as the beginning of the video sequence. In inter mode, HEVC will skip a block if it is similar to its reference, although this block may contain saliency pixels.

3.5.1.2. Arbitrary Directional Intra

HEVC support using Arbitrary Directional Intra (ADI) to predict movement. The coding efficiency is improved by using more prediction directions. The technology has been modified as following: there are amount 35 intra prediction modes available in HEVC, of which 33 of them are direction predictions. One is the traditional DC prediction 0 with a filter and one is the planar model 1. The mode 2-18 are horizontal prediction mode and 19-34 are denoted as vertical prediction mode. The ADIs are available in block size from 4×4 to 32×32.In an intra frame, Arbitrary Directional Intra (ADI) can be extracted from partial decoding the bitstream in order to reconstruct the prediction data. Unlike block depth, ADI is not a common feature shared by all frames. HEVC uses two different way to reconstruct the prediction data. ADI is the one used in intra blocks. In intra frame, all the blocks are encoded using ADI, while in inter frame only part of the blocks is encoded in ADI, the others are encoded in inter frame.

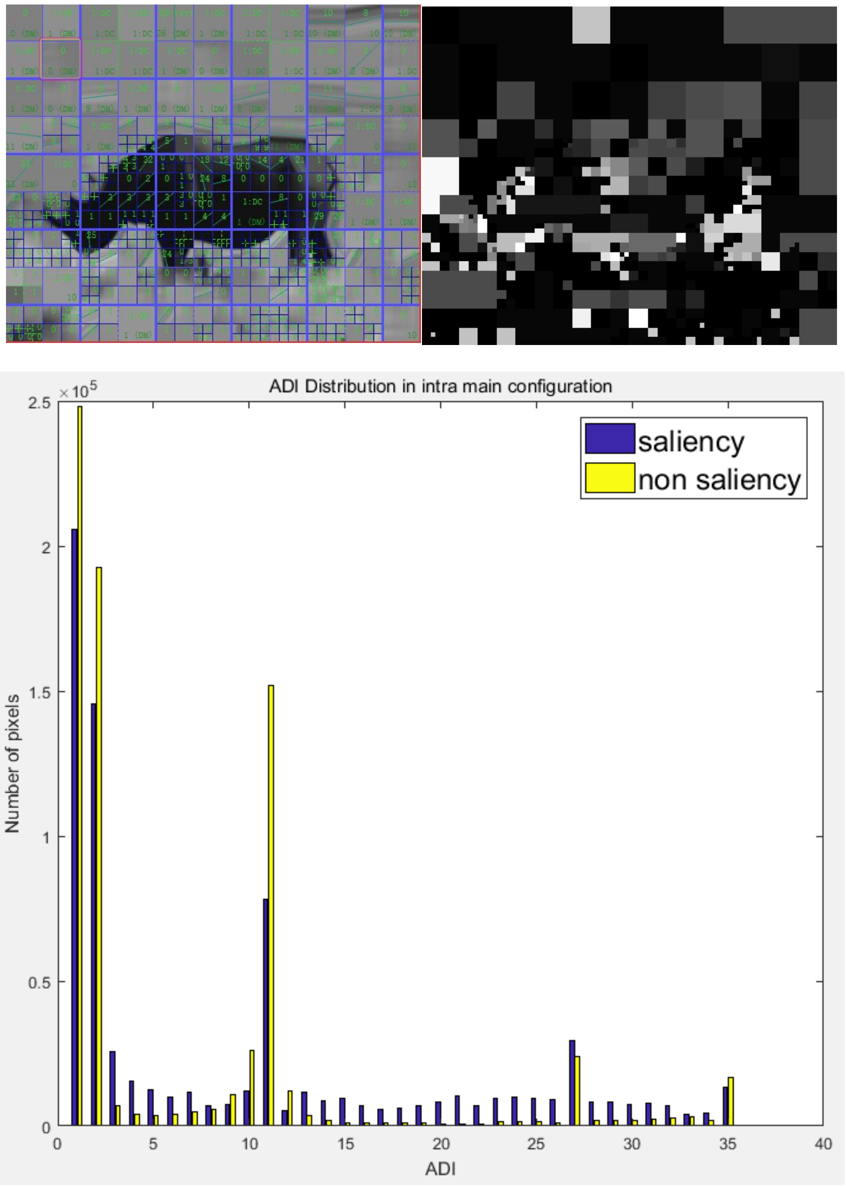


Figure 3-13 ADI feature analysis. Top left: an Original frame with ADI information. Top right: ADI feature map. Bottom: ADI value distribution in saliency and non-saliency pixels.

In figure 3-13, the top left image is the original frame from a video, the top right image is the feature map plotted by ADI value. There is no apparent correlation between ADI value and saliency map at a glance. The bottom chart is the histogram of ADI value from saliency pixels and non-saliency pixels from a 70 frames video encoded with all intra configuration. In the histogram shown in this video, most blocks are encoded with the value 0, 1 and 11. However, the number of saliencies and non-saliency pixels in ADI value 0 and 1 are very close to each other. This means, taking a random block that is encoded with ADI value 0, there is a 50% chance that this block could be saliency block and vice versa. Other ADI values show a better correlation between salient. For example, pixels with ADI value ranges from 12 to 26, 27 to 32 are more likely to be saliency pixels.

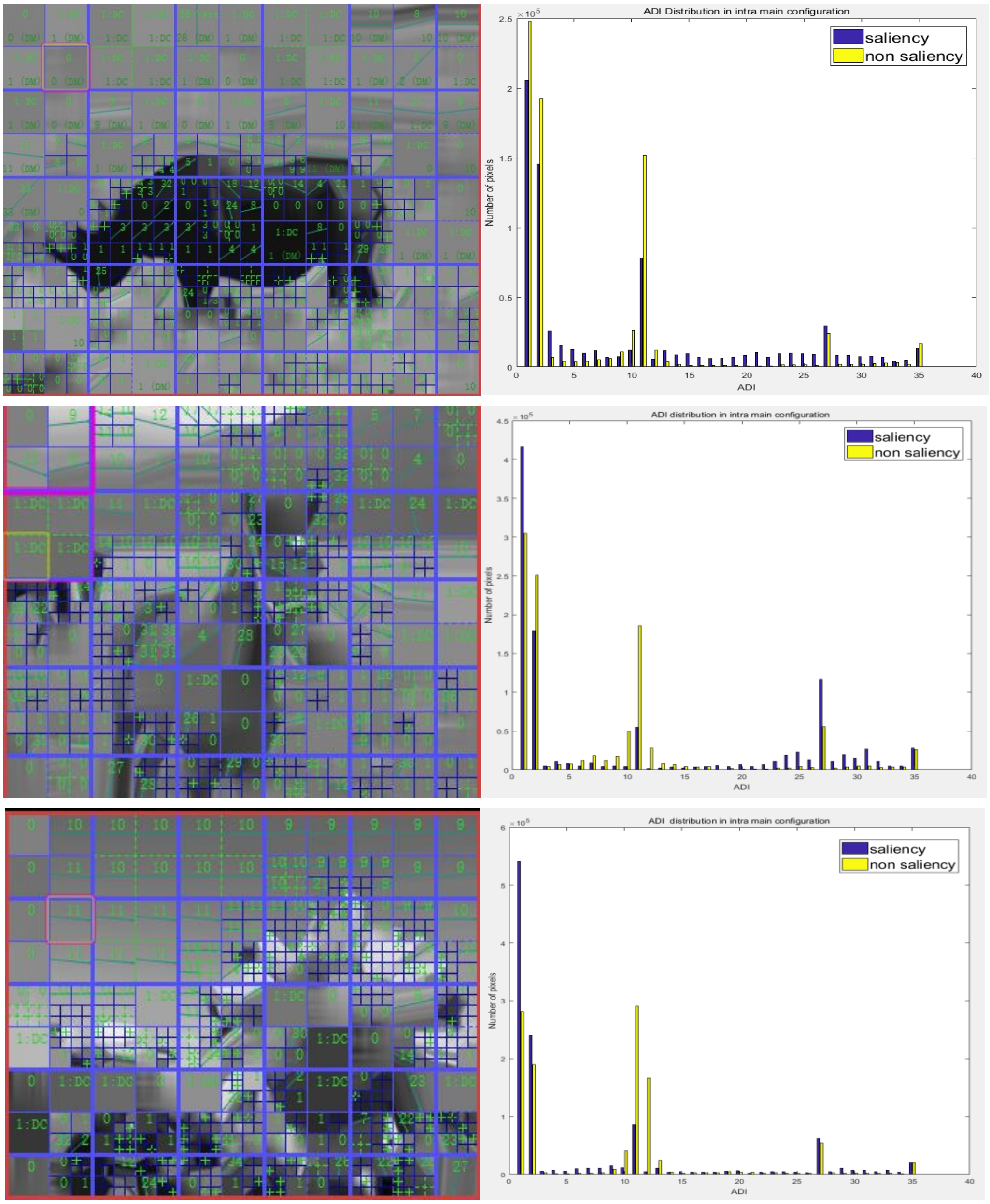


Figure 3-14 ADI and Saliency distribution in different videos.

It is also noticeable from figure 3-14, that the ADI value distribution of saliency and non-saliency pixels various in different videos. To train a machine-learning model using raw ADI data can be tricky. And a lot of various types of video should be used to extracted the training data to avoid overfitting. Future more, in inter frame, only a small number of blocks are encoded using ADI, the rest of them are encoded using motion compensation.

3.5.1.3. Motion vectors

The motion compensated prediction is performed in prediction blocks (PBs). The displacement between the reference frame and the current PB is the motion area between the reference frame and the current frame. The motion vectors are not the exact motion areas because of some rate distortion, but motion vectors are the good prediction of the true motion. In HEVC, there are two type of inter frames. P frames of motion compensation use single prediction reference, only one reference list is available. B frames of motion compensation use bidirectional one or two references, two reference lists available., The motion vector is another feature can be extracted from the bitstream. It describes the pixels movement from reference frame to current pixel.

In paper [71] Xu et.al. introduce a way of saliency detection utilizing the motion vector feature alone with other HEVC features. Under the assumption that motion is an obvious cue [72] of salient regions, saliency is calculated by motion vector values. However, it should be noticed there are drawbacks of using motion vector as a feature to estimate saliency. One obvious drawback is the motion vector only exists in inter frame. An intra frame saliency cannot be calculated using motion vector due to motion vector does not exist in intra frames.

The motion vector is also affected by the movement of the camera and saliency object. As shown in the images below both videos are shot by a standing camera. Compared to the moving person in video on the right, the sunflower in video on the left has less motion vector value. This indicates motion vector has less correlation with saliency in videos where the saliency object has a small amount of movement.

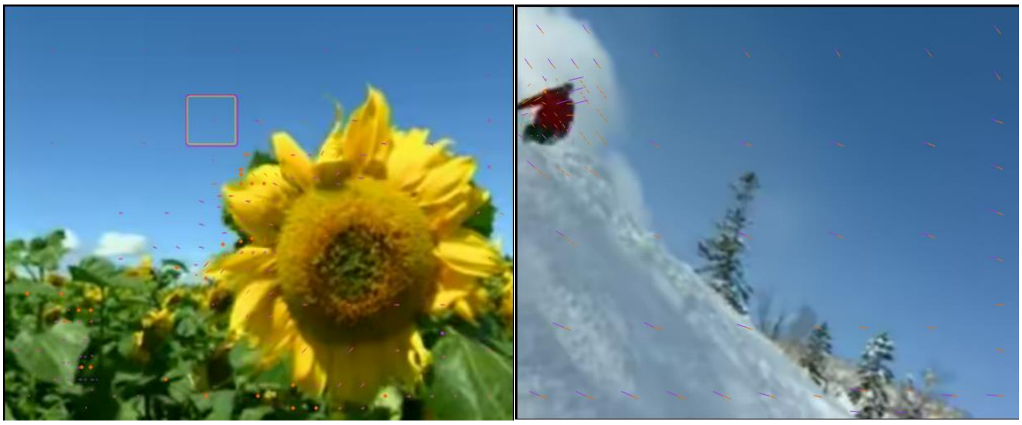


Figure 3-15 Motion vector movement comparison. Left: Image of saliency object with small amount of movement. Right: Image of saliency object with large amount of movement.

Furthermore, motion vectors are also affected by HEVC encoder configuration. The following two images are the same frame from one video encoded with two different configurations. The orange and purple lines are the motion vectors. As we can see, the basketball on the left image has much significate motion vector compares to the one on the right. The reason behind this is one of the configuration enabled the random access feature while the other one did not. Although the two images are the same frame of a video, they use a different frame as reference frame.

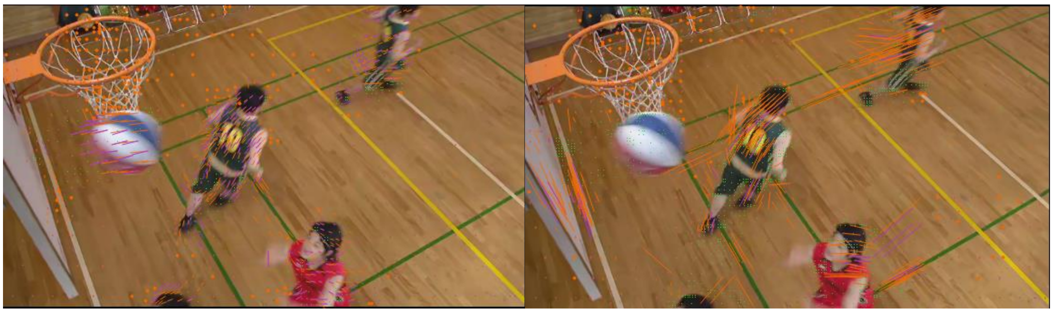


Figure 3-16 Inter frame motion vector in video encoded with different configurations.

In conclusion, two major factors affect the result of the motion vector. Firstly, the movement of the saliency object in a video. The motion vector is good at detecting a moving object in a video. However, the motion vector is less effective if the object is moving with the camera or the object only has a small amount of movement. Secondly, the motion vector can be affected by the configuration of the HEVC encoder. If a random access feature is enabled, motion vector will no longer present the movement of an object in temporally space.

As discussed above, all three low-level features have their limitations. This is mainly because of using different configurations in HEVC encoder. HEVC is a video compression algorithm based on the concept of using the references to reduce data redundancy. By using low-level features alone without considering its reference will lose the key information of the visual context.

To overcome those limitations, high-level features are extracted from the video sequences and a saliency detection model is designed by using the combination of low-level and high-level features.

3.5.2. High-level features

Before fully reconstruct the encoded video, some high-level features are calculated by the decoder. Those intermediate values are prediction data and residual data. Unlike the low-level features, the intermediate values are no longer reference-dependent. By using those data, we can reconstruct the frame without any data from other frames. This indicates that high-level features contain more visual context related information than the low-level features as discussed before.

3.5.2.1. Residual

Residual is the difference between the prediction blocks and the reconstruction of the coding blocks in the transform blocks (TBs) [55]. It describes how close is the predicted frame from the original frame. The residual is transformed and the quantized into residual coefficients, then entropy encoded into a bitstream

After inversed transformation at the decoder part, residual can be obtained. Similar to the low-level features discussed before, residual data is also affected by the HEVC encoder configuration [56]. The same frame in a video may have different residual data with different HEVC encoder configurations.

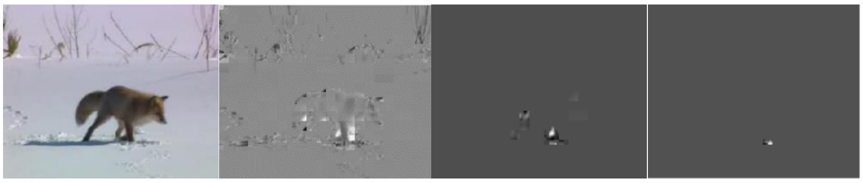


Figure 3-17 Residual of a frame in different video configuration. Left to right: Original frame; All intra configuration; Low delay configuration; Random access configuration.

As we can see from the figure 3-16, for a given frame, residual data changes dramatically in different configurations. Especially in random access configuration, the residual almost has no information, because the prediction data is very close to the original frame. Due to the fact, the residual data is not robust against configuration changing. It is not a desired feature to estimate residual with in inter frame. However, it can be used in intra-frame saliency prediction.

3.5.2.2. Prediction data

Prediction data is another high-level feature in HEVC. Unlike the residual data, prediction data contains the major part of a frame. The prediction data in HEVC is designed to be as close to the original data as possible. Figure3-18 below shows prediction data in the same frame from different configurations.

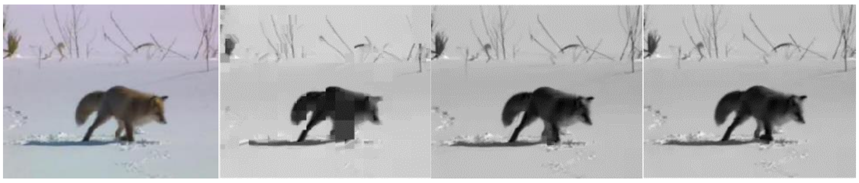


Figure 3-18 : Prediction of a frame in different video configuration. Left to right: Original frame; All intra configuration; Low delay configuration; Random access configuration.

By visual assessment, the prediction data is very close to each other and the original frame. It contains the most visual information compared to other features. Researchers such as Itti [11], have proven by using the features extracted from the image can successfully estimate saliency in an image. However ,using the original frame can be time-consuming , as the whole frame needs to be fully decoded. To overcome this problem, the prediction data is used to estimate saliency region. Comparing the original frame, prediction data contains sufficient data to estimate the saliency regions and require less computation to achieve. Compard to fully decoding, only decoding the prediction data require fewer steps. Time-consuming steps are skipped such as deblocking filters and Sampling Adaptive Offset (SAO).

In this section, low-level features such as block depth and high-level features such as residual data and prediction data have been discussed. Low-level features show a level of correlation with salient objects. However, the drawback of low-level features is quite obvious. When the target frame is very close to its reference frame, the video compression algorithm will skip the overlap part to achieve high compression rate. Without using the reference data, the low-level features can be useless. This situation is particularly common in videos when the cameras focused on the saliency object and saliency object moves slowly between each frame. In the other hand, the original data contains the most useful information to estimate the saliency region, but it is time-consuming to extract an original frame from a video.

3.6. Proposed method

As discussed in chapter 3.5, the following challenges need be overcomed when trying to design a model that can predict saliency with HEVC features.

- ◆ Features may be different in different configurations.

  The features extracted from one frame in one video varies when using different configurations. The HEVC configuration has a significant impact on the features ,especially the low-level features extracted from inter frames. The fact that the features extracted from one frame can be different in different configurations makes it difficult to make hand craft classifier for saliency prediction.

- ◆ Features do not exist in all the frames.

  Features like motion vector and ADI only exist in particular type of frames. The incompletion of features makes it difficult to design a universal model for saliency prediction.

To overcome those two challenges, two different saliency prediction models are designed.

3.6.1.Beysian theory

In general, the probability of event A under the condition of event B (occurrence) is not the same as the probability of event B under the condition of event A [74]. However, these two have deterministic relationship, the Bayesian theory is the statement of the relationship.

For a given independent variable set:

$$x = \{x_i\}_{i=1}^{|x|} \quad (3\text{-}1)$$

Where x_i is a an implementation of a random variable x.

For the Bayesian theory, the conditional probability $\ominus$ given by evidence x is:

$$prob(\ominus\,|\mathrm{x}) = \frac{prob(x|\ominus)*prob(\ominus)}{prob(x)} \quad (3\text{-}2)$$

(2) can be expressed as:

$$posterior = \frac{likelihood*prior}{evidence} \quad (3\text{-}3)$$

3.6.2. Naïve Bayesian classification

Naïve Bayesian classification is a general classification algorithm. This algorithm is based on the Bayesian theorem; it is collectively referred to as Bayesian classification. This classification model assigns class tags to eigenvalues that are given by the problem instance, and the class tags are taken from a finite set. For some types of probability models, a very good classification effect can be obtained by supervised learning. In many practical applications, the naive Bayesian model parameter is estimated using the maximum likelihood method. The assumption of variables are independent. It is only necessary to estimate each variable without the entire covariance matrix.

For a feature set x and a category set c:

$$x = \{a_1, a_2, \dots, a_m\} \quad (3\text{-}4)$$

$$c = \{y_1, y_2, \dots, y_n\} \quad (3\text{-}5)$$

If $P(y_k|x) = \max\{P(y_1|x), P(y_2|x), \dots, P(y_n|x)\}$,

We can know $x \in y_k(k\ \ is\ \ any\ \ category)$.

So the important step is calculation of $P(y_1|x), P(y_2|x), \dots, P(y_n|x)$.

So the probability of each feature under each category is obtained, which is:

$$P(a_1|y_1), P(a_2|y_1), P(a_3|y_1), P(a_1|y_2), P(a_2|y_2), P(a_3|y_2)$$

According with Bayesian theory [74],

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)} \quad (3\text{-}6)$$

Because the denominator is constant for all categories, we only need to maximized the $P(x|y_i)P(y_i)$ in equation (3-6). In addition, all features are independent of the conditions, so:

$$P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_i) \dots P(a_m|y_i)P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j|y_i) \quad (3\text{-}7)$$

Where m is total number of features.

3.6.3.Support vector machine

Support vector machine, referred as SVM was proposed by Cortez and Vapnik in 1995. [76] It shows many unique advantages in small samples classification and high dimensional pattern recognition. SVM also can be applied to other machine learning problems such as the algorithm of fitting.

In statistical theory, machine learning is an approximation of the real model. The vectors are mapped in higher dimensional space and are divided into different categories by hyper plane. The essence of the SVM algorithm is to find a hyper plane that maximizes a margin which is the minimum distance between the hyper plane and all the training samples. These hyper planes can be represented by functions:

$$\underset{w}{\rightarrow}.\underset{x}{\rightarrow} - b = 1. \quad (3\text{-}7)$$

$$\underset{w}{\rightarrow}.\underset{x}{\rightarrow} - b = -1. \quad (3\text{-}8)$$

When a new point x needs to be predicted which category to belong to, we can use $sgn(f(x))$. $sgn(f(x))$ represents the symbol function. When $f(x) > 0$,

$sgn(f(x)) = 1$. When $f(x) < 0$, $sgn(f(x)) = -1$.

The distance from point x to the hyper plane is:

$$\gamma = \frac{f(x)}{\|\underset{w}{\rightarrow}\|} = \frac{2}{\|\underset{w}{\rightarrow}\|}. \quad (3\text{-}9)$$

$\|\underset{w}{\rightarrow}\|$ is the Euclid norm of $\underset{w}{\rightarrow}$.

To maximize the distance γ, we need to minimize $\|\underset{w}{\rightarrow}\|$. This is a Lagrange optimization problem. The weight vector W and bias b can be obtained by the Lagrange multiplier method.

In order to make the problem easy to handle, our method is to integrate all the constraints into a new Lagrange function. The optimal point can be found by this new function. The Lagrange formula is:

$$\mathrm{L}(\omega, \beta) = f(\omega) + \sum_{i=1}^{l} \beta_i h_i(\omega). \quad (3\text{-}10)$$

ω is known as the Lagrange operator. Then we let:

$$\frac{\delta L}{\delta \omega_i} = 0; \ \frac{\delta l}{\delta \beta_i} = 0. \quad (3\text{-}11)$$

According with the previous function (3-10), the Lagrange formula can be written as:

$$L(\omega, \alpha, \beta) = f(\omega) + \sum_{i=1}^{k} \alpha_i g_i(\omega) + \sum_{i=1}^{l} \beta_i h_i(\omega). \quad (3\text{-}12)$$

The α_i *and* β_i are Lagrange operators.

So we let:

$$\theta_p(\omega) = ma\,x\, L(\omega, \alpha, \beta). \quad (3\text{-}13)$$

In this case, p presents as primal. The primal constraints are:

$$g_i(\omega) \leq 0; \quad i = 1, 2, \ldots, k. \quad (3\text{-}14)$$

$$h_i(\omega) = 0; \quad i = 1, 2, \ldots, l. \quad (3\text{-}15)$$

So we can draw the following function:

$$\theta_p(\omega) = \begin{cases} f(\omega) & if\ \omega\ satisfied\ primal\ constractions. \\ \infty & otherwise, \end{cases} \quad (3\text{-}16)$$

We can fine the boundary and solve the problem.

In our research, we have two categories: saliency (S) and non-saliency (NS). We can set saliency (S) as 1 and non-saliency (NS) as -1.

3.6.4.K nearest neighbors algorithm

The most basic of instance based learning methods is the k nearest neighbors (KNN) algorithm. This algorithm assumes that all instances correspond to points in n-dimensional Euclidean space a_n. The nearest neighbor of an instance is defined by the standard Euclidean distance. The arbitrary instance x is represented as the following eigenvectors:

$$a_1(x), a_2(x), a_3(x), \ldots, a_n(x) \quad (3\text{-}17)$$

Where $a_r(x)$ represents the "r"th property of the instance x. Then the distance between two instances x_i and x_j is defined as $\mathrm{d}(x_i, x_j)$.

$$\mathrm{d}(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(a_r(x_i) - a_r(x_j))^2}. \quad (3\text{-}18)$$

In the k nearest neighbor learning, the objective function can be discrete or real. KNN is an instance based learning algorithm. In KNN classification, the output is a taxonomic group. The classification of an object is determined by the majority of its neighbors. k equals a positive integer, usually small. If $k = 1$, the object's category is given directly by the nearest node.

In KNN regression, the output is the object's property value. This value is the average of the k nearest neighbors.

For a given instance x_q, selecting the closest instances of x_q in training examples and presenting as $x_1, x_2, \ldots, x_k$. Then regression:

$$\widehat{f(x_q)} \leftarrow \arg max \sum_{i=1}^{k} \delta(v, f(x_i)). \quad (3\text{-}19)$$

3.6.5.Decision tree

The decision tree is a tree structure (either binary or non-binary). Each of its leaf nodes represents a characteristic test. Moreover, each leaf node stores a category. The process of using decision tree is to test the corresponding attribute from the root node and select the output branch. The category stored in the leaf node is taken as the decision result.

Iterative Dichotomiser 3 (ID3) algorithm is to calculate the gain of each attribute, and then select the attribute with the highest gain to split. D is the classification of training samples. The entropy of D is expressed as:

$$info(D) = -\sum_{i=1}^{m} p_i log_2(p_i). \quad (3\text{-}20)$$

Where p_i denotes the probability that the "i"th category appears in the training set. Assuming training group D is divided following attribute A, the expected information of A on D is:

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} info\left(D_j\right). \quad (3\text{-}21)$$

The information gain is the difference between (3-20) and (3-21):

$$gain(A) = info(D) - info_A(D). \quad (3\text{-}22)$$

3.6.6. Intra-frame prediction model

An intra-frame prediction model is proposed to estimate the saliency region in intra frame mode.

As discussed in section 3.5, the intra frame contains low-level features block depth and ADI as well as two high-level features prediction data and residual data.

3.6.6.1. Features training

Based on the feature analyzed in chapter 3.5, the features having the highest correlation with saliency are chosen to train a naïve Bayesian model.

In our case, we use a training set including intra features: block depth:B, the prediction in YUV color space: $PrediY$, $PrediU$ and $PrediV$, and residual: R. All the training data are classified into two categories: saliency: S and non-saliency:NS.

From the naïve Bayesian theory, we can show:

The feature set x in our algorithm is:

$$x = \{B, PrediY,, PrediU, PrediV, R\} \quad (3\text{-}23)$$

The category set c in our algorithm is:

$$c = \{S, NS\} \quad (3\text{-}24)$$

Our dataset includes two parts: the training set and the testing set. 80% of videos are used as a training set and rest are in the testing set. The manually segmented ground truth is used to label each entry in the training set. From the training set and ground truth we can know the probability of each kind of features for saliency and the probability of saliency for each kind of features:

$P(S|B)$, $P(S|PrediY)$, $P(S|PrediU)$, $P(S|PrediV)$, $P(S|R)$, $P(B|S)$, $P(PrediY|S)$,$P(PrediU|S)$, $P(PrediV|S)$, $P(R|S)$

3.6.6.2. Features testing

For the testing set, we want to calculate the probability of features for saliency, therefore use the equation below

$$P(S|B,ADI,Predi,R) = \frac{P(B,ADI,Predi,R|S)P(S)}{P(B,ADI,Predi,R)} \quad (3\text{-}25)$$

Because each feature is individual, we have:

$$P(S|B,PrediY,PrediU,PrediV,R) = \frac{P(B,,PrediY,PrediU,PrediV,R|S)P(S)}{P(B,PrediY,PrediU,PrediV,R)} = \frac{P(B|S)P(PrediY|S)P(PrediU|S)P(PrediV|S)P(R|S)P(S)}{P(B)P(PrediY)P(PrediU)P(PrediV)P(R)} \quad (3\text{-}26)$$

For the SVM classification, the weight vector W and bias b can be obtained by the Lagrange multiplier method.

For the decision tree algorithm, the features can be classified by the gain determination. Furthermore, to reduce the complexity of reconstruct prediction data from ADI in intra frame. A simplified prediction recovery method is designed. Compared to the original algorithm, the proposed prediction recovery has two major changes.

1) The low pass filter for block size Nc > 4 is removed.

   The low pass filter is used to smooth the edge of each block. However, in saliency predication, the smoothness of the edge of the block does not have huge impact on the prediction result. Removing the low pass filter can speed up the prediction data reconstruction.

2) Remapping the 35 directions to 10 directions

$$ADI_{new} = \begin{cases} 0 & ADI = 0 \\ 1 & ADI = 1 \\ \frac{ADI}{4} & ADI > 1 \end{cases} \quad (3\text{-}27)$$

Figure 3-19 shows the comparison between original reconstructed prediction data and simplified reconstructed prediction data.

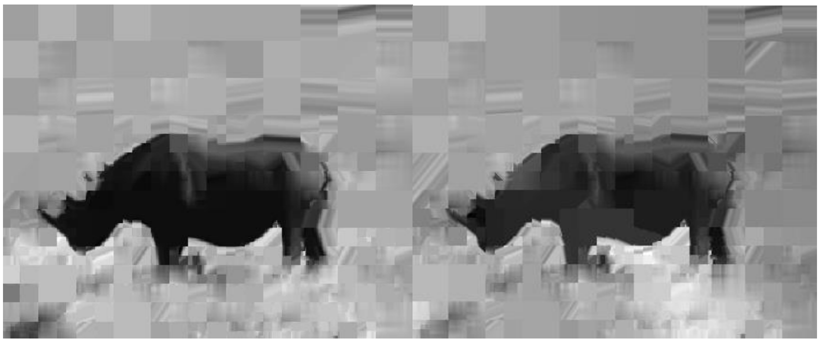


Figure 3-19 : Comparison between original reconstructed prediction data and simplified reconstructed prediction data. Left: original reconstructed prediction frame. Right: simplified reconstructed prediction frame.

As it is shown in the figure 3-19, the simplified reconstructed prediction frame has more noise than the original one. However, the shape of the saliency object is still visually recognizable.

3.6.7. Inter-frame saliency prediction model

The inter-frame prediction faces more challenge than the intra-frame saliency prediction. As discussed in section 3.5, the features of inter frames suffer the issue of inconsistency and incompletion. The blocks in an inter frame are normally a mix of both ADI encoded block and motion vector encoded block. Due to this reason, those features can not be used in a machine-learning model. Because most of the state of art classifier request the input matrix to be in the same size. Another major issue is the HEVC encoder might skip the saliency object if the object in the frame is in the same location as itself in the reference frame.

One option is to use the prediction data as a feature to training a machine-learning model as we did in the intra frame. However, it is time-consuming to decode the prediction data from the bitstream as the prediction data is at a late stage of the decoding process. To overcome this problem, we propose a reference based saliency reconstruction method.

When a HEVC video is decoding, the saliency map of the first intra frame is calculated by the intra-frame saliency prediction model in chapter 3.6. Then this saliency map is feedback to the decoder to be used as an input for the inter-frame saliency prediction model. The inter-frame saliency prediction model calculates the saliency map of an inter frame by using its ADI, motion vector and the saliency map of its reference frame.

$$Saliency = \begin{cases} ADI\left(Saliency_{ref}, adi\right) & block = intra\ \ block \\ MV\left(Saliency_{ref}, mv\right) & block = inter\ \ block \end{cases} \quad (3\text{-}28)$$

Where ADI and MV are the standard HEVC operation mentioned in book [58].

The two proposed model over comes the challenge we meet when trying to work with different configurations. The inter-frame saliency prediction model overcomes the problem caused by encoder skipping the saliency region. As a tradeoff , the proposed model needs to decode the prediction data from a frame which is time-consuming. However , the reference based inter-frame saliency prediction model avoids decoding prediction data from every frame, only the prediction data from intra frame is decoded. The performance of the proposed model are discussed in the next chapter.

3. 7. Experimental Result Analysis

The proposed method contains two different model, intra frame model, and inter frame model. The intra frame model contains a machine-learning model trained by a combination of low-level HEVC features and high-level HEVC features. The inter frame model uses HEVC low feature and the result from intra frame model to estimate the inter frame saliency. The results will be analyzed in two sections.

3.7.1. Intra frame model

Although the proposed method is designed for video saliency estimation, the result of the proposed model is frame based. It is worth to compare the result with both video saliency estimation models and image saliency estimation models.

The figure below compares the proposed method with Itti's [12] image saliency estimation model , gbvs [73] image estimation model and a manually segmented ground truth.

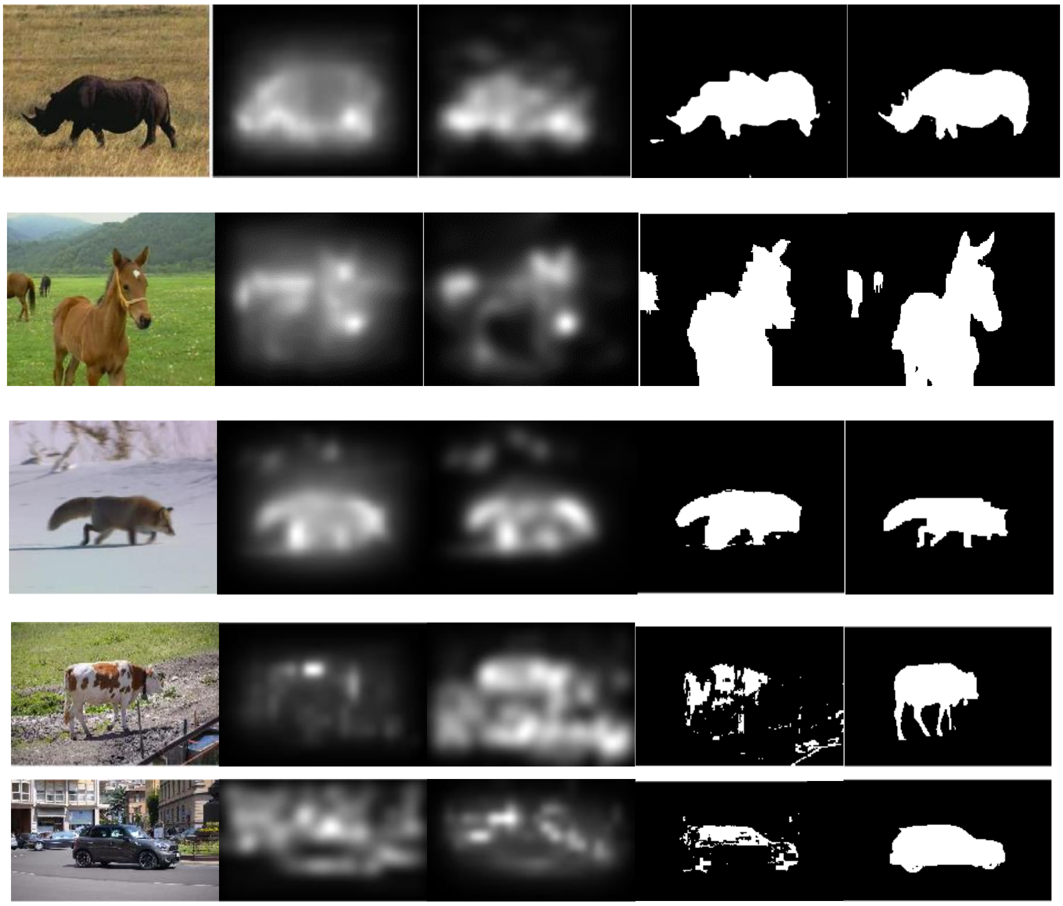
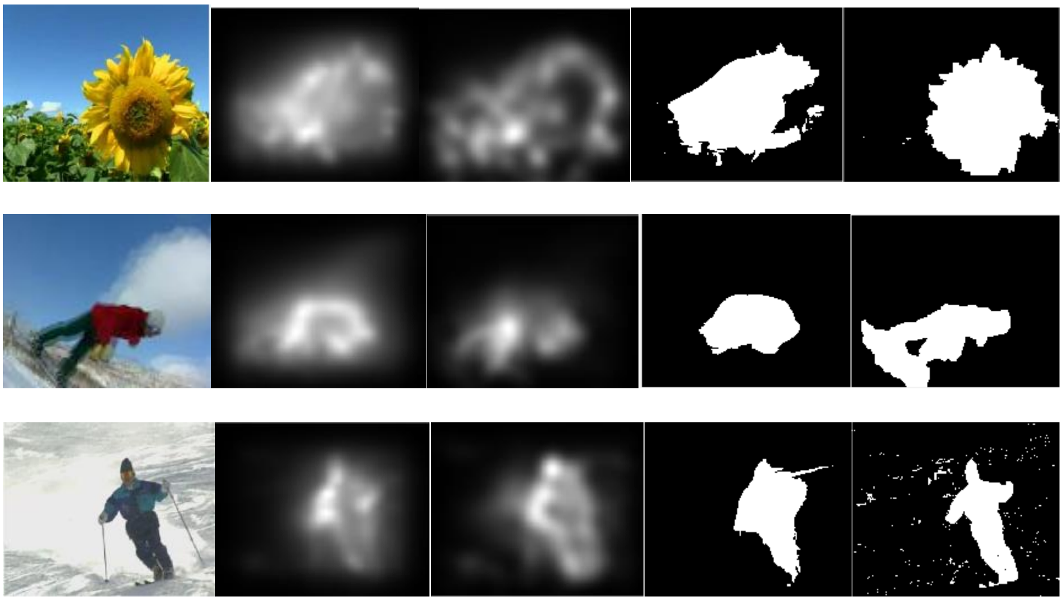


Figure 3-20 Intra-frame saliency results of static background set. First column : Original frame. Second column: GBVS image saliency. Third column: Itti's image saliency. Fourth column: Proposed method (Bayesian model). Fifth column: Manual segmented ground truth.



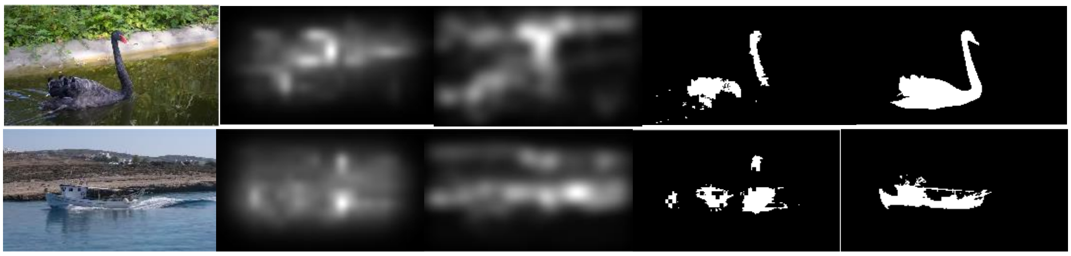


Figure 3-21 Intra-frame saliency results of dynamic background set. First column : Original frame. Second column: GBVS image saliency. Third column: Itti's image saliency. Fourth column: Proposed method (Bayesian model). Fifth column: Manual segmented ground truth.

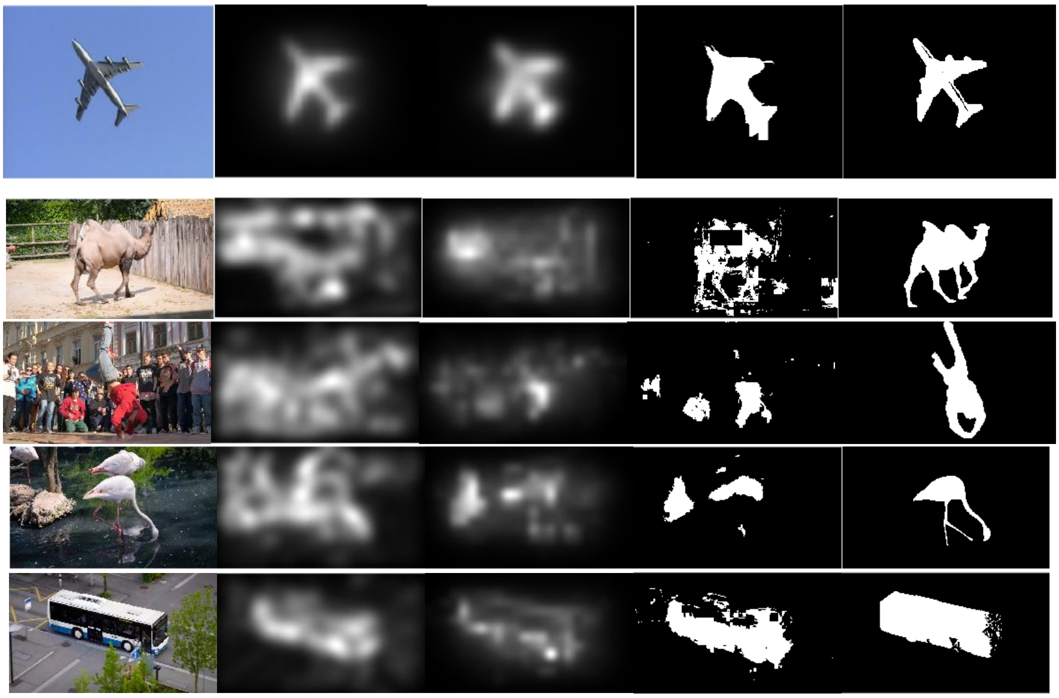
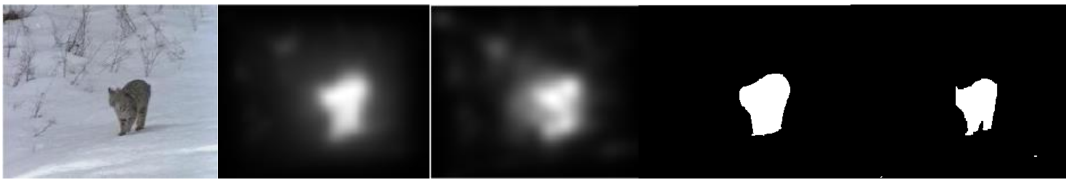


Figure 3-22Intra-frame saliency results of objects moving with dynamic background set. First column : Original frame. Second column: GBVS image saliency. Third column: Itti's image saliency. Fourth column: Proposed method (Bayesian model). Fifth column: Manual segmented ground truth.



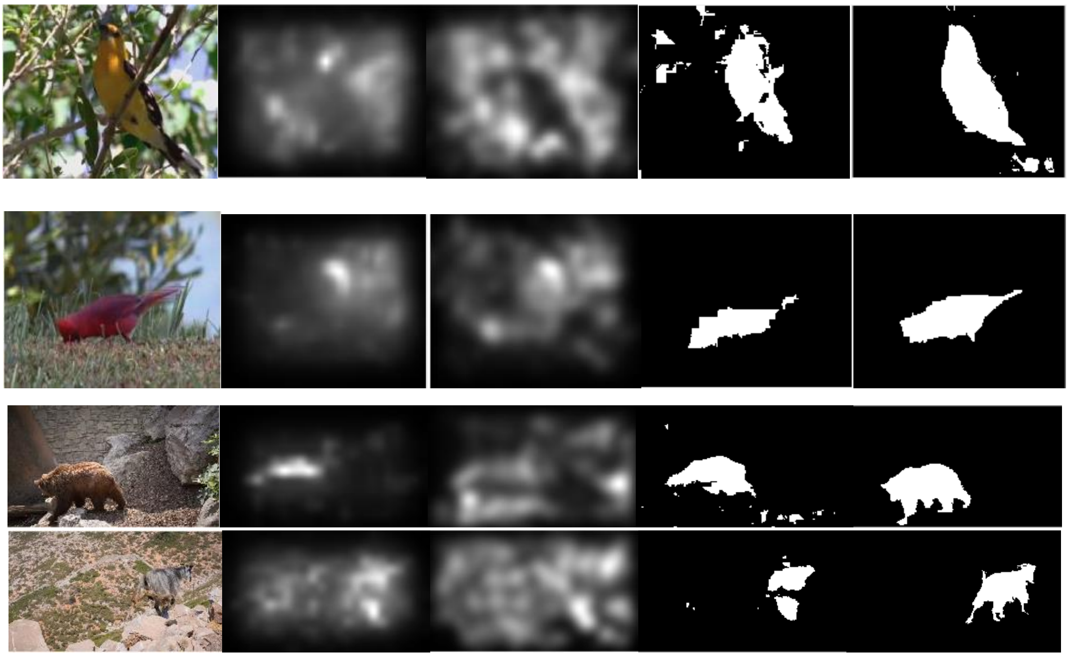


Figure 3-23 Intra-frame saliency results of moving camera. First column: Original frame. Second column: GBVS image saliency. Third column: Itti's image saliency. Fourth column: Proposed method (Bayesian model). Fifth column: Manual segmented ground truth.

We separate the dataset into four groups: (1) static background, (2) dynamic background, (3) objects moving with dynamic background, and (4) moving camera. From figure 3-20, 3-21, 3-22, 3-23 we can see, saliency cannot improved by setting groups. Some videos in the static background group still have unclear results. Saliency maps are not good when videos contain complex background.

With visual assessment, the proposed method shows satisfied performance comparing with the human segmented ground truth. When comparing with the gbvs model and the Itti's model, the proposed method is a binary classifier and will give a much sharper edge.

To analyze the performance of the proposed model, ROC accuracy (ACC) and computational time cost are shown in the table below.

Where ACC= (True Positive +True Negative) / total number of samples

	GBVS	Itti	Proposed method (Bayesian)	Proposed method (SVM)	Proposed method (KNN)	Proposed method (Decision tree)

ROC ACC	0.9417	0.9372	0.9561	0.918	0.962	0.909
Time cost(in seconds)	0.375	0.12	0.015	0.869	0.017	0.056

Table 3-2 Intra frame performance comparison between existing model and proposed model.

The proposed methods achieved the best accuracy between Itti’s model and GBVS. Compared with four different classification methods, we can see the Bayesian model and KNN have better accuracy. however, this difference is not significant. Unlike the proposed method, both GBVS method and Itti’s method are not binary classifiers. To achieve such accuracy, a manually fine tuned threshold need to be applied. Another advantage of the proposed method is it uses much less time to compute the saliency map comparing to other methods. This is mainly because the proposed method uses pre-trained machine-learning models instead of computing each saliency channels on the fly.

A ROC chart is plotted below to evaluate different saliency models. As shown in the chart, although the proposed has slightly less true positive rate than Itti’s model, it has a less false positive rate. In the mean while, the gbvs model has the worst true positive rate, but the lowest false positive rate. Since we set k=1 in the KNN model, the ROC is only a point. Since the results of SVM and decision tree are in lower accuracy than Bayesian and KNN. Furthermore, fine FNN only can plot one point ROC. We decided to choose Bayesian model as final classification method.

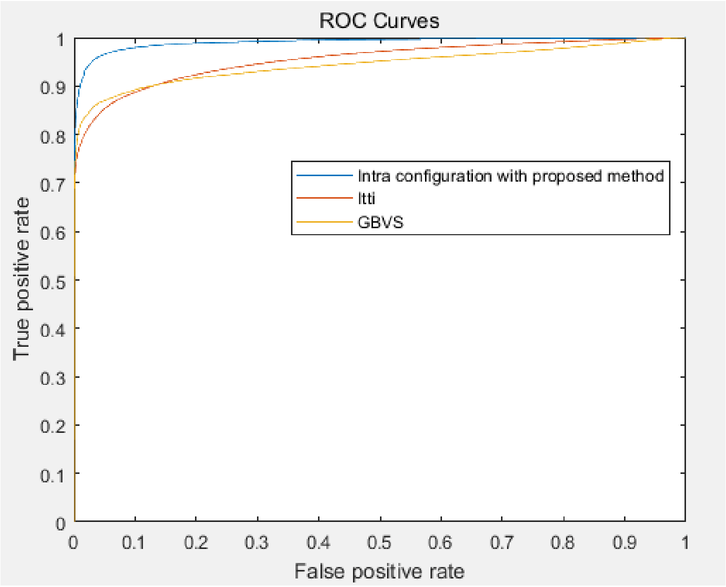
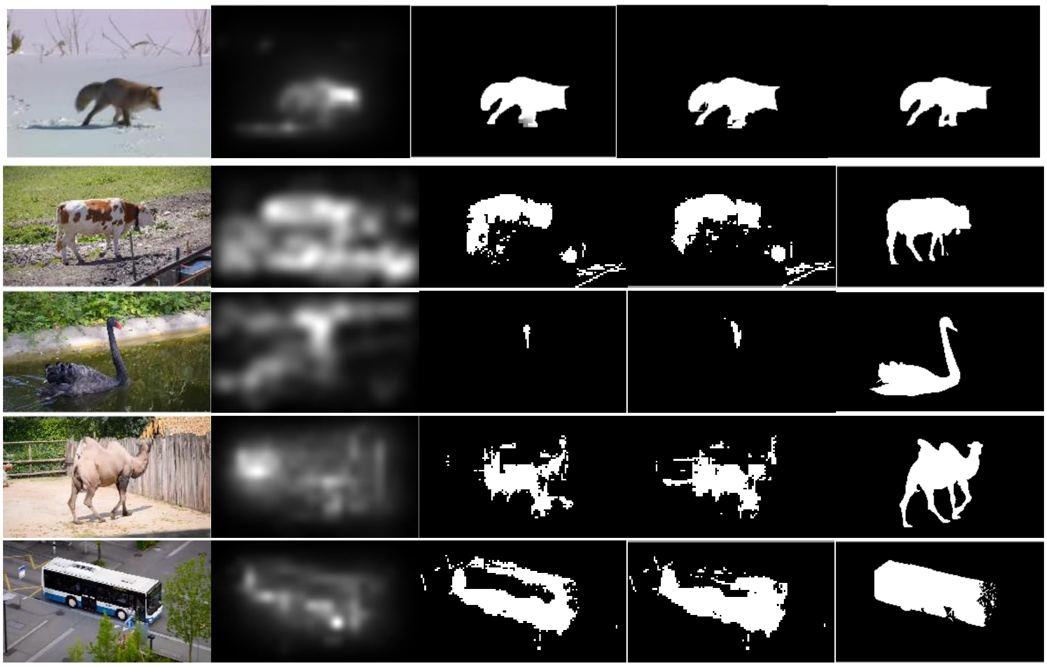


Figure 3-24 Intra frame ROC comparison between existing model and proposed model.

3.7.2. Inter frame model

Unlike the intra frame model, the proposed inter frame model uses the result from an intra model and combines the HEVC level feature to estimate the saliency region. As discussed before, for the same video, the different configuration will produce different low-level features. It is important the model is evaluated under all the configurations.



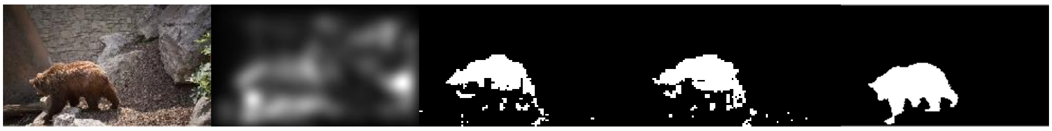


Figure 3-25 Inter-frame saliency results. First column: Original frame. Second column: Itti motion saliency. Third column: Proposed method with low delay configuration. Fourth column: Proposed method with random access configuration. Fifth column: Manually segmented ground truth from dataset.

The proposed method shows satisfactory result compared to the ground truth. However, the video encoded with the random access configuration show less accurate result. This is because the random access configuration enables the random access feature in the encoding step. Each frame and their reference frame are not necessarily temporally continued, thus the movement of saliency object is no longer continued as well. This may cause the encoder unable to find the nearest similar block and calculated the motion vector. The encoder will encode that block using ADI instead of the motion vector. The nature of ADI will cause the estimated saliency map has saliency pixels filling the entire block instead of showing the shape of the saliency object. From the ROC figure below, the proposed method works better with low delay configuration than the random access. However, random access configuration has its own advantage. Compares to low delay configuration, the video encoded with random access configuration will have one intra frame in every n frames based on the setting. Because the inter frame saliency is computed using reference saliency with prediction features. Having an intra frame in every n frames will calibrate the inter saliency accuracy rate.

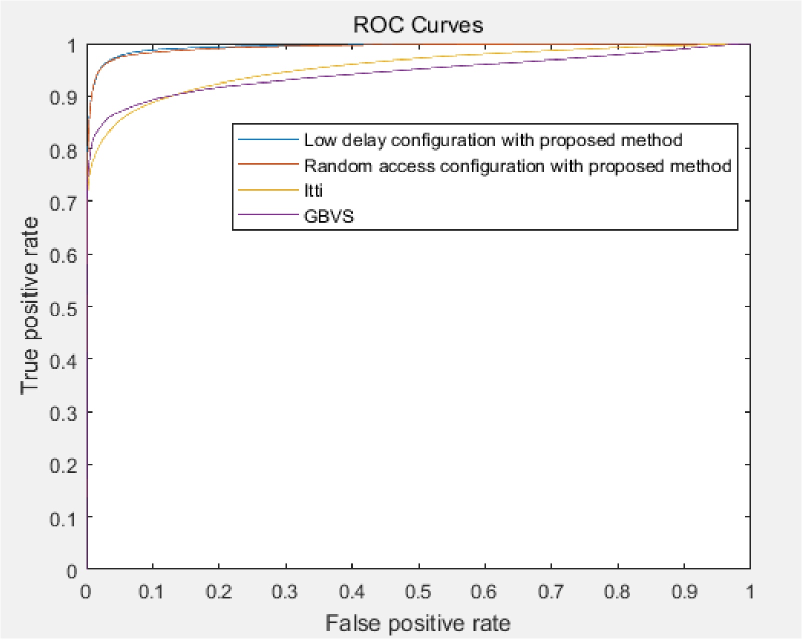


Figure 3-26 ROC curve of one video sequence.

3.8. Compression ratio

We set different compression ratio in the configuration of HEVC. The compression ratio is controlled by quantization parameter (QP) in HEVC. We set QP as three different values: 16, 32 and 51. The compression ratio becomes higher when QP is higher.

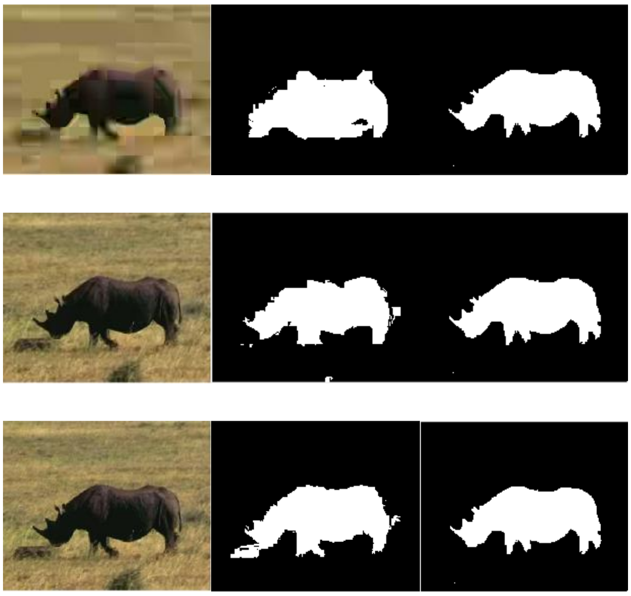


Figure 3-27 Intra-frame saliency results. First column: Original frame. Second column: Proposed method (Bayesian model). Third column: Manual segmented ground truth. First row: qp=16. Second row: qp=32. Third row: qp=51.

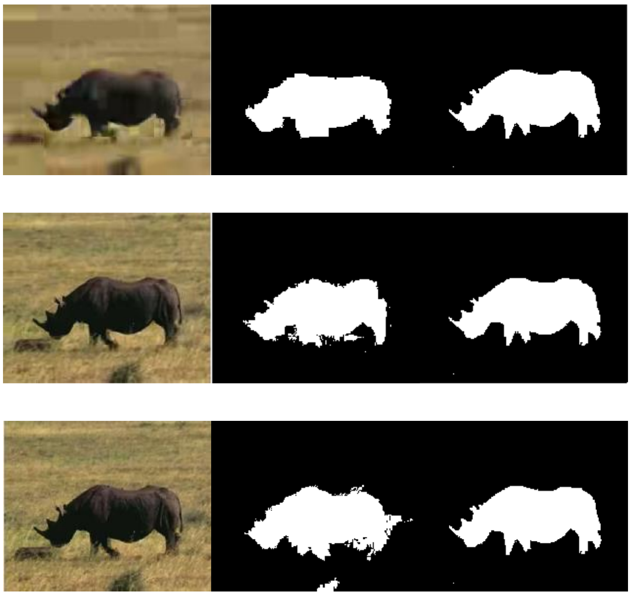


Figure 3-28 Inter-frame saliency results. First column: Original frame. Second column: Proposed method (Bayesian model). Third column: Manual segmented ground truth. First row: qp=16. Second row: qp=32. Third row: qp=51.

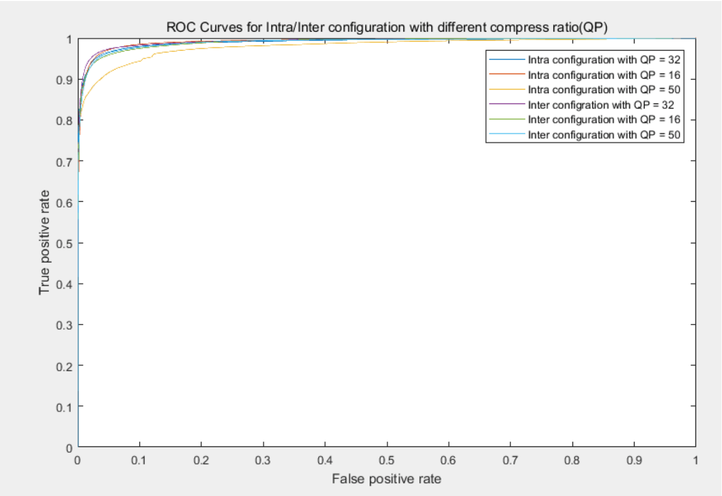


Figure 3-29 ROC curve of different QP values.

Figure 3-29 shows the accuracy of saliency in different compression ratio. When QP is high, the compression ratio is high, and the accuracy is high.

In this chapter, results of the proposed method are compared with state of art image saliency detection model. The proposed show better results in terms of accuracy and speed. The proposed method shows a better result in intra frame than inter frame. Because the intra frame has more robust features than inter frames. Due to the way of the HEVC encoder working, features extracted from inter frames have less correlation with saliency than the features extracted from intra frame. The proposed method is a generalized saliency estimate model that show robust result with different HEVC encoding configuration. The proposed method also show drawback when estimating saliency in frames where many blocks are encoded with ADI. The intra saliency was trained by 10 different videos. However, the videos are filmed in a similar theme, where saliency object is surrounded with natural texture such as snow, grass, and sand. No artificial background was included in the training set such as building, street, and indoor texture. This may make the machine learning model fail to estimate the saliency object in such backgrounds.

Chapter 4. Conclusion and further work

In this thesis, a saliency estimation algorithm by using Bayesian classification model in HEVC compressed domain is achieved. In the chapter, the summary of our work, the main contribution, and the further work are presented.

4.1. Conclusion

The background of human visual system and visual attention models in compressed domain is introduced. After performances analysis, the gap of the existing methods is: The effect of features in compressed domain on attracting visual saliency cannot be found. In addition, the relationship between the features in compression domain and the visual saliency is unclear. So a novel visual attention model is established. The relationship between HEVC features and visual saliency has been analyzed and discussed. Intra mode saliency (spatial saliency) is achieved by applying Bayesian modeling of relative intra features. Then inter saliency is achieved by adding the spatial saliency from reference block with temporal saliency from motion vector. Several classification methods such as SVM, FNN and decision tree are used to testing features. The variety of compression ratio has been discussed. The saliency results are good when the compression ratio is high.

The main advantage of this method is to achieve saliency map in HEVC compressed domain without fully decoded bitstream. The proposed method is a generalized saliency estimate model that show robust result with different HEVC encoding configuration in terms of speed. However, our methods may unable to achieve videos that include complex background.

4.2. Further work

The Bayesian model is used to train and classify the features and ground-truth. In recent research, the neural networks have good prospects in machine learning. The artificial neural network can produce an automatic identification system by comparing the local situation. The symbolic systems such as naive Bayesian model, they also have inference function based on a collection of algorithms. For the further work, the Bayesian model can change as the neural networks.

Reference

[1] H.E. Egeth and S. Yantis, “Visual Attention: Control, Representation, and Time Course,” *Ann. Rev. Psychologogy*, vol. 48, pp. 269-297, 1997.

Doi: 10.1146/annurev.psych.48.1.269

[2] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini, “Human neural systems for face recognition and social communication,” *Biol Psychiatry* 51(1), pp. 59–67, 2002.

[3] D. Flynn *et al*., "Overview of the Range Extensions for the HEVC Standard: Tools, Profiles, and Performance," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 4-19, Jan. 2016.

Doi: 10.1109/TCSVT.2015.2478707

[4] C. Gale and A. F. Monk, “Where am i looking? the accuracy of video-mediated gaze awareness,” *Percept Psychophys* 62(3), pp. 586–595, 2000.

[5] C. Koch and S. Ullman, “Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry,” *Human Neurobiology*, vol. 4, no. 4, pp. 219-227, 1985.

[6] D.J. Simons and D.T. Levin, “Failure to Detect Changes to Attended Objects,” *Investigative Ophthalmology and Visual Science*, vol. 38, no. 4, p. 3273, 1997.

[7] S. Treue, “Neural Correlates of Attention in Primate Visual Cortex,” *Trends in Neurosciences,* vol. 24, no. 5, pp. 295-300, 2001.

[8] E.T. Rolls and G. Deco, “Attention in Natural Scenes: Neurophysiological and Computational Bases,” *Neural Networks,* vol. 19, no. 9, pp. 1383-1394, 2006.

[9] Q. Wan, K. Panetta and S. Agaian, "A video forensic technique for detecting frame integrity using human visual system-inspired measure," *2017 IEEE International Symposium on Technologies for Homeland Security (HST)*, Waltham, MA, 2017, pp. 1-6.

doi: 10.1109/THS.2017.7943466

[10] Y. Yang, M. Yang, S. Huang, Y. Que, M. Ding and J. Sun, "Multifocus Image Fusion Based on Extreme Learning Machine and Human Visual System," in *IEEE Access*, vol. 5, no. , pp. 6989-7000, 2017.

doi: 10.1109/ACCESS.2017.2696119

[11] K. Koch, J. McLean, R. Segev, M.A. Freed, M.J. Berry, V. Balasubramanian, and P. Sterling, “How Much the Eye Tells the Brain,” *Current Biology*, vol. 25, nos. 16-14, pp. 1428-34, 2006.

[12] L. Itti, C. Koch, and E. Niebur, “A Model of Saliency-Based Visual Attention for Rapid Scene Analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[13] M. C. Park and S. Mun, "Overview of Measurement Methods for Factors Affecting the Human Visual System in 3D Displays," in *Journal of Display Technology,* vol. 11, no. 11, pp. 877-888, Nov. 2015.
doi: 10.1109/JDT.2015.2389212

[14] G. J. Suaning, "Strategic circuits for neuromodulation of the visual system," *2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC)*, Chiba, 2017, pp. 291-294.
doi: 10.1109/ASPDAC.2017.7858336

[15] A. E. Nidecker, P. Y. Shen, J. H. Pettey, B. Dhillon, M. K. Reddy and K. Khaderi, "The visual system as a proxy for evaluation of brain function," *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Budapest, 2016, pp. 004171-004175.
doi: 10.1109/SMC.2016.7844886

[16] Wenbo Li, Haiwei Pan, Xiaoqin Xie, Zhiqiang Zhang and Qilong Han, "MICS: Medical image classification visual system," *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Shenzhen, 2016, pp. 1032-1039.
doi: 10.1109/BIBM.2016.7822664

[17] A. Hazan, Y. Harel and R. Meir, "Learning an attention model in an artificial visual system," *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, Eilat, 2016, pp. 1-5.
doi: 10.1109/ICSEE.2016.7806115

[18] D. E. Broadbent, "Perception and ergonomics," in *IEEE Transactions on Professional Communication*, vol. PC-21, no. 1, pp. 34-37, March 1978.

doi: 10.1109/TPC.1978.6592439

[19] R. W. Conners and C. A. Harlow, "Some theoretical considerations concerning texture analysis of radiographic images," *1976 IEEE Conference on Decision and Control including the 15th Symposium on Adaptive Processes*, Clearwater, FL, USA, 1976, pp. 162-167.

doi: 10.1109/CDC.1976.267723

[20] B. Ratnaparkhi, L. Katore and J. S. Umale, "Improved student psychology prediction & recommendation strategy using 2 state data analysis," *2015 Global Conference on Communication Technologies (GCCT)*, Thuckalay, 2015, pp. 869-873.

doi: 10.1109/GCCT.2015.7342786

[21] J. Polcari, "An Informative Interpretation of Decision Theory: Scalar Performance Measures for Binary Decisions," in *IEEE Access*, vol. 2, no. , pp. 1456-1480, 2014.

doi: 10.1109/ACCESS.2014.2377593

[22] D. Chinn and K. Martin, "Work in Progress: Adapting the Treisman Model to Computer Science," *Proceedings. Frontiers in Education. 36th Annual Conference*, San Diego, CA, 2006, pp. 23-24.

doi: 10.1109/FIE.2006.322521

[23] X. Song, Z. Zhou, H. Guo, X. Zhao and H. Zhang, "Adaptive Retinex Algorithm Based on Genetic Algorithm and Human Visual System," *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, 2016, pp. 183-186.

doi: 10.1109/IHMSC.2016.27

[24] A. Buades and R. Grompone von Gioi, "Visual system inspired algorithm for contours, corner and T-junction detection," *2016 6th European Workshop on Visual Information Processing (EUVIP)*, Marseille, 2016, pp. 1-6.

doi: 10.1109/EUVIP.2016.7764586

[25] A. Azaza, L. Kabbai, M. Abdellaoui and A. Douik, "Salient regions detection method inspired from human visual system anatomy," *2016 2nd International*

Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, 2016, pp. 155-160.

doi: 10.1109/ATSIP.2016.7523087

[26] M. Khalil, Jian-Ping Li, K. Kumar, Xiao-Long Tang and Ping Kuang, "Color constancy models inspired by human visual system: Survey paper," *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP),* Chengdu, 2015, pp. 432-435.

[27] Q. Wan and K. Panetta, "A facial recognition system for matching computerized composite sketches to facial photos using human visual system algorithms," *2016 IEEE Symposium on Technologies for Homeland Security (HST)*, Waltham, MA, 2016, pp. 1-6.

doi: 10.1109/THS.2016.7568945

[28] Di Zhang, Ligu Zhu, Chengcheng Wang and Lei Zhang, "TimeSpiral, an enhanced interactive visual system for time series data," *2016 2nd International Conference on Information Management (ICIM)*, London, 2016, pp. 127-133.

doi: 10.1109/INFOMAN.2016.7477546

[29] K. Muthuswamy and D. Rajan, "Salient Motion Detection in Compressed Domain," in *IEEE Signal Processing Letters*, vol. 20, no. 10, pp. 996-999, Oct. 2013.

doi: 10.1109/LSP.2013.2277884

[30] Wen-Huang Cheng, Wei-Ta Chu, Jin-Hau Kuo and Ja-Ling Wu, "Automatic video region-of-interest determination based on user attention model," *2005 IEEE International Symposium on Circuits and Systems*, 2005, pp. 3219-3222 Vol. 4.

doi: 10.1109/ISCAS.2005.1465313

[31] J. J. Anaya, A. Sánchez, J. J. Giménez and D. Ruiz, "High Frame Rate Compression Efficiency and Backwards Compatibility using HEVC," *SMPTE 2015 Annual Technical Conference and Exhibition*, Loews Hollywood Hotel, Hollywood, CA, 2015, pp. 1-21.

doi: 10.5594/M001642

[32] S. Parikh, D. Ruiz, H. Kalva and G. Fernández-Escribano, "Content dependent

intra mode selection for medical image compression using HEVC," *2016 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, 2016, pp. 561-564.

doi: 10.1109/ICCE.2016.7430731

[33] A. Heindel; E. Wige; A. Kaup, "Low-Complexity Enhancement Layer Compression for Scalable Lossless Video Coding based on HEVC," in *IEEE Transactions on Circuits and Systems for Video Technology* , vol.PP, no.99, pp.1-1

doi: 10.1109/TCSVT.2016.2556338

[34] V. Sanchez, M. Hernandez-Cabronero, F. Auli-Llinàs and J. Serra-Sagristà, "Fast lossless compression of whole slide pathology images using HEVC intra-prediction," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 1456-1460.

doi: 10.1109/ICASSP.2016.7471918

[35] H. Zhang, Q. Zhou, N. Shi, F. Yang, X. Feng and Z. Ma, "Fast intra mode decision and block matching for HEVC screen content compression," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 1377-1381.

doi: 10.1109/ICASSP.2016.7471902

[36] C. H. Chan, H. S. Ji and W. J. Tsai, "Adaptive accordion transformation based video compression method on HEVC," *2016 Visual Communications and Image Processing (VCIP)*, Chengdu, 2016, pp. 1-4.

doi: 10.1109/VCIP.2016.7805586

[37] J. J. Anaya and D. Ruiz, "HEVC Mezzanine Compression for UHD Transport over SDI and IP infrastructures," *SMPTE 2016 Annual Technical Conference and Exhibition*, Los Angeles, CA, 2016, pp. 1-21.

doi: 10.5594/M001701

[38] K. Naser, V. Ricordel and P. Le Callet, "A foveated short term distortion model for perceptually optimized dynamic textures compression in HEVC," *2016 Picture Coding Symposium (PCS)*, Nuremberg, 2016, pp. 1-5.

doi: 10.1109/PCS.2016.7906311

[39] S. Li; M. Xu; Y. Ren; Z. Wang, "Closed-form Optimization on Saliency-guided Image Compression for HEVC-MSP," in *IEEE Transactions on Multimedia* , vol.PP, no.99, pp.1-1

doi: 10.1109/TMM.2017.2721544

[40] D. Springer, F. Simmet, D. Niederkorn and A. Kaup, "Robust Rotational Motion Estimation for efficient HEVC compression of 2D and 3D navigation video sequences," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, pp. 1379-1383.

doi: 10.1109/ICASSP.2013.6637877

[41] D. Springer, M. Frank, F. Simmet, D. Niederkorn and A. Kaup, "Spiral search based fast rotation estimation for efficient HEVC compression of navigation video sequences," *2013 Picture Coding Symposium (PCS)*, San Jose, CA, 2013, pp. 201-204.

doi: 10.1109/PCS.2013.6737718

[42] J. Ström *et al.*, "High quality HDR video compression using HEVC main 10 profile," *2016 Picture Coding Symposium (PCS)*, Nuremberg, 2016, pp. 1-5.

doi: 10.1109/PCS.2016.7906372.

[43] X. He, X. Li, L. Qing and S. Su, "Study on Segmentation-Based HEVC Compression Performance," *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, 2016, pp. 417-420.

doi: 10.1109/ISCID.2016.2104.

[44] S. Parikh; D. Ruiz; H. Kalva; G. Fernandez-Escribano; v. Adzic, "High Bit-Depth Medical Image Compression with HEVC," in *IEEE Journal of Biomedical and Health Informatics* , vol.PP, no.99, pp.1-1， 2017

doi: 10.1109/JBHI.2017.2660482.

[45] S. Parikh, H. Kalva and V. Adzic, "Evaluation of HEVC compression for high bit depth medical images," *2016 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, 2016, pp. 311-314.

doi: 10.1109/ICCE.2016.7430625

[46] T. K. Tan *et al.*, "Video Quality Evaluation Methodology and Verification Testing of HEVC Compression Performance," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 76-90, Jan. 2016.
doi: 10.1109/TCSVT.2015.2477916

[47] R. Weerakkody, M. Mrak, V. Baroncini, J. R. Ohm, T. K. Tan and G. J. Sullivan, "Verification testing of HEVC compression performance for UHD video," *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Atlanta, GA, 2014, pp. 1083-1087.
doi: 10.1109/GlobalSIP.2014.7032288

[48] S. Vítek, L. Krasula, M. Klíma, V. Hvêzda and M. H. Martínez, "Influence of HEVC compression on event detection in security video sequences," *2013 47th International Carnahan Conference on Security Technology (ICCST)*, Medellin, 2013, pp. 1-6.
doi: 10.1109/CCST.2013.6922066

[49] F. Saab, I. H. Elhajj, A. Kayssi and A. Chehab, "Energy analysis of HEVC compression with stochastic processing," *MELECON 2014 - 2014 17th IEEE Mediterranean Electrotechnical Conference*, Beirut, 2014, pp. 170-176.
doi: 10.1109/MELCON.2014.6820526

[50] D. Springer, F. Simmet, D. Niederkorn and A. Kaup, "Motion vector analysis based homography estimation for efficient HEVC compression of 2D and 3D navigation video sequences," *2013 IEEE International Conference on Image Processing*, Melbourne, VIC, 2013, pp. 1742-1746.
doi: 10.1109/ICIP.2013.6738359

[51] Y. Yuan; D. Li; M. Q. H. Meng, "Automatic Polyp Detection via A Novel Unified Bottom-up and Top-down Saliency Approach," in *IEEE Journal of Biomedical and Health Informatics* , vol.PP, no.99, pp.1-1
doi: 10.1109/JBHI.2017.2734329

[52] M. Hossny, S. Nahavandi, D. Creighton, C. Lim and A. Bhatti, "Enhanced decision fusion of semantically segmented images via local majority saliency map,"

in *Electronics Letters*, vol. 53, no. 15, pp. 1036-1038, 7 20 2017.

doi: 10.1049/el.2016.4709

[53] N. Mu, X. Xu and X. Zhang, "A superpixel-based saliency model for robust autofocus in low contrast images," *2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*, Taipei, Taiwan, 2017, pp. 77-78.

doi: 10.1109/ICCE-China.2017.7991003

[54] M. Fujimura, K. Imamura and H. Kuroda, "Application of saliency map to restraint scheme of attack to digital watermark using seam carving," *2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*, Taipei, Taiwan, 2017, pp. 347-348.

doi: 10.1109/ICCE-China.2017.7991138

[55] R. Y. Xiao and M. H. Yeh, "A new method with saliency detection for image quality assessment," *2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*, Taipei, Taiwan, 2017, pp. 75-76.

doi: 10.1109/ICCE-China.2017.7991002

[56] D. Zhao, Y. Ma, Z. Jiang and Z. Shi, "Multiresolution Airport Detection via Hierarchical Reinforcement Learning Saliency Model," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 6, pp. 2855-2866, June 2017.

doi: 10.1109/JSTARS.2017.2669335

[57] Gonzalez, R., & Woods, R. *Digital image processing*, 3rd ed. Upper Saddle River, NJ: Pearson/Prentice Hall, 2008.

[58] Wien M, *High Efficiency Video Coding: coding tools and specification.* Springer, Heidelberg, 2015.

[59] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa, "Effcient hierarchical

graph-based video segmentation," in *Computer Vision and Pattern Recognition (CVPR)*, 2010 IEEE Conference on. IEEE, 2010, pp. 2141-2148.

[60] Y. Fang, W. Lin, Z. Chen, C. M. Tsai and C. W. Lin, "A Video Saliency

Detection Model in Compressed Domain," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 27-38, Jan. 2014. doi: 10.1109/TCSVT.2013.2273613

[61] S. H. Khatoonabadi, N. Vasconcelos, I. V. Bajić and Yufeng Shan, "How many bits does it take for a stimulus to be salient?," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston*, MA, 2015, pp. 5501-5510. doi: 10.1109/CVPR.2015.7299189

[62] C. Guo and L. Zhang, "A novel multi-resolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[63] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.,* Jun. 2010, pp. 2376–2383.

[64] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[65] T. Wiegand, J.-R. Ohm, G. J. Sullivan, W.-J. Han, R. Joshi, T. K. Tan, and K. Ugur, "Special section on the joint call for proposals on High Efficiency Video Coding (HEVC) standardization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 12, pp. 1661–1666, Dec. 2010.

[66] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards—Including High Efficiency Video Coding (HEVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1668-1683, Dec. 2012.

[67] F. Bossen, B. Bross, K. Suhring, and D. Flynn, "HEVC complexity and implementation analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1684-1695, Dec. 2012.

[68] E.H. Yang and X. Yu, "Rate distortion optimization for H.264 inter frame coding: A general framework and algorithms," *IEEE Trans. Image Process.*, vol. 16, no. 7, pp.

1774-1784, Jul. 2007.

[69] B. Li, G. J. Sullivan, and J. Xu, "Compression performance of high efficiency video coding (HEVC) working draft 4," *in Proc. IEEE Int. Conf. Circuits Syst.,* May 2012, pp. 886-889.

[70] M. Flierl and B. Girod, "Generalized B pictures and the draft H.264/AVC video compression standard," IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 7, pp. 587-597, Jul. 2003.

[71] M. Xu, L. Jiang, X. Sun, Z. Ye and Z. Wang, "Learning to Detect Video Saliency With HEVC Features," in *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 369-385, Jan. 2017.

doi: 10.1109/TIP.2016.2628583

[72] C. Guo and L. Zhang, “A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression,” in I*EEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[73] Jonathan Harel , Christof Koch , Pietro Perona, “Graph-Based Visual Saliency,” *Proceedings of the 19th International Conference on Neural Information Processing Systems*, p.545-552, December 04-07, 2006.

[74] K. Wu and Jiang Ke, "A scheme of real-time traffic classification in secure access of power enterprise based on improved Naive Bayesian classification algorithm," *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS),* Beijing, 2016, pp. 1017-1021.

doi: 10.1109/ICSESS.2016.7883239

[75] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, “A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation” *2016 Computer Vision and Pattern Recognition (CVPR).*

[76] Cortes, C.; Vapnik, V. Support-vector networks. Machine Learning. 1995, 20 (3): 273–297. doi:10.1007/BF00994018.