International Journal of Research in Advent Technology, Special Issue, March 2019
E-ISSN: 2321-9637
International Conference on Technological Emerging Challenges (ICTEC-2019)
Available online at www.ijrat.org

Image Saliency Detection in Compressed Domain

[1]S.Spandana, [2]A.Navya Sai Sri, [3]S.Manitejaswini, [4]R.Hemalatha, [5]Y.Lakshmi Prasanna,[6]SK.Afroze
[1,2]Assistant Professor, Dept of ECE, Tirumala Engineering College, Narasaraopet, A.P, India
[3,4,5,6]B.Tech scholar, Dept of ECE, Tirumala Engineering College, Narasaraopet, A.P, India
sspandana2017@gmail.com

Abstract: Recently, many saliency detection models have been proposed for image in uncompressed (pixel) domain. However, image over Internet is always stored in compressed domains, such as JPEG2, H.264, and JPEG4 Visual. In this paper, we propose a novel video saliency detection model based on feature contrast in compressed domain. Four types of features including luminance, color and texture are extracted from the discrete cosine transform coefficients . The static saliency map of unpredicted frames (I frames) is calculated on the basis of luminance, color, and texture features. A new fusion method is designed to the static saliency maps to get the final saliency map for each image frame. Due to the directly derived features in compressed domain, the proposed model can predict the salient regions efficiently for image frames. Experimental results on a public database show superior performance of the proposed image saliency detection model in compressed domain.

Keywords : DCT, saliency.

1. INTRODUCTION

Visual attention is an important characteristic in the Human Visual System (HVS) for visual information
processing. With large amount of visual information, visual attention would selectively process the important part by filtering out others to reduce the complexity of scene analysis .These important visual information is also termed as salient regions or Regions of Interest (ROIs) in natural images. There are two different approaches in visual attention mechanism: bottom-up and top-down. Bottom-up approach, which is data driven and task independent, is a perception process for automatic salient region selection for natural scenes [1]–[8],while top-down approach is a task-dependent cognitive processing affected by the performed tasks, feature distribution of targets ,etc.
Over the past decades, many studies have tried to propose computational models of visual attention for various
Multimedia processing applications, such as visual retargeting , visual quality assessment , visual coding , etc. In these applications, the salient regions extracted from saliency detection models are processed specifically since they attract much more humans' attention compared with other regions. Currently, many bottom up saliency detection models have been proposed for2D images/videos.
processing applications in the past decades. Our previous study has also demonstrated that DCT coefficients can be adopted for effective feature representation in saliency detection . Therefore, we use DCT coefficients for feature extraction for image patches in this study.
In essence, the input stereoscopic image and depth map are firstly divided into small image patches. Color, luminance and texture features are extracted based on DCT coefficients of each image patch from the original image, while depth feature is extracted based on DCT coefficients of each image patch in the depth map. Feature contrast is calculated based on center surround feature difference, weighted by a Gaussian model of
Spatial distances between image patches for the consideration of local and global contrast. A new fusion method is designed to combine the feature maps to obtain the final saliency map for 3D images. Additionally, inspired by the viewing influence from centre bias and the property of human visual acuity in the HVS, we propose to incorporate the centre bias factor and human visual acuity into the proposed model to enhance the saliency map. The Centre-Bias Map (CBM) calculated based on centre bias factor and a statistical model of human visual sensitivity are adopted to enhance the saliency map for obtaining the final saliency map of 3D images.
Existing 3D saliency detection models usually adopt depth information to weight the traditional 2D saliency map or combine the depth saliency map and the traditional 2D saliency map simply to obtain the saliency map of 3D images. Different from these existing methods ,the proposed model adopts the low-level features of color, luminance, texture and depth for saliency calculation in a whole framework and designs a novel fusion method to obtain the saliency map from feature maps. Experimental results on
Eye-tracking databases demonstrate the superior performance of the proposed model over other existing methods
The remaining of this paper is organized as follows. Section II introduces the related work in the literature.

In Section III, the proposed model is described in detail. Section IV provides the experimental results on eye tracking databases. The final section concludes the paper.

2. RELATED WORK

Visual attention mechanism is an important characteristic in the human visual system (HVS). The selective attention may be stimulus-driven or goal-driven corresponding to bottom up and top-down approaches in perception process . Existing studies have explored visual attention mechanism from the various aspects such as psychology, biology, computer vision. In the 1980s, Treisman et al. proposed the famous Feature Integration Theory (FIT). According to this theory, the early selective attention mechanism leads some image regions to be salient for their different features (including color, intensity, orientation and so on) from their surroundings . Meanwhile, Koch et al. proposed a neurophysiological model of visual attention.

Recently, researchers in the area of computer vision have started to build computational models of visual attention for the emerging interest in the HVS. Itti et al. proposed a saliency detection model based on the neuronal architecture of the primates' early visual system. The saliency map is obtained through the calculation of multiscale center-surround differences by using three features including intensity, color, and orientation. Harel et al. proposed a graph-based saliency detection model based on the study . In this model, the saliency map is calculated based on two steps: forming activation maps on several features and the normalization of these feature maps . Hou et al. defined the concept of spectral residual to design a visual attention model. The spectral residual is computed based on the Fourier transform. Achanta et al. [6] tried to obtain more frequency information to get a better saliency measure. The difference of Gaussian (DoG) is used to extract the frequency information in that model . Goferman et al.designed a contextaware saliency detection model by including more context information in the final saliency map. The center-surround differences of patches are used for saliency detection.

Besides the saliency detection models for images are been proposed in this area. Guo et al. proposed a phase-based saliency detection model for image in [5]. This model obtains the saliency map through inverse Fourier transform on a constant amplitude and the original phase spectrum of input images based on the following features: intensity, color. Itti et al. [8] developed a model to detect the low-level surprising events in image; the surprising events are defined as the important information attracting human beings' attention in video. Zhai et al. [9] built a image saliency detection model by combining the spatial and temporal saliency maps. The color histograms of images are used for the spatial saliency detection, while the planar motion between images (estimated by applying RANSAC on point correspondences in the scene) is adopted for the temporal saliency detection [9]. In , the authors designed a dynamic visual attention model based on the rarity of features. The incremental coding length (ICL) is defined to measure the entropy gain of each feature for saliency calculation.

All these saliency detection models mentioned above are implemented in uncompressed domain. As to these saliency detection models, coded images have to be decompressed into spatial domain to extract features for saliency detection. In this paper, we propose a image saliency detection model in compressed domain. As mentioned in the first section, we calculate the saliency map of each image in 8 × 8 block level. The DCT coefficients of 8 × 8 image blocks in unpredicted frames (I frames).The luminance, color, texture of images are calculated from the DCT coefficients. Then, we calculate the static saliency map of unpredicted frames based on luminance, color, and texture features. Furthermore, we design a new fusion method to the static saliency map to obtain the final saliency map. Due to the directly derived features from the compressed domain, the proposed saliency detection model obtains promising results, as shown in the experimental section.

3. PROPOSED FRAMEWORK

In this section, we describe the proposed model in detail. The proposed framework is depicted in Fig. 1. Firstly, three features including luminance, color, and texture are extracted from the image bit stream for unpredicted frames (I frames). Then, the static saliency map is obtained based on the features of luminance, color, and texture for unpredicted frames. Finally, the static saliency map to get the final saliency map for each image. Here, we use JPEG4 ASP image to extract features in compressed domain. The features of other types of images such as JPEG2 image can be extracted in a similar way.

a. Feature Extraction From Image Bit stream:

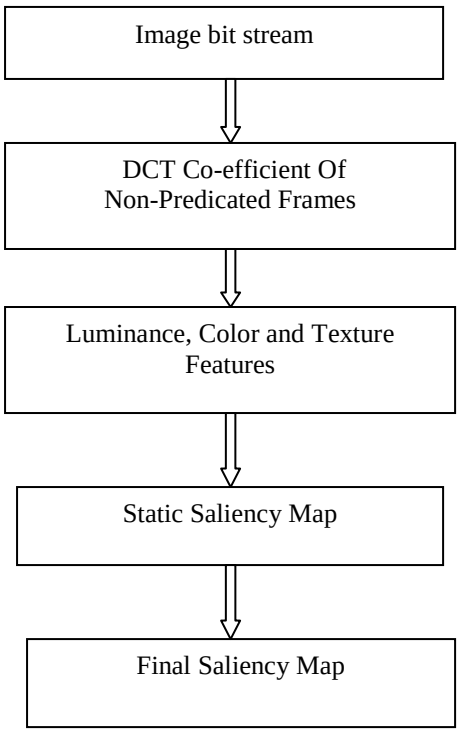
The proposed model uses DCT coefficients of unpredicted frames (I frames) to get luminance, color, and texture features. Here, we do not use DCT coefficients of the predicted frames since these DCT coefficients represent the interpredicted block residue information. These residue information cannot be used to obtain the static saliency map by the proposed method. Meanwhile, there are no motion vectors for unpredicted frames in video bit stream. Therefore, in this paper, the DCT coefficients of unpredicted frames are used to

calculate the static saliency map for these unpredicted frames.

b. DCT Coefficient Extraction From Image Bitstream:

A natural image object is composed of a sequence (at different time points) of 2-D representations, which are referred to as image object planes (VOPs). The VOPs are coded using macroblocks by exploiting both temporal redundancies and spatial redundancies. Usually, a VOP consists of one or several image packets (slices) and each image packet is composed of an integer number of consecutive macroblocks. In each macroblock, the DCT coefficients are coded. The coded DCT coefficients are the 64 DCT coefficients encoded by zig-zag scanning and run-length-encoded, and the VLC.

In a similar way, VLC tables of DCT coefficients are used to decode the coded DCT coefficients. The fixed length decoding is used to obtain the real DCT coefficients for image frames [10].



c. Feature Calculation Based on DCT Coefficients:

In JPEG4 ASP image, DCT coefficients in one 8 × 8 block are composed of one DC coefficient and 63 AC coefficients. In each block, the DC coefficient is a measure of the average energy for the 8 × 8 block, while other 63 AC coefficients represent detailed frequency properties of this block. As mentioned above, YCrCb color space is used in JEPG4 image bit stream. In the YCrCb color space, the Y channel represents the luminance component, while Cr and Cb represent the chroma components. Thus, the DC coefficients in DCT blocks from the Y, Cr, and Cb channels are used to represent one luminance feature and two color features for 8 × 8 blocks as follows:

$$L = DC\gamma \quad (1)$$

$$C_1 = DC_{Cr} \quad (2)$$

$$C_2 = DC_{Cb} \quad (3)$$

where L, C1, and C2 represent one luminance and two color features in each 8 × 8 DCT block, respectively, $DC\gamma$, $^{DC}c_r$, and $^{DC}c_b$ are DC coefficients from the Y, Cr, and Cb components in each DCT block, respectively. It is noted that four 8 × 8 luminance blocks share two 8 × 8 chrominance blocks in the 4:2:0 chrominance format.

As mentioned above, AC coefficients include the detailed frequency information and existing studies have shown that AC coefficients can represent texture information for image blocks [11], [12]. In YCrCb color space, Cr and Cb components mainly include color information and little texture information is included in these two channels. Thus, only the AC coefficients in Y component are used to represent the texture information for images. In one DCT block, most of the energy is included in the first several low-frequency coefficients, which are in the left-upper corner of the block. The AC coefficients in the

right-bottom corner of DCT blocks are equal to or close to zero and they are neglected during the quantization in coding process. In the progressive coding, the AC coefficients in one DCT block are ordered by zig-zag scanning. As the high-frequency AC coefficients include little energy for each DCT block, we use the first several AC coefficients to represent the texture feature of the DCT block. The existing study in [12] has shown that the first nine AC coefficients can represent most energy in each DCT block. Therefore, here, we use the first nine AC coefficients in each DCT block to represent the texture feature as follows

T = {AC01, AC10, AC20, AC11... AC30}. (4)

d. Saliency Detection in Compressed Domain:

Based on the above description, the luminance, color, and texture features (L, C1, C2, and T) for unpredicted frames can be extracted from the DCT coefficients. In this paper, we use these features to calculate the saliency map for image frames in compressed domain.

e. Static Saliency Map Calculation:

Existing studies have shown that observers will be attracted by the regions with different features from its surrounding when looking at a natural scene [1], [2]. The features which can be used to discriminate the image regions include intensity, color, motion, and so on. Based on the FIT [1], the center surround differences of 8 × 8 DCT blocks are used to detect salient regions for images. The features of luminance, color and texture extracted from the DCT coefficients are used to calculate the DCT block differences for saliency detection in this paper.

It is commonly accepted that the HVS is highly space variant since there are different densities of cone photoreceptor cells in the retina of human eyes [13]. On the retina, the fovea has the highest density of cone photoreceptor cells, and thus the focus area is perceived at the highest resolution. The visual acuity decreases with the increasing eccentricity from the fixation areas [13]. This means that the HVS is more sensitive to the center-surround differences from the blocks with nearer distance compared with those from the farther blocks. Here, we use a Gaussian model to simulate this mechanism for weighting the center-surround differences among image blocks for saliency detection. The feature map of each image frame is calculated as follows:

$$Si^k = \sum \alpha_{ij} D_{ij}^k \quad (6)$$

$$\sigma_{ij} = 1/\sigma\sqrt{2}\prod e^{-d^2/2\sigma^2} \quad (7)$$

where *Ski* indicates saliency value of the *i*th DCT block in the feature map with feature *k*; {*L*,*C*1,*C*2, *T, V*}; 9 σ is a parameter of the Gaussian distribution, *dij* is the Euclidean distance between DCT blocks *i* and *j*, *Dkij* is the feature differences between DCT blocks *i* and *j* with feature *k*.

As depicted in Fig. 1, the features of luminance, color and texture are used to calculate the static saliency map for unpredicted frames. The luminance and color features only include one DC coefficient value from the luminance and color channels; thus, the feature differences of luminance and color among DCT blocks are represented as DC coefficient differences from the luminance and color channels. Since the texture feature is represented as a vector including nine AC coefficients, the Euclidean distance between the vectors are used to compute the texture difference between DCT blocks. The static saliency map *Ss* for unpredicted frames is calculated as linear combination of four feature maps from the luminance, color, and texture features (*L, C*1, *C*2, *T*) as follows:

$$Ss = \sum \beta_\theta N_\theta$$

Where N is normalization operation; $\theta \in \{SK\}$; β_θ is the parameter determining the weight for each feature map. In this paper, we set β_θ =¼. The final saliency map calculated by the static saliency map Ss. We will describe how to compute the final saliency map as follows.

f. Final Saliency Map Calculation*:*

Based on the above description, we can obtain the static saliency map for unpredicted image frames (I frames). Meanwhile, the static saliency map of predicted frames cannot be computed since the DCT coefficients of these predicted frames represent the DCT block residue information and cannot be used to calculate the static saliency map. Here, we use the static/motion saliency map of the previous unpredicted/predicted frames to replace that of the current predicted/unpredicted ones based on the implicit memory theory [24], [25]. Existing studies have shown that the focal attention and eye movements are guided by the recently attended locations and the implicit memory traces of context cueing [24]–[26]. These studies demonstrate that the previous attended targets will trigger the attention traces utilized in the following several fixations. Thus, human beings will continue focusing on the similar locations in future frames with these from the previous frames without much context change. Generally, the content in the consecutive image frames will not change greatly, and thus the saliency maps (static or motion) of the consecutive image frames are very similar in image. Therefore, we can use the static or motion saliency map of the previous image frames to represent that of the current ones.

As there is no motion saliency map for unpredicted frames the motion saliency map of the previous predicted frame is adopted to represent that of the current unpredicted frame. Thus, the final saliency

map for unpredicted frames (I frames) is calculated as follows:

$$S = f(S_s, S_{mp}) \quad (9)$$

Where S is the final saliency map of the current unpredicted frame, Ss is the static saliency map of the current unpredicted frame ,f (S1, S2) is the fusion function to get the final saliency map from the saliency maps of S1 and S2.

Similarly, the static saliency map of the previous unpredicted frame is used to represent that of the current predicted frame, and thus the final saliency map of the predicted frames (P and B frames) is computed as follows:

$$S = f(Ssp, Sm) \quad (10)$$

Where *S* is the final saliency map of the current predicted frame, *Ssp* is the static saliency map of the previous unpredicted frame, *Sm* is the motion saliency map of the current predicted frame, *f* (*S*1, *S*2) is the fusion function to get the final saliency map from the saliency maps of *S*1 and *S*2.

According to (9) and (10), we can calculate the final saliency map for video frames based on the static saliency and motion saliency maps. We will describe the fusion method *f* [in (9) and (10)] for the static and motion saliency maps in the next section.

Saliency Map Fusion*:* Currently, there are many fusion methods for combining the static saliency and motion saliency maps into the final saliency map [33]. In this paper, we have tried several common fusion methods in [33] for combing the static saliency and motion saliency maps as follows.

1.Normalized and sum (NS): the most simple fusion method that normalizes the static saliency and motion
saliency maps to the same dynamic range, and then sums these two maps to obtain the final saliency map as
follows [33]:

$$S = \sum_i N(S_i) \quad (11)$$

2) Normalization and maximum (NM): the fusion method that normalizes the static saliency map and motion
saliency map to the same dynamic range, and then uses the maximum value as the final saliency value at each location.

$$S = \max_i N(S_i) \quad (12)$$

where max is the maximum operator.

3) Normalization and product (NP): the fusion method that normalizes the static saliency map and motion saliency map to the same dynamic range, and then products the static saliency and motion saliency maps for the final saliency map

$$S = \prod_i N(S_i) \quad (13)$$

These three methods are the common fusion methods of the research area. However, there is no spatial competition between the static saliency map and the motion saliency map with the above fusion methods. In these fusion methods, the static saliency and motion saliency maps are considered with the same weighting whatever the differences between these two maps. To address the drawbacks with these fusion methods, we propose a new fusion method of parameterized normalization, sum and product (PNSP) based on the characteristics of the static saliency map and the motion saliency map. The final saliency map from the proposed fusion method PNSP for video frames is calculated as follows:

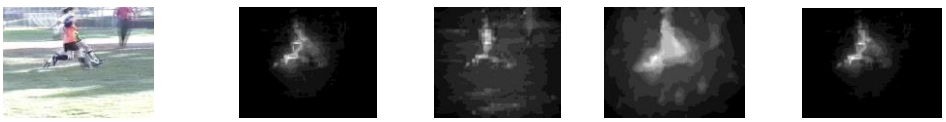
$$S = \gamma_1 S_s + \gamma_2 S_m + \gamma_3 S_s S_m \quad (14)$$

where *S* is the final saliency map for video frames, *Ss* is the static saliency map, S_m is the motion saliency map, $\gamma 1$, $\gamma 2$, and $\gamma 3$ are the parameters to determine the weighting of each component.

The weighting parameters in (14) are determined by the characteristic of the static saliency and motion saliency maps. A good saliency map should include the small compact salient regions rather than the spread salient points. If the salient regions in the motion saliency map are small and compact, the motion contrast of this frame can be considered strong. In this case, we can use a large weighting parameter $\gamma 2$ to weight the motion saliency map. Thus, the motion saliency map contributes much to the final saliency map. Similarly, a large weighting parameter $\gamma 1$ should be used to weight the static saliency map if the feature contrast is also strong in the static saliency map. As we can see, the weighting parameter $\gamma 3$ is used to measure the importance of these regions which both static saliency and motion saliency maps detect as salient.

Here, we set $\gamma 3 = (\gamma 1 + \gamma 2)/2$.

4. RESULTS



fig(a)

fig(b) fig(c) fig(d) fig(e)

Figures:(a)compressed image,(b)luminance image,(c)color image,(d)texture image,(e)final saliency image.

5. CONCLUSION

In this paper, we have proposed a image saliency detection model in compressed domain based on three types of features: luminance, color, texture. These four types of features are extracted from the DCT coefficients and motion vectors in video bit stream. The static saliency map is calculated from the features of luminance, color, and texture. A new fusion method has been designed to combine the static saliency and motion saliency maps to get the final saliency map for video frames. Experimental results based on a public database show that the proposed image saliency detection model outperforms the relevant existing ones.

It is noted that existing image saliency detection models are implemented in uncompressed domain. Compared with the video saliency detection in uncompressed domain, the proposed image saliency detection model in compressed domain can be used more conveniently in the Internet-based multimedia applications such as image retargeting, video quality assessment. Therefore, the proposed saliency detection model in compressed domain is significant in this research area. As the next step of the paper, we will explore various multimedia applications of the proposed image saliency detection model in compressed domain.

REFERENCES

[1] L. Jansen, S. Onat, and P. Konig, "Influence of disparity on fixation and saccades in free viewing of natural scenes," J. Vis., vol. 9, no. 1, p. 29, 2009

[2] J. Wang, P. Le Callet, S. Tourancheau, V. Ricordel, and M. Perreira Da Silva, "Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli," J. Eye Movement Res., vol. 5, no. 5, pp. 1–11, 2012.

[3] T. Jost, N. Ouerhani, R. V. Wartburg, R. Muri, and H. Hugli, "Contribution of depth to visual attention: Comparison of a computer model and human," in Proc. Early Cognitive Vis. Workshop, 2004, pp. 28.5–1.6.

[4] F. Shao, G. Jiang, M. Yu, K. Chen, and Y.-S. Ho, "Asymmetric coding of multi-view video plus depth based 3D video for view rendering," IEEE Trans. Multimedia, vol. 14, no. 1, pp. 157–167, Feb. 2012.

[5] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics," IEEE Trans. Image Process., vol. 22, no. 5, pp. 1940–1953, May 2013.

[6] Q. Huynh-Thu, M. Barkowsky, and P. Le Callet, "The importance of visual attention in improving the 3D-TV viewing experience: Overview and new perspectives," IEEE Trans. Broadcast., vol. 57, no. 2, pp. 421–431, Jun. 2011.

[7] P. Tseng, R. Carmi, I. Cameron, D. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes,"